UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA
DEPARTAMENTO DE COMUNICAÇÕES

Este exemplar corresponde à redação final da tese defendida por João Batista Destro Filho e aprovada pela Comissão Julgadora em 24/06/1994

Orientador

BASE TEÓRICA PARA O PROCESSAMENTO NEURAL-ADAPTATIVO DE SINAIS

Orientado: João Batista Destro Filho

Banca examinadora
Orientador: Prof. Dr. João Marcos Travassos Romano (presidente).
Prof. Dr. Max Gerken - DEE/EPUSP (membro).
Prof. Dr. Márcio Luiz de Andrade Netto - FEE/UNICAMP (membro).
Prof. Dr. João Bosco Ribeiro do Val - FEE/UNICAMP (suplente).

Tese apresentada à Faculdade de Engenharia Elétrica - FEE/UNICAMP como parte dos requisitos exigidos para a obtenção do título de *Mestre em Engenharia Elétrica.*

24 de Junho de 1994

"Não nasce a planta perfeita, não nasce o fruto maduro, e para ter a colheita é preciso semear" (Olavo Bilac).

"O reino de Deus é como um homem que lançou a semente à terra: ele dorme e acorda, de noite e de dia, mas a semente germina e cresce, sem que ele saiba como. A terra por si mesma produz fruto: primeiro a erva, depois a espiga e, por fim, a espiga cheia de grãos. Quando o fruto está no ponto, imediatamente se lhe lança a foice, porque a colheita chegou." (Marcos 4, 26 - 29)

"Se quiser lucrar em um ano, plante arroz; em 10 anos, plante uma árvore, em 100 anos, eduque" (Provérbio chinês).

"O maior investimento que se pode fazer é no próprio homem" (Gandhi).

"Gracias a la vida, que me ha dado tanto
me ha dado la risa, y me ha dado el llanto." (Violeta Parra).

"Canción y huayno para bailar,
canción y saya para bailar ..." (bis)

("Canción y Huayno" - Carnavalito - M. Nuñez - Bolívia).

...
"Huye de ese mortal desasosiego
que interroga a las sombras del Destino
la vida es ciega y el anos es ciego
pero nunca equivocan el camino
...
Renueva el corazón a cada hora
y aprende a renacer cada mañana
como el paisage al despuntar da aurora
como el sol que amanece en tu ventana"

(R. LEON)

DEDICATÓRIA

Este tese é dedicada a seus verdadeiros autores, ou seja, aqueles que constituem a razão de todo este trabalho:

A João Baptista;
A Adelaide (Pandy - xióngmão);
A José Paulo (Mec.);
A LH (Cê, Lú,..., etc);
A Leopoldina (a Imperatriz);

E

A
uma
pessoa
MUITO
MUITO
especial,
ainda desconhecida,
sempre presente.

"Estando trancadas as portas, veio Jesus, pôo-se no meio deles e disse: "A paz esteja convosco !". Depois disse a Tomé: "Põe aqui o teu dedo, e vê as minhas mãos. Põe a tua mão no meu lado. Não sejas incrédulo, mas homem de fé." Respondeu-lhe Tomé: "Meu Senhor, e meu Deus!". Disse-lhe Jesus: "Creste porque me viste. Felizes aqueles que crêem sem ter visto!" (João 20, 26-29).

AGRADECIMENTOS

A Deus pela saúde, pela vocação, pela família e pelo trabalho.

À CAPES, pelo financiamento deste trabalho.

Ao Prof. João Marcos, pela compreensão, paciência, orientação e pela amizade.

Aos Professores Sylvie Marcos (LSS - Supélec) e Paulo Diniz (COPPE - UFRJ) pelas referências bibliográficas, necessárias a esta tese. Ao Prof. Francesco Langoni (IB - UNICAMP) pela revisão da parte de neurofisiologia e à Profa. Adelaide Breda Destro (FFCLH - S. J. Rio Pardo) pelas sugestões quanto à parte de psicologia cognitiva.

A todos os professores, responsáveis direta ou indiretamente por esta tese, através dos cursos de graduação, pós-graduação ou do simples intercâmbio de idéias. Em particular, na UNICAMP, gostaria de agradecer a todos os docentes da FEE; aos Profs. Luiz K. Hotta, Ary O. Chiacchio e Maria Alice B. Grou (IMECC); à Profa. Maria Isabel Felisberti (IQ) e ao Prof. Sérgio Bordalo (FEM). Na USP, obrigado ao Profs. Jacyntho J. Angerami, Max Gerken e Paulo Boulos (EPUSP); aos Profs. William T. H. Liu e Leila M. Vespoli de Carvalho (IAG) e à Neuza Paes (CRAAE - INPE).

A todos aqueles que me ensinaram novas culturas: Professores Rancey P. Portela, Egon E. Zink e Professoras Christiane Maillart, Renata F. Arruda, Li Bin Bin; Ana Luiza Z. Degelo, Maria Salete e Fumiko Takasu (IEL - UNICAMP).

A todos os amigos do DECOM e da UNICAMP, colegas e funcionários. Em particular ao Sr. Airton (desenho); à Ademildes (impressão); ao Sr. Washington e Mari (xerox); à paciência da Sra. Wilma durante o curso de graduação e ao Sr. Motoyama (almoxarife).

A todos os professores e amigos de São José do Rio Pardo, responsáveis pela minha formação básica.

Ao Marcos A. Gallego pela iniciação no rádio.

A todos os membros da Seção de Língua Portuguesa da Rádio Internacional de Beijing, Rep. Popular da China, pelo intercâmbio cultural e pelo entretenimento.

A todos aqueles que participaram desta tese, direta ou indiretamente; e aos queridos Tio Ary, Tia Esther, Dani e Dri (SP).

RESUMO

Nesta tese realiza-se um estudo cujo objetivo é estabelecer uma base teórica para o processamento neural-adaptativo de sinais, uma nova técnica que visa conjugar as potencialidades intrínsecas das propriedades coletivas emergentes de redes neurais ao sólido formalismo matemático da filtragem adaptativa. Isto possibilita, ao mesmo tempo, uma análise matemática mais aprofundada dos princípios básicos do processamento de informação neural (auto-organização e processamento paralelo distribuído) e a generalização da aplicação de filtros adaptativos a situações mais complexas (por exemplo, no caso de aplicações que envolvam ruído não-gaussiano).

A base teórica proposta nesta tese está fundamentada numa série de analogias matemáticas e conceituais existentes entre as redes neurais e a filtragem adaptativa, que envolvem estruturas, algoritmos de treinamento e princípios básicos. Evidencia-se como redes neurais podem ser fundamentadas pelo formalismo matemático associado à filtragem adaptativa, em termos da equação de Wiener-Hopf, da predição linear, do algoritmo do gradiente estocástico e da desconvolução cega. Simultaneamente, demonstra-se como filtros adaptativos podem ser relacionados aos princípios básicos de redes neurais, por exemplo, ao sistema nervoso vertebrado, à sinapse de Hebb, ao processamento paralelo distribuído, à auto-organização e à psicologia cognitiva.

Apresenta-se uma metodologia de trabalho para o desenvolvimento do processamento neural-adaptativo de sinais e discutem-se alguns resultados já alcançados por esta nova abordagem, que consistem na análise matemática simultânea do processamento paralelo distribuído intrínseco a uma rede neural Perceptron multi-camadas (linear e parcialmente interconectada) e à cascata de filtros adaptativos transversais. Com base nesta análise, propõe-se uma versão modificada do algoritmo do gradiente estocástico na forma cascata, cujo desempenho é avaliado para a predição linear de um sinal auto-regressivo. Simulações evidenciam que, para este caso, o novo algoritmo é mais rápido e mais independente das condições iniciais que sua versão original.

ABSTRACT

We propose the theoretical foundations of the "Neural Adaptive Signal Processing", an emerging technique which establishes an useful co-operation between the collective properties of neural networks and the solid mathematical theory connected to adaptive filtering. Neural adaptive signal processing enables, at the same time, a deeper mathematical analysis of neural networks basic principles (eg. parallel distributed processing and self-organization) and the efficient application of adaptive filters to more complex tasks (eg. signal processing in the presence of non-Gaussian noise).

The theoretical foundations are based upon several mathematical and conceptual analogies between neural networks and adaptive filtering structures, training algorithms and basic principles. We point out how some adaptive filtering theories and equations (such as Wiener-Hopf equation, linear prediction, stochastic gradient algorithm and blind deconvolution) can be applied as an useful formalism to neural networks mathematical analysis. Conversely, we show how adaptive filters can be related to neural networks basic principles (for exemple, vertebrate nervous system, Hebbian synapsis, parallel distributed processing, self-organization and cognitive psychology).

We present a research methodology for neural adaptive signal processing development and we discuss some results attained by making use of it. We analyse mathematically the parallel distributed processing of two systems: a linear partially-interconnected multi-layer Perceptron and a cascade of transversal adaptive filters. Based upon this analysis, we propose an alternative cascade-form stochastic gradient algorithm, and we evaluate its performance when the cascaded filters are applied to the linear prediction of an auto-regressive signal. For this case, simulations outline that the new algorithm seems to be faster and more independent of system initial conditions than its original counterpart.

ÍNDICE

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Capítulo 8

CAPÍTULO 1

INTRODUÇÃO

As redes neurais correspondem, atualmente, a uma das técnicas mais promissoras em diversos campos, por exemplo, no reconhecimento de padrões, na robótica e na inteligência artificial [1]. As arquiteturas de redes neurais são baseadas em modelos matemáticos do neurônio biológico, sendo que um dos primeiros modelos a ser definido foi o "neurônio formal" de McCulloch & Pitts [2], na década de 40. Entretanto, apenas na década de 60, Widrow e Hoff [3] apresentaram o "Adaline", uma das primeiras aplicações práticas mais conhecidas de redes neurais.

A pesquisa em redes neurais possui, portanto, duas características singulares. É extremamente jovem e multidisciplinar, tanto no sentido da aplicação em diversas áreas, como no sentido de ser fundamentada por conceitos associados a múltiplas ciências (por exemplo, neurofisiologia [4-6], psicologia cognitiva [7], mecânica estatística [8-9] e processos estocásticos [5]). Consequentemente, faz-se necessário um estudo multidisciplinar para a formação de uma base teórica mínima, com o objetivo de compreender e utilizar redes neurais.

Deve-se destacar, também, que existe uma certa disparidade entre os resultados já alcançados pela vertente de pesquisa aplicada de redes neurais (os quais evidenciam a grande potencialidade da nova ferramenta) e a pesquisa básica. De fato, o caráter empírico do projeto e do treinamento destas estruturas pode ser de certa forma justificado pelo fato da "matemática das redes neurais encontrar-se ainda em sua infância" [9].

As técnicas de processamento de sinais (e, dentre estas, a filtragem adaptativa), ao contrário de redes neurais, já são aplicadas industrialmente em diversos campos, como por exemplo os de telecomunicações, geofísica e engenharia biomédica [10]. Deve-se notar que estas técnicas estão fundamentadas pelas teorias de comunicações, do controle automático e pela teoria estatística do sinal [11], desenvolvidas a partir dos trabalhos pioneiros de Shannon [12] e de Wiener [13] na década de 40. Não se deve deixar de mencionar também o filtro de Kalman [14], uma das contribuições mais marcantes ao processamento de sinais, desenvolvido na década de 60 (contemporâneo, portanto, ao Adaline).

Constata-se, assim, que as redes neurais e a filtragem adaptativa foram desenvolvidas de forma independente, porém paralela. O objetivo desta tese é proporcionar uma base teórica que permita a unificação dos dois campos, através do estabelecimento de uma "linguagem comum" a partir de diversas analogias conceituais e matemáticas existentes entre as duas áreas. Tal unificação é aqui denominada como "processamento neural-adaptativo de sinais". A seguir, descreve-se resumidamente o processo de desenvolvimento desta base teórica na presente tese.

Inicialmente, apresenta-se uma revisão dos conceitos básicos e dos principais modelos matemáticos que fundamentam as redes neurais.

No capítulo 2, discutem-se alguns princípios biológicos associados ao sistema nervoso vertebrado, bem como o processamento paralelo distribuído, o princípio da mínima perturbação e as propriedades coletivas emergentes de redes neurais. Analisam-se as principais aplicações destas estruturas, definindo-se o conceito de tarefa cognitiva.

No capítulo 3, analisam-se os principais modelos matemáticos e os algoritmos de treinamento de redes neurais, bem como seus problemas práticos e os conceitos biológicos fundamentais associados a cada modelo. Estudam-se o Perceptron [15], o Adaline [3], o Perceptron multi-camadas [4], o mapa auto-organizativo de Kohonen [5] e a rede função radial de base [16]. Comentam-se também as principais aplicações de redes neurais ao processamento de sinais.

Formado então um corpo de conceitos multidisciplinares e de modelos matemáticos associados às redes neurais, necessários para sua análise e utilização, evidencia-se a existência de uma inter-relação entre a filtragem adaptativa e as redes neurais, a nível de princípios básicos, estruturas e algoritmos de treinamento. Isto é realizado nos capítulos 4 e 5.

No capítulo 4, analisam-se as principais diferenças entre os dois campos em termos do treinamento supervisionado. A seguir, apresentam-se diversas analogias matemáticas e conceituais que podem ser estabelecidas entre a filtragem adaptativa e as redes neurais. Desta forma, pode-se demonstrar como o formalismo matemático que embasa a filtragem adaptativa (o qual envolve a teoria de otimização, a equação de Wiener-Hopf, a predição linear, a filtragem espacial e a desconvolução cega [10,17-18]) permite relacioná-la às redes neurais.

O capítulo 5 realiza o processo inverso ao do capítulo 4. Partindo-se dos princípios básicos do processamento de informação neural (apresentados nos capítulos 2 e 3, e que consistem no sistema nervoso vertebrado [19], na lei de Hebb [20], no processamento paralelo distribuído [4], na auto-organização [5] e na psicologia cognitiva [21]), demonstra-se como tais princípios permitem relacionar as redes neurais aos filtros adaptativos. Conclui-se também que estes últimos podem ser considerados redes neurais rudimentares, pois apresentam formas simples de propriedades coletivas emergentes.

Finalmente, no capítulo 6, define-se o processamento neural-adaptativo de sinais. Para isto, a inter-relação existente entre as redes neurais e a filtragem adaptativa é estabelecida, com base na discussão dos capítulos 4 e 5. Em seguida, discute-se a complementariedade das duas técnicas, o que motiva a proposta de uma cooperação entre estas. Define-se então o princípio básico do processamento neural-adaptativo de sinais, bem como uma metodologia de trabalho baseada na inter-relação anteriormente estabelecida. Comentam-se algumas recentes publicações [22-25] a respeito da cooperação entre as redes neurais e a filtragem adaptativa. Finalmente, apresenta-se alguns resultados já alcançados, que correspondem à análise matemática de alguns aspectos de um Perceptron multi-camadas linear parcialmente interconectado (como a superfície de erro quadrático médio e o processamento paralelo distribuído), e à proposição de uma versão modificada do algoritmo do gradiente estocástico para o treinamento da cascata de filtros adaptativos transversais.

As principais conclusões desta tese são resumidas no capítulo 7, que também apresenta possíveis extensões.

CAPÍTULO 2

REDES NEURAIS: PRINCÍPIOS BÁSICOS

2.1 - HISTÓRICO.

A pesquisa na área de redes neurais iniciou-se há aproximadamente cinquenta anos atrás e pode ser dividida em dois períodos distintos.

Durante 1943-1959, a pesquisa concentrou-se no modelamento do cérebro humano como computador, sendo os neurônios considerados como unidades de processamento elementares. As principais contribuições foram as seguintes:

- McCulloch & Pitts [2] (1943): Análise de processos neurofisiológicos a partir da proposição de um modelo matemático do neurônio (considerado como porta lógica digital), com base em resultados experimentais da neurobiologia e na lógica proposicional.

- Hebb [20] (1949): Proposição da primeira lei neurofisiológica de adaptação sináptica, conhecida por "Lei de Hebb" ou "Sinapse de Hebb". Análise comparativa da neurofisiologia e da neuropsicologia.

- Rosenblatt [15] (1958): Apresentação de um modelo matemático simplificado para o neurônio (o "Perceptron") e de seu algoritmo de treinamento.

A partir de 1960, e até os dias de hoje, duas linhas de pesquisa originaram-se do propósito inicial: a vertente básica (que visa a compreensão do funcionamento do cérebro humano e do sistema nervoso a nível multidisciplinar - teoria de sistemas, neurofisiologia, inteligência artificial, psicologia cognitiva, processos estocásticos, caos -) e a aplicada, a qual utiliza em

parte os resultados alcançados pela vertente anterior com objetivo de desenvolver métodos de treinamento e estruturas de redes neurais para execução de determinadas tarefas. Como principais contribuições, pode-se citar:

- Widrow & Hoff [3] (1960): proposição do Adaline, estrutura associada a um modelo matemático do neurônio, e de seus algoritmos de treinamento (regra delta e algoritmo do gradiente estocástico). Discussão do aprendizado supervisionado (que utiliza um sinal de referência) e da minimização recursiva da função de custo pelo método "steepest descent".

- Hopfield [26] (1982): proposição de uma rede neural com realimentação e análise matemática de suas propriedades sistêmicas especiais, decorrentes da interação simultânea dos diversos neurônios constituintes.

- Powell [16] (1985): proposição de uma estrutura neural treinada por um procedimento não-supervisionado (o qual utiliza funções gaussianas multivariáveis denominadas "funções radiais de base") e, posteriormente, por um procedimento supervisionado.

- Sejnowski & Rosenberg [27] (1986): implementação de um sistema de conversão texto-fala em tempo real utilizando redes neurais, denominado "NETTalk".

- Robinson & Fallside [28] (1988): definição do modelo de rede neural recorrente a tempo discreto, que corresponde a uma versão aprimorada da rede proposta por Hopfield [26].

- Kohonen [5] (1989): definição do mapa auto-organizativo, rede composta por neurônios treinados sem acesso a um sinal de referência, fundamentada pelos conceitos biológicos associados à percepção e à memória.

- Grupo PDP [4] (1989): publicação do "Algoritmo de Treinamento por Retropropagação", utilizado para o aprendizado da rede

Perceptron multi-camadas.

- Williams & Zipser [29] (1989): proposição do "Algoritmo de Retropropagação Truncada no Tempo", para o treinamento de redes recorrentes.

2.2 - DEFINIÇÃO E ASPECTOS GERAIS.

Rede neural é um sistema composto pela interconexão de diversas unidades de processamento elementares (denominadas neurônios ou nós, em geral não-lineares) de acordo com uma determinada arquitetura pré-estabelecida. A operação conjunta de tais unidades confere propriedades especiais à rede neural, o que justifica sua capacidade de execução de tarefas extremamente complexas (multivariáveis e descritas por parâmetros ambíguos) em vários campos: controle de robôs e reconhecimento de padrões [30], conversão texto-fala [27], previsão do tempo e diagnóstico clínico [31]. A rede neural é definida a partir de modelos matemáticos simplificados do neurônio, da percepção e da memória dos animais vertebrados.

O neurônio (fig. 2.1) é uma estrutura que possui uma saída e N entradas, cada uma destas associada a um coeficiente denominado "peso sináptico". O sinal de entrada de cada nó, proveniente do ambiente externo ou de outros neurônios, é denominado "Padrão" ou "Estímulo de Entrada", processado em duas fases. Na primeira, realiza-se uma operação pré-estabelecida (denominada "Regra de Propagação") entre o padrão e o conjunto de pesos sinápticos do neurônio. Em seguida, o resultado desta primeira fase é mapeado por uma função matemática denominada "Função de Ativação", gerando a saída do nó. A saída também é geralmente referenciada como ATIVIDADE NEURAL DE SAÍDA ou ainda estado de ativação do nó.

O conjunto de pesos sinápticos de cada neurônio armazena informação simultânea sobre os vários estímulos de entrada. O

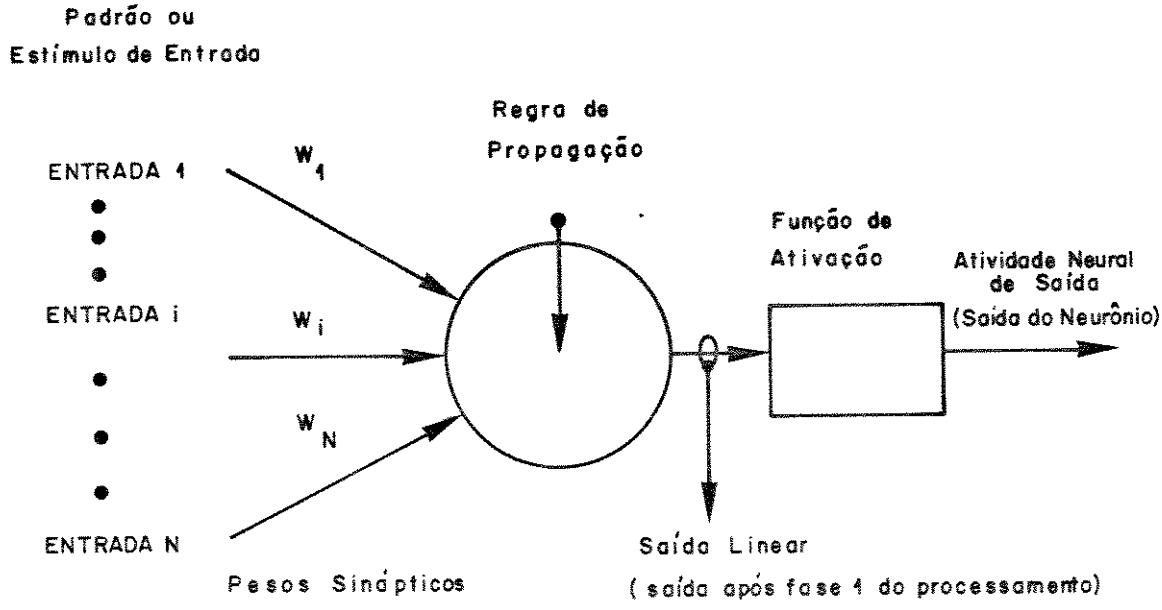


Figura 2.1: Neurônio ou Nó.

agrupamento de todos estes conjuntos do sistema representa, portanto, o conhecimento da rede neural sobre o sinal externo que sensibiliza suas entradas. Este, por sua vez, pode assumir naturezas extremamente genéricas. Por exemplo, o padrão de entrada da rede pode representar tanto o conjunto de variáveis fisiológicas que caracterizam o estado clínico de um paciente, como também fonemas de uma base de dados ou até mesmo "pixels" de uma imagem. Por esta razão, na presente tese, os termos "padrão" e "estímulo" serão considerados diferentes de "sinal" (o qual diz respeito aqui a uma série temporal).

O conjunto de valores de saída de todos os nós constituintes do sistema em um determinado instante de tempo define o estado de ativação da rede, cuja evolução temporal é denominada "Padrão de Atividade Neural". Além disso, a arquitetura do sistema é definida pela quantidade de neurônios, sua organização geométrica e tipo de interconexão.

Sob o ponto de vista tradicional [1,4,31,32], que corresponde às aplicações ao reconhecimento de padrões e à classificação, a operação de redes neurais é realizada em três fases: adaptação, teste e utilização. Supõe-se disponível um conjunto finito de padrões, divididos em dois grupos: estímulos de aprendizado (grupo G1) e de teste (grupo G2).

A primeira fase corresponde ao treinamento da rede neural através de um algoritmo adaptativo, que estabelece modificações no conjunto de pesos sinápticos do sistema. O grupo de estímulos de aprendizado G1, suposto representativo das características estatísticas do universo de padrões a serem processados pela rede, é apresentado ao sistema durante esta fase. Se um sinal de referência é provido externamente, o erro da rede pode ser calculado e o aprendizado é do tipo "supervisionado". Do contrário, diz-se "auto-organizativo". Para o primeiro tipo, a fase de adaptação termina quando o erro (ou uma função deste) assumir valor inferior a um limiar máximo pré-estabelecido. Para o

caso auto-organizativo, fixa-se uma quantidade de iterações para a realização do treinamento.

Na segunda fase, os pesos sinápticos são mantidos constantes e a rede é então testada através da apresentação dos estímulos do grupo G2, o que permite avaliar o desempenho do sistema na execução da tarefa desejada. Caso este não seja satisfatório, a rede é retreinada conforme explicado no parágrafo anterior, e seu desempenho é novamente avaliado.

Repete-se várias vezes o procedimento fase1 (adaptação) - fase2 (teste) até se alcançar o mínimo desempenho exigido pela aplicação. Neste instante, inicia-se a terceira fase de operação da rede neural, que corresponde à sua utilização.

Os algoritmos de treinamento supervisionados podem ser divididos em dois tipos: regras de gradiente e regras de correção de erro. As regras de gradiente alteram os pesos sinápticos do sistema com objetivo de minimizar o erro quadrático médio. A correção do erro, neste caso, é proporcional a uma estimativa do gradiente da função de custo (minimização por "Steepest-Descent" [18]). São aplicados às redes neurais com função de ativação contínua. Como exemplo deste primeiro tipo de algoritmo, pode-se citar o algoritmo do gradiente estocástico [18] e o algoritmo de retropropagação [4].

As regras de correção de erro são aplicadas ao treinamento de redes neurais com função de ativação descontínua, caso em que não é possível deduzir matematicamente uma estimativa do gradiente. Minimizam o erro da saída, cuja correção é proporcional ao próprio erro. Como exemplo, temos o Algoritmo Madaline II [31] ("Adaptação por Tentativa"), onde os pesos sinápticos de um neurônio são alterados somente se for constatada diminuição da amplitude do erro atual, após a aplicação de uma perturbação externa aleatória nos próprios pesos. O processo pode envolver tanto neurônios isolados como em grupos.

Deve-se notar que a dinâmica do aprendizado da rede neural depende basicamente da arquitetura e da função de ativação, como também da escolha e da sistemática de apresentação do grupo de padrões de treinamento G1. Esta técnica de separação dos estímulos externos em grupos de teste e de aprendizado é conhecida por "Método da Validação Cruzada" [1].

Existem quatro funções básicas desempenhadas por uma rede neural, onde se supõe que os padrões anteriormente aprendidos durante o treinamento e em seguida degradados por ruído, sejam apresentados à entrada da rede durante a fase de utilização. Como sistema auto-associativo, deve ser capaz de regenerar o estímulo original. Como sistema associador, deve gerar como saída o padrão associado ao estímulo anteriormente aprendido. Como classificador, deve identificar corretamente a classe do padrão de entrada. A quarta função corresponde ao sistema "Detector de Regularidades", onde a rede objetiva estimar características estatísticas do conjunto de estímulos de treinamento, apresentados com uma determinada probabilidade durante o aprendizado. Em geral, redes neurais que operam como sistema auto-associativo ou detector de regularidades são treinadas de forma auto-organizativa.

2.3 - CONCEITOS BIOLÓGICOS: SISTEMA NERVOSO, NEURÔNIOS E MEMÓRIA.

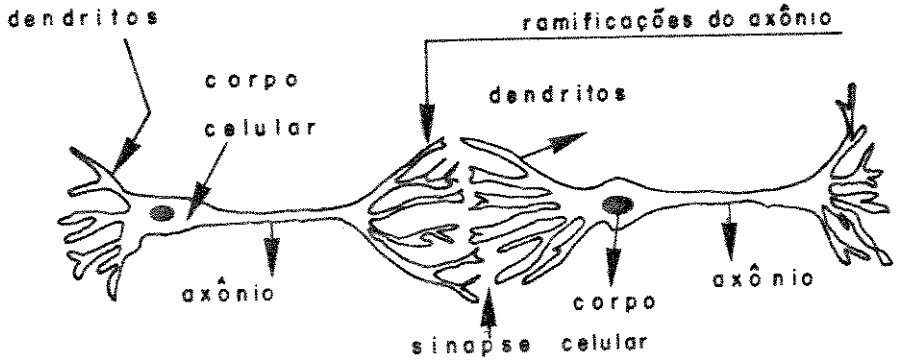
A inclusão de elementos neurofisiológicos (extensão de conceitos biológicos que inspiraram a definição de redes neurais) nos modelos do sistema nervoso analisados pela linha de pesquisa básica de redes neurais mostrou-se necessária para propiciar resultados coerentes com a realidade biológica [4,5]. Portanto, tais elementos não podem ser desconsiderados na análise de redes neurais.

Segundo Kohonen [5], o principal objetivo de qualquer sistema nervoso é "monitorar e controlar as condições de vida do organismo em relação ao meio-ambiente no qual está inserido".

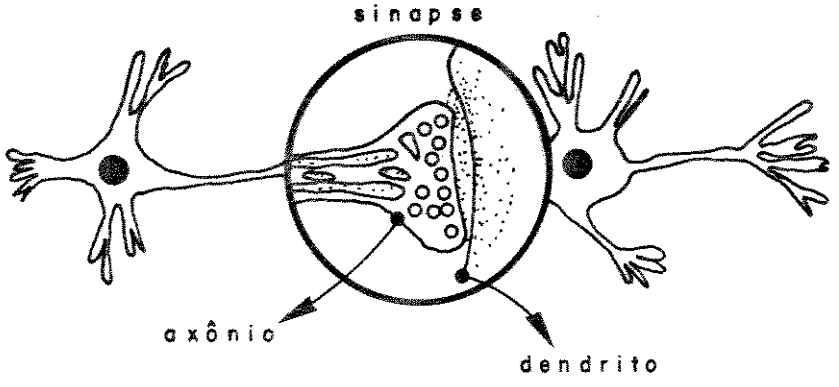
O sistema nervoso dos vertebrados [19] apresenta duas divisões. A primeira delas corresponde ao sistema nervoso SOMÁTICO, dedicado ao controle da ações associadas à relação do organismo com o ambiente externo (por exemplo, percepção). Os elementos estruturais desta divisão estão localizados centralmente (encéfalo e medula espinhal) e perifericamente (nervos e receptores sensoriais). A segunda divisão corresponde ao sistema nervoso VISCERAL, responsável pelo controle das variáveis comprometidas com a manutenção da constância do meio interno (por exemplo, temperatura, pressão arterial, batimentos cardíacos).

Convém ressaltar que o processamento principal de informação pelo sistema nervoso vertebrado ocorre no sistema somático central, notavelmente nos núcleos (agrupamentos de corpos celulares de neurônios) e nos córtices (correspondentes a agrupamentos de corpos celulares neuronais em camadas), sendo que estes últimos podem ser do tipo cerebelar ou cerebral. O cortex cerebral, parte mais externa do encéfalo, está diretamente envolvido neste processo de controle e na percepção consciente. O sistema nervoso somático recebe informações através de impulsos nervosos transmitidos pelas fibras nervosas (axônios), que são processadas para desencadear ações de comando e retransmitidas por estas mesmas vias a todas as partes do organismo.

A fig. 2.2(a) apresenta esquematicamente uma célula nervosa ou neurônio. É constituída pelo corpo celular (citoplasma e núcleo do neurônio), por dendritos (expansões citoplasmáticas curtas que estendem a partir do corpo celular) e pelo axônio (neurofibrilas envoltas por duas membranas: a celular - mais interna - e a bainha de mielina, substância lipídica que funciona como isolante elétrico). O comprimento do axônio varia de fração de milímetros a mais de um metro. Um nervo é formado por centenas ou milhares de axônios, cada um deles associado a um neurônio diferente. Num



(a)



(b)

Figura 2.2 - Neurônio Biológico [19].

(a) - Estrutura;

(b) - Liberação de Substâncias Mediadoras da Sinapse.

nervo não se encontram corpos celulares, que estão localizados no encéfalo ou na medula.

O sistema nervoso é formado por uma complexa cadeia de neurônios interconectados. Denomina-se sinapse a conexão existente entre duas células nervosas, que permite a passagem do impulso nervoso. Quando este atinge a extremidade do axônio, induz a liberação de um mediador químico (por exemplo, a acetilcolina), o qual se difunde pelo espaço existente entre a membrana do axônio e a membrana do neurônio pós-sináptico (vide fig. 2.2(b)). Quando a molécula do neurotransmissor (ou mediador químico) se liga a seu receptor, altera-se a polarização da membrana do neurônio pós-sináptico, permitindo então a propagação do impulso nervoso. Portanto, a intensidade da sinapse pode ser regulada pela quantidade do mediador químico liberada.

O neurônio gera trem de impulsos com frequência proporcional àquela do sinal sensorial de entrada, ou seja, sob excitação contínua, oscila com frequências maiores para entradas excitatórias e menores para entradas inibitórias. A geração de sinal de saída ocorre somente se a intensidade de despolarização da membrana dos dendritos for superior a um limiar mínimo. A partir deste limite, a amplitude do impulso de saída é proporcional à intensidade da excitação externa, até que um limiar de saturação seja atingido. Após a geração de cada impulso, o neurônio fica insensível a sinais externos durante alguns milissegundos.

Alterações na relação de entrada-saída de neurônios ocorrem através de mudanças estruturais ou funcionais. O primeiro caso, observado predominantemente em organismos jovens, corresponde ao desenvolvimento de terminações nervosas adicionais. O segundo caso ocorre principalmente em adultos e corresponde a alterações no mecanismo bioquímico que regula a sinapse.

De fato, a adaptatividade da sinapse justifica a eficiência do sistema nervoso frente a situações críticas como crescimento do

organismo, alterações abruptas do meio-ambiente e recuperação de lesões. Hebb [20] estabeleceu a primeira lei fisiológica para explicar a adaptatividade da sinapse, resumida logo a seguir.

LEI DE HEBB (ou SINAPSE DE HEBB)

"Sejam dois neurônios vizinhos A e B. Se um estímulo de elevada magnitude, proveniente do axônio de A, induzir atividade neural intensa em B, e se este processo persistir repetidamente, a sinapse A-B será cada vez mais eficiente".

A eficiência significa que a amplitude da excitação proveniente de A influencia cada vez mais a atividade de saída em B, relativamente aos demais estímulos externos que excitam B.

Embora Hebb não houvesse especificado o processo biológico de alteração da eficiência sináptica, nem os mecanismos por ela responsáveis, a comprovação experimental da validade desta lei ocorreu 30 anos após sua publicação [5]. Portanto, a lei de Hebb constitui elemento fundamental para a análise da adaptação sináptica.

A organização espacial da atividade neural do encéfalo (e, particularmente, do cortex), estudada há mais de cem anos, pode ser detalhadamente especificada [33]. A observação de distúrbios neurológicos após a lesão de determinadas áreas do cortex (acidentais ou induzidas pela aplicação de correntes elétricas de pequena amplitude e curta duração) sugerem que a atividade neural de regiões específicas do encéfalo estão associadas a determinadas operações de percepção. Cada uma destas regiões é denominada "Mapa Cerebral de Características" (ou "Área de Projeção Primária", em termos da neurofisiologia [19]) para um sinal sensorial específico, e podem ser estabelecidas por técnicas de processamento de imagens biomédicas. Como exemplo, tem-se mapas visuais, tonotópicos (auditivos) e motores (movimentos musculares) [33].

Deve-se notar que a atividade neural, em uma área do cortex delimitada por um mapa cerebral de características, também está organizada espacialmente. Características particulares do sinal sensorial processado induzem ativação em grupos específicos de neurônios, os quais constituem os denominados "Campos de recepção". Por exemplo, nos mapas visuais existem campos de recepção para cores e para linhas [34], enquanto que nos mapas tonotópicos cada campo está associado a um "pitch" do sinal auditivo [35].

A codificação da informação no cérebro dos animais vertebrados é organizada, portanto, de forma espacial, através dos mapas de características e dos campos de recepção.

Uma das funções mais importantes do sistema nervoso humano é a memória, que envolve o armazenamento de dados para posterior utilização. Segundo Guillaume [36]: "Sem memória não existiria vida psíquica propriamente dita; o ser não adquiriria hábitos nem sentimentos. Não teria imaginação, nem representação, nem vida interior, pois esses termos designam modalidades da memória; não teria vontade, pois não poderia pensar nos atos antes de executá-los". Portanto, a memória biológica representa a base fundamental dos processos psicológicos (e dentre eles, o conhecimento e o aprendizado).

A memória caracteriza-se por participar de quase todas funções cerebrais de forma espacialmente distribuída (sem gerar atividade neural localizada). Entretanto, o armazenamento de informação ocorre através da codificação espacial pelos mapas de características e pelos campos de recepção. Como consequência, a memória realiza processamento associativo, isto é, a recuperação dos dados por ela guardados ocorre pela identificação do padrão de atividade neural que possuir máxima correlação com o estímulo externo. A informação associada a tal padrão é biunivocamente correspondente ao dado armazenado. (Notar que este mecanismo explica a eficiência da memória mesmo em condições ambientais desfavoráveis à percepção, como por exemplo a visão noturna).

A memória também armazena, de certa forma, a influência do grupo social no qual o indivíduo está inserido no funcionamento de seu sistema nervoso.

2.4 - PRINCÍPIOS FUNDAMENTAIS: O PROCESSAMENTO PARALELO DISTRIBUÍDO, AS PROPRIEDADES COLETIVAS EMERGENTES E O PRINCÍPIO DA MÍNIMA PERTURBAÇÃO.

Independente da estrutura considerada, uma rede neural corresponde a um conjunto de várias unidades de processamento (matematicamente simples), interligadas conforme determinada arquitetura, e que se influenciam mutuamente em resposta a um estímulo externo. O processamento do padrão de entrada ocorre através da interação simultânea dos nós constituintes, que enviam e recebem sinais excitatórios ou inibitórios entre si. Portanto, o tratamento de informação por redes neurais pode ser descrito pelo princípio do "Processamento Paralelo Distribuído" [4], que afirma que o comportamento da rede resulta do complexo mecanismo de interações existentes entre todas as unidades que a compõem, o qual depende da arquitetura e da sistemática de treinamento.

O termo "distribuído" está associado a três características fundamentais do processamento de informação neural:

- A tarefa realizada pelo sistema é igualmente repartida entre seus diversos neurônios (os quais atuam no padrão de entrada de forma "paralela", ou seja, simultânea), não existindo unidades que possam desempenhar papel mais importante que outras.

- A rede produz uma "Representação Interna Distribuída" dos padrões de aprendizado. Isto significa que cada estímulo de treinamento associa-se biunivocamente a um padrão de atividade neural. Desta forma, cada neurônio está envolvido na

caracterização de todos os estímulos de aprendizado (os quais, em uma representação sistêmica local, estariam armazenados em unidades de processamento específicas).

- O conhecimento da rede neural é distribuído, expresso matematicamente pelo conjunto de todos os pesos sinápticos. O aprendizado corresponde a encontrar valores dos pesos de cada nó de modo que padrões de atividade neural convenientes sejam produzidos em função das condições de entrada.

Como consequência imediata deste tipo de processamento de informação, redes neurais apresentam "Propriedades Coletivas Emergentes" [26,4], que podem ser definidas como capacidades sistêmicas, oriundas da interação microestrutural existente entre os neurônios, as quais justificam a eficiência da rede na execução de tarefas extremamente complexas (multivariáveis e descritas por parâmetros ambíguos). Tais propriedades, que representam a diferença fundamental entre redes neurais e computadores, serão estudadas detalhadamente na próxima seção.

O principal paradoxo no treinamento de redes neurais consiste em ensinar ao sistema novos estímulos de entrada, garantindo a manutenção do conhecimento anterior. Procura-se alterar o menor número possível de pesos sinápticos, mantendo-se a amplitude da modificação reduzida (o que diminui a velocidade de convergência do treinamento). Existe, portanto, um compromisso a ser satisfeito entre magnitude da adaptação e quantidade de parâmetros alterados versus velocidade de treinamento, de modo a permitir a modificação mais suave possível da estrutura de conhecimento da rede.

Widrow e Lehr [31] denominam este compromisso de "Princípio da Mínima Perturbação", que consiste em adaptar para reduzir o erro de treinamento, com mínima perturbação às respostas já aprendidas. Tal princípio representa fundamento básico de todas as regras de treinamento. Por exemplo, para o algoritmo do gradiente

estocástico [31] (análogo ao aplicado para a filtragem adaptativa), manifesta-se matematicamente de forma explícita no passo de adaptação μ, que controla a estabilidade e a velocidade de convergência do algoritmo. A limitação tradicionalmente imposta à magnitude de μ (necessária para garantir a estabilidade da regra de treinamento) significa que a modificação sináptica não pode ser arbitrariamente elevada.

2.5 - MOTIVAÇÃO PARA O USO DE REDES NEURAIS: TAREFAS COGNITIVAS E PROPRIEDADES COLETIVAS EMERGENTES

Tarefas cognitivas correspondem a aplicações ou processos matematicamente mal definidos, as quais exigem geração de decisões em tempo real. O sistema dedicado a tarefas cognitivas utiliza, em geral, um volume excessivo de informação, composta por dados multivariáveis, simbólicos (de difícil caracterização matemática) e perturbados por ruído. Consequentemente, a decisão é baseada na consideração simultânea de inúmeras condições ambíguas, imperfeitamente especificadas, e cujo grau de importância é variável no decorrer do processamento. Tendo em vista a exigência de tempo real e o tipo de informação tratada pelo sistema, as tarefas cognitivas envolvem volume excessivo de cálculo, o que sugere a necessidade de computação paralela massiva.

A sobrevivência dos animais vertebrados representa um exemplo de tarefa cognitiva. Para desempenhá-la eficientemente, o sistema nervoso realiza um controle multitarefa de processos vitais complexos (metabolismo, imunidade, movimentos musculares, percepção), através de ações resultantes do processamento das condições de sobrevivência do meio (parâmetros físico-químicos, presença de predadores ou alimento) e do estado global do organismo. Deve-se destacar a exigência de decisões em tempo real, o volume de informação envolvido e o fato dos sinais naturais e sensoriais internos serem não-estacionários e perturbados por ruído.

Outros exemplos de tarefas cognitivas são o mecanismo de controle automático de movimentação de robôs, reconhecimento de padrões em meio a ruído (imagens, sons), previsão de tempo e classificação étnica de pessoas.

Um contra-exemplo de tarefa cognitiva poderia ser o cálculo da transformada de Fourier tridimensional de uma imagem. Todos os dados podem ser exatamente caracterizados em termos matemáticos, gerando saídas que representam condições perfeitamente especificadas para um processo de decisão.

Embora não possam calcular a transformada tridimensional a velocidades compatíveis com computadores atuais, as pessoas (e até mesmo insetos ou peixes), são extremamente mais eficientes no reconhecimento de imagens e de sons que os mais modernos sistemas de classificação de padrões [31]. Por exemplo, animais reconhecem facilmente imagens em diversos ambientes (mesmo obstruídas por outros objetos), sob várias condições de luminosidade, de posição e de orientação espacial. Da mesma forma, pessoas são capazes de compreender palavras pronunciadas por diferentes interlocutores, a diversos "pitches" e volumes, mesmo na presença de ruído.

Como então superar a deficiência de arquiteturas computacionais atuais em aplicações cognitivas?

De forma análoga ao sistema nervoso dos vertebrados, o processamento paralelo distribuído de redes neurais propicia propriedades coletivas emergentes, que há trinta anos demonstram ser necessárias para realizar eficientemente tarefas cognitivas. A seguir, as principais propriedades coletivas emergentes são apresentadas e relacionadas às redes neurais e às suas aplicações.

1) Compressão de Informação (ou Abstração)

Capacidade do sistema em armazenar apenas as características coletivas fundamentais do conjunto de dados de entrada, de forma a desprezar informações muito específicas ou irrelevantes. Em termos matemáticos, isto significa que o sistema é capaz de rapidamente caracterizar a "estrutura estatística média" do conjunto de estímulos de entrada (ou seja, estimar parâmetros estatísticos que sejam os mais representativos possíveis do conjunto de estímulos considerado como um todo), por mais complexos e distintos entre si que os padrões externos possam ser.

Esta propriedade justifica a eficiência de redes neurais no processamento de estímulos externos complexos, como por exemplo a classificação de padrões com estrutura estatística desconhecida [32] e a equalização de sinais perturbados por ruído não-linear, não-gaussiano (interferência de cocanal e desvanecimento [37]) ou transmitidos por canais de comunicação de fase não-mínima [38]. Nestes tipos de aplicações, as técnicas tradicionais de classificação de padrões e de processamento de sinais apresentam baixo desempenho.

A propriedade de abstração também justifica a eficiência da aplicação de redes neurais à compressão de sinais de vídeo [39] e à quantização vetorial [40].

Em termos da Inteligência Artificial, os padrões de atividade neural da rede podem ser identificados com símbolos. Estes, por sua vez, estão associados a estruturas conceituais ou a grandezas físicas, os quais podem abranger múltiplos sentidos e idéias. Portanto, as redes neurais podem ser usadas eficientemente na expansão de estruturas conceituais básicas para níveis mais elevados, bem como para a compactação de conceitos genéricos e complexos em um conjunto de idéias sucintas. Esta representação flexível de premissas básicas e de interrelações entre conceitos propicia um estabelecimento rápido e preciso de regras relacionais, operação básica para a Inteligência Artificial [4].

2) Recuperação Espontânea do Sistema

Capacidade de auto-regeneração do sistema após uma danificação estrutural. Um exemplo típico é a recuperação de organismos animais que sofreram lesões. Tal comportamento é também constatado na simulação computacional das redes neurais. Assim, sob determinados limites, a eliminação de alguns neurônios e a alteração aleatória de seus pesos sinápticos não implicam necessariamente numa sensível degradação do desempenho, em regime permanente. Isto deve ser comparado, por outro lado, aos efeitos causados pela danificação de circuitos aritméticos (ou de memória) na performance de um microcomputador.

Esta propriedade representa uma clara manifestação do processamento paralelo distribuído, onde a ausência de uma unidade é compensada pelas demais, pois todas desempenham o mesmo papel no processamento do estímulo de entrada.

3) Processamento Associativo (ou Memorização Endereçada por Conteúdo)

Capacidade de um sistema de armazenamento de dados (memória) recuperar um item (informação por ele armazenada), correspondente a um determinado estímulo externo, a partir de dados de entrada que referenciem indiretamente este estímulo.

As memórias seriais computacionais, que guardam cada item em um endereço físico específico e pressupõem conhecimento deste endereço para a recuperação do item, realizam processamento associativo de forma ineficiente. É necessária uma varredura extensiva e demorada de toda a memória, pois os dados de entrada não representam diretamente um estímulo por hipótese, o que impossibilita conhecer o endereço do item. Além disso, não conseguem gerar respostas corretas para informações de entrada errôneas.

As memórias neurais realizam um processamento associativo em tempo real, através da geração de padrões de atividade neural coerentes a partir dos dados de entrada. Além disso, são mais econômicas (por exemplo: a memória óptica associativa para armazenamento de imagens [5], de tecnologia fotônica) devido à possibilidade de representação simbólica e ao fato do armazenamento de novos itens exigir apenas alterações sinápticas, e não uma alocação de "hardware" adicional ou pré-processamento da informação (por exemplo, a compressão de imagens).

4) Processamento Simbólico

Capacidade de processamento de informação de entrada multivariável, composta por dados numéricos e simbólicos (por exemplo, alguns parâmetros meteorológicos, atributos físicos de pessoas, etc). As redes neurais não impõem restrições sobre a natureza do padrão de entrada, sendo capazes de executar determinadas tarefas mais eficientemente que computadores. Alguns exemplos típicos de tarefas que envolvem dados simbólicos são os seguintes: modelamento do processo psicobiológico de reconhecimento de palavras e de objetos [4]; estabelecimento rápido e eficiente de relações semânticas, hierárquicas ou de outras categorias especificadas pela Inteligência Artificial [4].

5) Representações Invariantes (ou Codificação Esparsa)

Uma característica importante do sistema visual animal é a grande capacidade de reconhecimento de objetos independentemente das condições de luminosidade, do ambiente e da posição destes em relação ao sistema perceptivo. O que possibilita tal eficiência é a codificação esparsa, que significa que um conjunto de características físicas do objeto é armazenado na atividade neural de um grupo específico de nós, sendo que o efeito propiciado por uma unidade representa a média estatística do efeito de suas

vizinhas. Assim, para um certo posicionamento do objeto em relação ao sistema visual, pelo menos um neurônio do grupo é sempre ativado, desencadeando o padrão de atividade correspondente ao objeto. Este mecanismo explica a eficiência de redes neurais aplicadas à visão computacional.

6) Generalização (ou Extrapolação, Interpolação, Familiaridade)

Capacidade de um sistema com parâmetros variantes no tempo de processar corretamente dados de entrada diferentes daqueles de treinamento, quando seus parâmetros são mantidos constantes. Esta propriedade permite à rede neural (em regime permanente) reconstruir estímulo de treinamento, tendo por entrada uma versão degradada deste, por exemplo no caso de um padrão corrompido por ruído, de um conjunto limitado ou desordenado de amostras do estímulo e de imagens rotacionadas ou translacionadas. No contexto de reconhecimento de padrões, a generalização é definida como a capacidade de uma rede neural classificar corretamente padrões diferentes daqueles de aprendizado (devido a ruído ou a não-estacionariedades), quando os valores de seus pesos sinápticos são fixados.

A fig. 2.3 apresenta um exemplo de aplicação desta propriedade. Kohonen [5] treinou uma rede neural com a imagem da fig. 2.3(a). Parte da imagem original (fig. 2.3(b)) foi posteriormente apresentada à rede já treinada, originando uma saída (fig. 2.3(c)) idêntica ao padrão de aprendizado. As figuras 2.4(a) e 2.4(b) [5] apresentam exemplos de supressão de ruído e de interpolação, respectivamente, obtidas pela mesma rede.

7) Codificação Espacial

Corresponde exatamente ao conceito de organização espacial da atividade neural do cortex, discutida na seção 2.3. Esta

Figura 2.3: Demonstração da Propriedade Coletiva de Generalização aplicada ao Processamento Digital de Imagens [5].

(a) - Imagem aprendida pela rede neural durante o treinamento;

(b) - Imagem de entrada incompleta apresentada à rede;

(c) - Imagem de saída da rede.

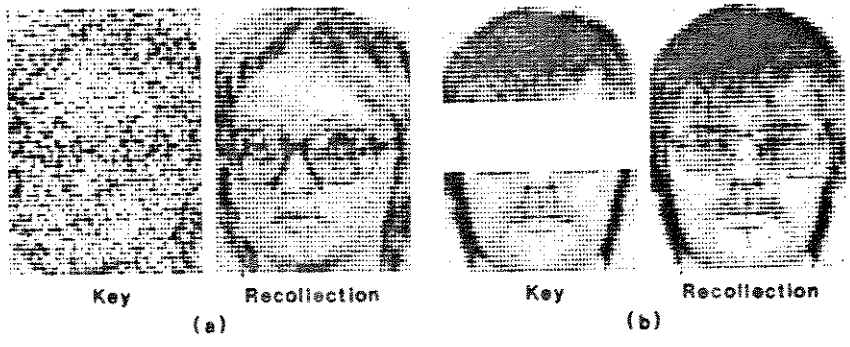


Figura 2.4: Demonstração da Propriedade Coletiva de Generalização aplicada ao Processamento Digital de Imagens [5].

(a) - Extrapolação;

(b) - Interpolação.

propriedade é exclusiva da rede neural mapa auto-organizativo de Kohonen [5], o qual pode ser considerado um mapa cerebral de características simplificado. Através de treinamento adequado, demonstra-se [5] que o conjunto dos pesos sinápticos converge para uma função da densidade de probabilidade dos estímulos de aprendizado.

Deve-se notar que as propriedades (1)-(4) podem ser consideradas "propriedades coletivas emergentes primárias", pois representam consequências imediatas do processamento paralelo distribuído. As propriedades seguintes podem ser consideradas "secundárias", visto que constituem extensões das quatro anteriores. Além disso, (5),(6) e (7) são extremamente dependentes da sistemática de treinamento.

2.6 - APLICAÇÕES TÍPICAS DE REDES NEURAIS E COMPARAÇÃO COM COMPUTADORES.

A partir da análise das propriedades coletivas emergentes, concluiu-se que as redes neurais são especializadas na execução de tarefas cognitivas, presentes principalmente nas áreas de controle multivariável, robótica, inteligência artificial, reconhecimento e classificação de padrões, bem como processamento de sinais (por exemplo: voz e vídeo). Além disso, tem-se observado a possibilidade de modelamento matemático de processos psicobiológicos através de redes neurais, tais como movimentos musculares [41] e percepção auditiva [40], que apresentaram resultados coerentes com a biofísica.

Podem-se destacar duas implementações de redes neurais que marcaram profundamente a comunidade científica internacional: o sistema de conversão texto-fala NETTalk [27] (a ser analisado na seção 3.7) e o robô "Darwin" [30].

O Darwin é composto por um olho (sensor visual) e um braço multijuntas, que contém vários sensores de toque e de velocidade. Movimenta-se em um ambiente bidimensional. Seu controle é realizado por uma rede neural de 50 nós. Este robô possui comportamento autônomo, pois desenvolve coordenação sensomotora através da interação com o próprio meio-ambiente. É capaz de acompanhar movimentos de objetos com o olho; carregar e classificar objetos a partir de sensações visuais e táteis; movimentar-se por entre obstáculos de diversos tamanhos, formas e orientações, espalhados aleatoriamente no meio-ambiente.

Estas capacidades e a habilidade do Darwin de geração de ações corretas quando submetidos a múltiplos estímulos externos, graças ao controle neural, são inovadoras na área de robótica. Tradicionalmente [30], robôs são pré-programáveis e as decisões são fruto do processamento de diversas regras relacionais (ineficazes para um ambiente desconhecido ou variante no tempo), baseadas em informações obtidas pela classificação de padrões, a qual pressupõe conhecimento matemático prévio sobre o ambiente e seus objetos. De fato, a implementação de outro Darwin com base nestas técnicas, evidenciou a lentidão de resposta do autômato comparativamente à metodologia neural [30].

Deve-se mecionar ainda que as simulações envolvendo a rede neural controladora do Darwin (considerada como um "sistema nervoso rudimentar"), o meio-ambiente e suas interações mútuas permitiram não somente a análise do comportamento do autômato, com também o estudo da aquisição de capacidades sensoriais e motoras por animais vertebrados, apresentando resultados coerentes com as teorias biofísicas modernas [30].

Enquanto os computadores processam dados serialmente através de circuitos digitais extremamente mais rápidos que neurônios, as

redes neurais tratam padrões de entrada ruidosos através de unidades adaptativas, que se utilizam do paralelismo distribuído (mais massivo ainda que o conceito de "computação concorrente paralela" [5]). Além disso, as redes neurais apresentam as propriedades coletivas emergentes que a capacitam a interagir com o mundo real e a manipular estímulos externos complexos, enquanto os computadores apresentam apenas formas rudimentares de processamento simbólico e de codificação espacial. Por outro lado, visto que os computadores são especializados na resolução de problemas matemáticos, geram quase sempre resultados ótimos do ponto de vista teórico, ao passo que as redes neurais conduzem em geral a resultados sub-ótimos, tendo em vista as características das tarefas cognitivas.

2.7 - CONCLUSÃO.

Revisaram-se neste capítulo os principais conceitos associados às redes neurais, estruturas definidas a partir de modelos matemáticos simplificados do sistema nervoso de animais vertebrados. Tais estruturas realizam eficientemente tarefas cognitivas, aplicações caracterizadas pelo processamento de um volume excessivo de dados de entrada (multivariáveis e simbólicos, quase sempre perturbados por ruído e de de difícil caracterização matemática) e, consequentemente, pela grande quantidade de cálculo envolvida para resposta em tempo real e pela geração de resultados sub-ótimos sob o ponto de vista teórico. Como exemplos de tarefas cognitivas, pode-se citar a sobrevivência biológica, o reconhecimento de padrões e a robótica.

Portanto, rede neurais são especializadas na interação com o mundo real, graças às suas propriedades coletivas emergentes (associadas ao processamento paralelo distribuído e que dependem

da sistemática de aprendizado, bem como da arquitetura da rede), dentre as quais devem ser destacadas a compressão de informação (capacidade de rápida estimação da "estrutura estatística média" do complexo conjunto de dados de entrada) e a generalização. A representação do estímulo externo ocorre através de padrões de atividade neural e o aprendizado de novos estímulos deve induzir mínima perturbação na estrutura de conhecimento anteriormente formada.

Deve-se ainda ressaltar que o cortex cerebral, centro de processamento de informação do sistema nervoso dos animais vertebrados (que realiza notavelmente controle multivariável e percepção), pode ser considerado como uma rede neural extremamente complexa. Além disso, a adaptatividade da sinapse bioquímica, que pode ser analisada pela lei de Hebb, representa o princípio básico que fundamenta o aprendizado das redes neurais, bem como seu caráter de sistema com parâmetros variantes no tempo.

Com base nos princípios fundamentais apresentados neste capítulo, analisam-se a seguir os principais modelos matemáticos de redes neurais.

CAPÍTULO 3

REDES NEURAIS: ESTRUTURAS, ALGORITMOS E APLICAÇÕES

Neste capítulo, os principais modelos matemáticos e algoritmos de treinamento de neurônios (Perceptron, Adaline, Unidade Adaptativa Básica de Memória ou Neurônio de Kohonen [5]) e de redes neurais (Perceptron multi-camadas, mapa auto-organizativo de Kohonen, função radial de base) são apresentados. A capacidade de classificação, os principais problemas práticos e os conceitos biológicos fundamentais de cada modelo são discutidos e comparados. Comentam-se as principais aplicações de redes neurais ao processamento de sinais.

3.1 - MODELO MATEMÁTICO SIMPLIFICADO DO NEURÔNIO.

A fig. 3.1 representa o modelo matemático do neurônio conceitual, mostrado na fig. 2.1. O sistema é regido pelas seguintes equações:

$$X_i^T = [x_{i1}\ x_{i2}\ \ldots\ x_{iN}] \qquad (3.1a)$$

$$W_i^T = [w_{i1}\ w_{i2}\ \ldots w_{iN}] \qquad (3.1b)$$

$$s_i = \sum_{j=1}^{N} w_{ij} \cdot x_{ij} = W_i^T \cdot X_i \qquad (3.2)$$

$$y_i = f_i(s_i - \theta_i) \qquad (3.3)$$

Onde (especificam-se entre parêntesis os respectivos elementos correspondentes do modelo biológico do neurônio [5], apresentado na fig. 2.2) :

i = índice representativo do neurônio considerado.

N = quantidade de conexões externas (dendritos) do neurônio

X_i = estímulo de entrada (x_{ij} : amplitude do estímulo externo que atinge o j-ésimo dendrito).

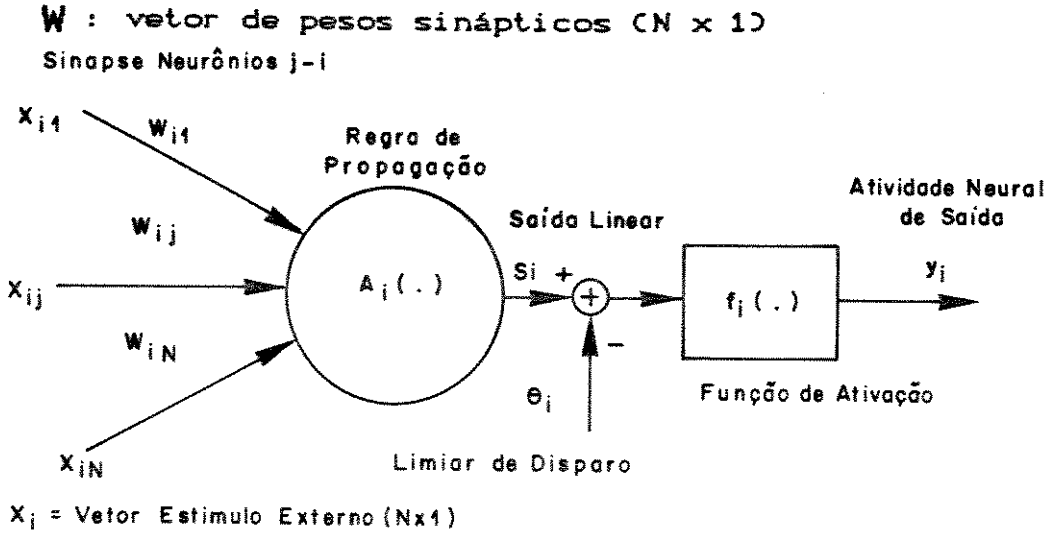


Figura 3.1 : Modelo matemático simplificado do neurônio.

W_i = conjunto de pesos sinápticos.

w_{ij} = peso sináptico que conecta a j-ésima excitação externa x_{ij} ao neurônio i (eficiência - ou intensidade [20]- da sinapse entre o axônio do neurônio j e o dendrito j do neurônio i).

$A(.)$ = produto vetorial como regra de propagação.

s_i = saída linear ou saída após primeira fase de processamento.

θ_i = limiar de disparo do neurônio. Em geral assume valor nulo, exceto para neurônio Adaline [3]. (Mínima amplitude do estímulo externo para induzir atividade neural na célula nervosa i).

f_i = função de ativação do nó.

y_i = saída do neurônio ou estado de ativação, após segunda fase de processamento (amplitude da atividade neural de saída).

As funções de ativação normalmente utilizadas são a função tangente hiperbólica, sigmoidal ou sinal, definidas respectivamente pelas seguintes equações (vide fig. 3.2):

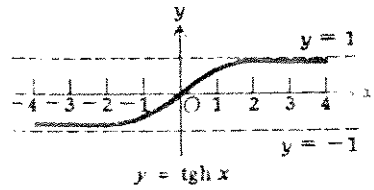
$$tgh(s) = (1 - e^{-\beta s})/(1 + e^{-\beta s}) \qquad (3.4)$$

$$sigm(s) = 1/(1 + e^{-\beta s}) \qquad (3.5)$$

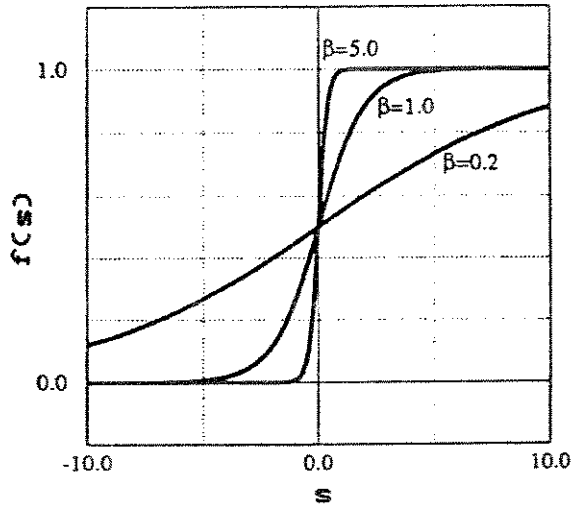
$$sign(s) \cong \lim_{\beta \to \infty} sigm(s) = \begin{cases} 0 \text{ se } s<0 \\ 1 \text{ se } s \geq 0 \end{cases} \qquad (3.6)$$

O parâmetro β determina a suavidade da derivada da função de ativação.

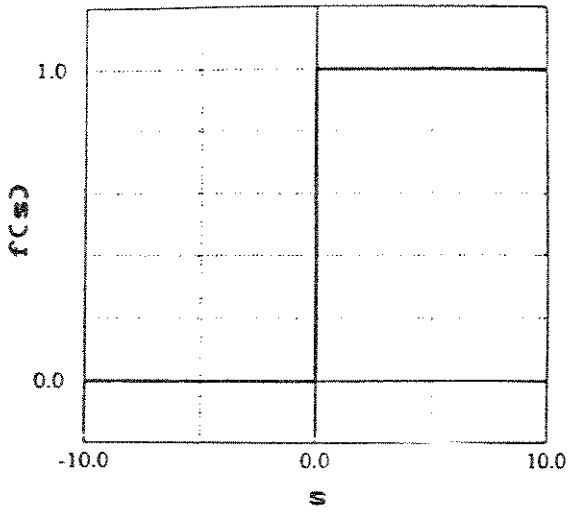
Historicamente, a primeira função de ativação utilizada foi a função sinal, presente nos primeiros modelos matemáticos do neurônio [2,20]. Posteriormente, introduziu-se a função de ativação sigmoidal, uma espécie de aproximação contínua e diferenciável da primeira, características necessárias para a dedução de algoritmos de treinamento baseados no método "steepest descent" [17,18].



(a)



(b)



(c)

Figura 3.2 : Tipos comuns de função de ativação [1].
(a) função tangente hiperbólica;
(a) função sigmoidal;
(a) função sinal.

Este modelo não leva em consideração as modificações temporais dos elementos biológicos f_i, x_{ij}, N e θ_i, associadas à adaptatividade do sistema nervoso. A realimentação estabilizadora do neurônio está sendo considerada pela regra de propagação do tipo produto vetorial [5]. Note-se que a saída linear do nó mede a correlação entre o estímulo externo X_i e o conhecimento armazenado em w_{ij}. Portanto, um neurônio pode ser matematicamente considerado como um filtro correlador não-linear e variante no tempo (devido à adaptação dos parâmetros w_{ij}). Além disso, desde que a função de ativação limita a amplitude da saída linear, o neurônio também pode ser considerado de certa forma como sistema estável.

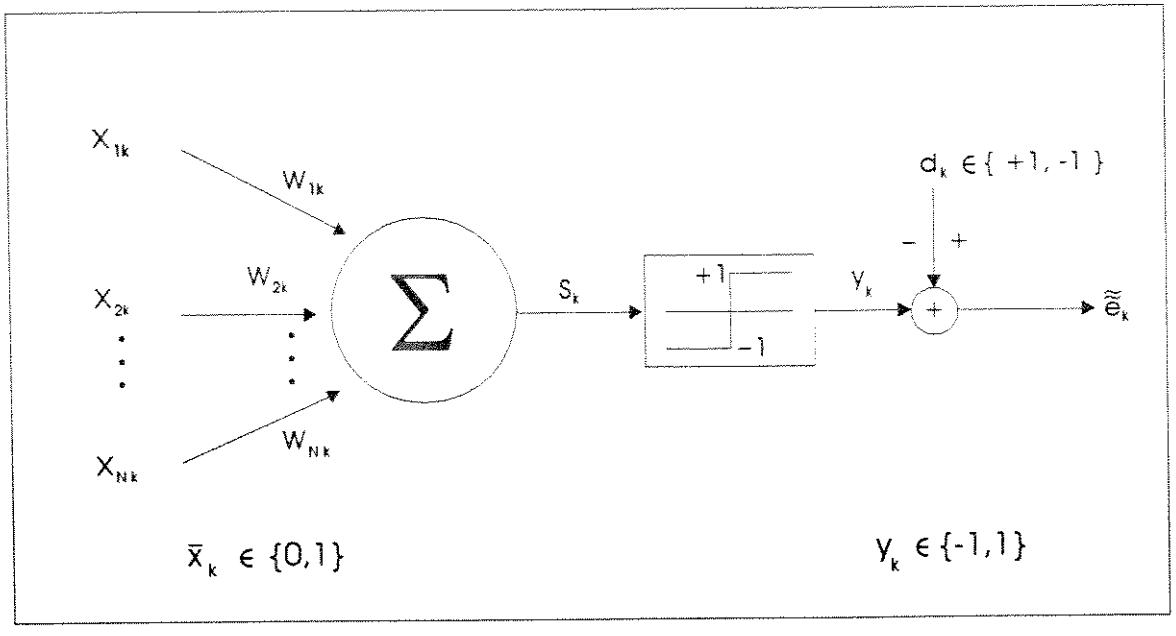
3.2 - PERCEPTRON e ADALINE.

O Perceptron [15] corresponde historicamente a um dos primeiros modelos matemáticos do neurônio, desenvolvido independentemente do Adaline. Embora as duas estruturas sejam idênticas, suas diferenças residem na natureza dos sinais de entrada (binários ou não-quantizados), nos algoritmos de treinamento e no propósito de concepção. Enquanto o Perceptron representa, de certa forma, um modelo matemático (decorrente dos trabalhos de Hebb [20] e de McCulloch & Pitts [2]) que permite a discussão de fenômenos neurofisiológicos e cognitivos, treinado por regra de aprendizado empírica; o Adaline [3] corresponde a uma estrutura com parâmetros variantes no tempo, utilizada para aplicações práticas (tais como a classificação adaptativa de padrões binários), e treinada pelo algoritmo do gradiente estocástico, que é deduzido matematicamente em [3].

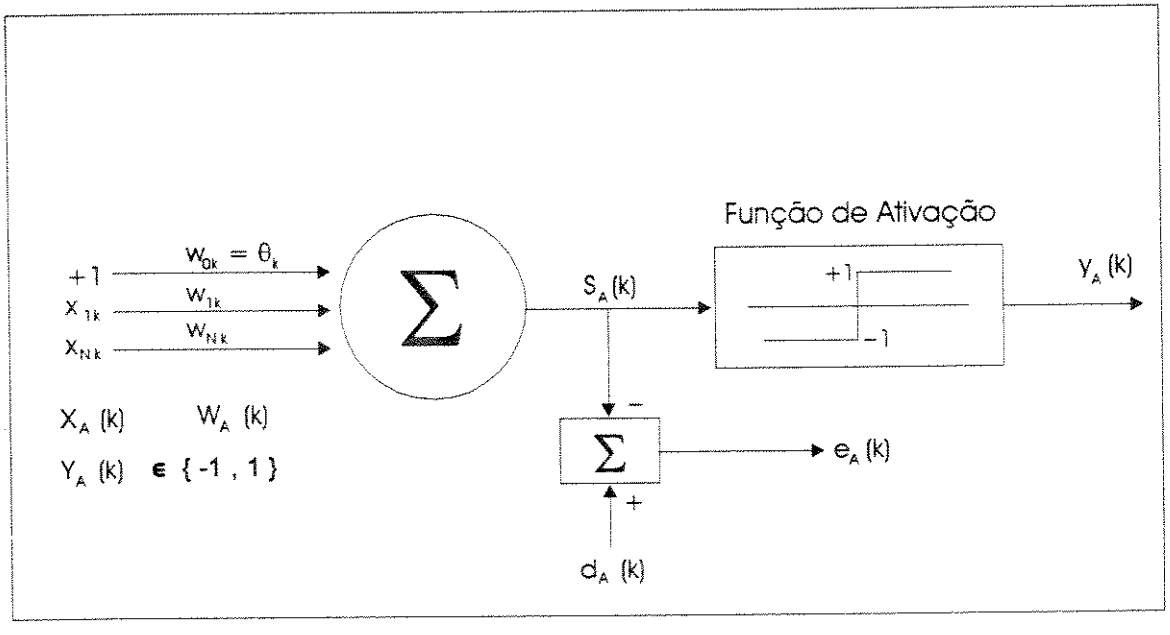
A fig. 3.3(a) apresenta o neurônio Perceptron [15], cuja estrutura é descrita pelas seguintes equações:

$$y_k = 2.\text{sign}(s_k) - 1 \qquad (3.7a)$$

$$\theta_k = 0 \ \forall k \qquad (3.7b)$$



(a)



(b)

Figura 3.3 : Modelos matemáticos do neurônio.
(a) Perceptron [15];
(b) Adaline [3].

O índice k denota a iteração temporal. A entrada X_k é suposta binária (assumindo valores {0,1}), bem como y_k ({-1,1}). O treinamento do Perceptron é realizado, em geral, através de uma regra de aprendizado do tipo correção de erro (sem dedução matemática, portanto), expressa pelas seguintes equações:

$$\tilde{\tilde{e}}_k = d_k - y_k \qquad (3.8a)$$

$$W_{k+1} = W_k + \alpha.(\tilde{\tilde{e}}_k/2).\ X_k \qquad (3.8b)$$

A eq. (3.8b) é denominada "Algoritmo de Rosenblatt" [15]. O sinal d_k reprsenta o sinal de referência (binário por hipótese, podendo assumir apenas um dos valores do conjunto {-1,1}) e α é a constante de adaptação. Utiliza-se em geral α=1, cujo valor não influencia a estabilidade do algoritmo.

A estrutura do neurônio Adaline [3] (abreviação inglesa que significa "Elemento Linear Adaptativo", embora os próprios autores também o referenciem como "Elemento Adaptativo Neural" em [3]), mostrada na fig. 3.3(b), é descrita pelas seguintes equações:

$$s_A(k) = X_A(k)^T.\ W_A(k) \qquad (3.9a)$$

$$X_A(k)^T = [+1\ x_1(k)...x_N(k)] \qquad (3.9b)$$

$$W_A(k)^T = [w_0(k)\ w_1(k)...w_N(k)] \qquad (3.9c)$$

$$y_A(k) = 2.\mathrm{sign}(s_A(k)) - 1 \qquad (3.9d)$$

$$\theta_k = w_0(k) \qquad (3.9e)$$

Nas equações anteriores, o índice A refere-se ao Adaline. A componente $x_o(k)$ é fixada em 1 e o coeficiente $w_0(k)$ controla o limiar de disparo do neurônio, que pode ser ajustado. Supõe-se por hipótese que $y_A(k)$ seja binário, podendo assumir um dos valores do conjunto {+1,-1}, e que $X_A(k)$ seja formado por valores não-quantizados. Em termos estruturais, a única diferença entre o Perceptron e o Adaline reside no limiar de diparo θ_k do nó, o qual é constante para a primeira estrutura.

Para o treinamento do Adaline, utiliza-se em geral o algoritmo do gradiente estocástico (ou LMS) [3], expresso pela equação (3.11):

$$e_A(k) = d_A(k) - s_A(k) \qquad (3.10)$$

$$W_A(k+1) = W_A(k) + \mu . e_A(k) . X_A(k) \qquad (3.11)$$

Onde $d_A(k)$ é o sinal de referência, suposto não-quantizado por hipótese, e μ é o passo de adaptação. O passo μ influencia tanto a velocidade de convergência quanto a estabilidade do LMS. A convergência deste algortimo pode ser demonstrada impondo-se a seguinte condição [18]:

$$0 < \mu < 2/E[X_A(k)^T . X_A(k)] \cong 2/N.\sigma_x^2 \qquad (3.12)$$

Onde σ_x^2 denota a potência do sinal associado aos vetores $X_A(k)$, supostos estatisticamente independentes no tempo. Demonstra-se também que, caso $X_A(k)$ e $d_A(k)$ sejam respectivamente vetores e sinal, ambos de média nula e estacionários no sentido amplo (pelo menos para o conjunto de amostras considerado), então os coeficientes ótimos de Wiener do Adaline podem ser expressos pelo vetor $W_A(k)^*$ [31]:

$$W_A(k)^* = R_x(k)^{-1} . R_{xd}(k) \qquad (3.13a)$$

$$R_x(k) \triangleq E[X_A(k) . X_A(k)^T] \qquad (3.13b)$$

$$R_{xd}(k) \triangleq E[d_A(k) . X_A(k)] \qquad (3.13c)$$

Uma comparação entre o algoritmo de Rosenblatt (eq. (3.8b)) e o do gradiente estocástico (eq. (3.11)) revela que a diferença entre eles reside na natureza do erro considerado, linear para o segundo ($e_A(k)$) e não-linear para o primeiro ($\tilde{e}_k$).

O Adaline também pode ser treinado pela "Regra Delta" [18], uma regra de correção de erro expressa pelas seguintes equações:

$$W_A(k+1) = W_A(k) + \mu(k).e_A(k).X_A(k) \qquad (3.14a)$$

$$\mu(k) = \delta/\|X_A(k)\|^2 \qquad (3.14b)$$

$$0 < \delta < 2 \qquad (3.14c)$$

Onde δ é o passo de adaptação constante. A regra delta possui equação semelhante ao algoritmo do gradiente estocástico e, de fato, demonstra-se [18] que ela também minimiza o erro quadrático médio e que converge sob a condição imposta pela equação (3.14c).

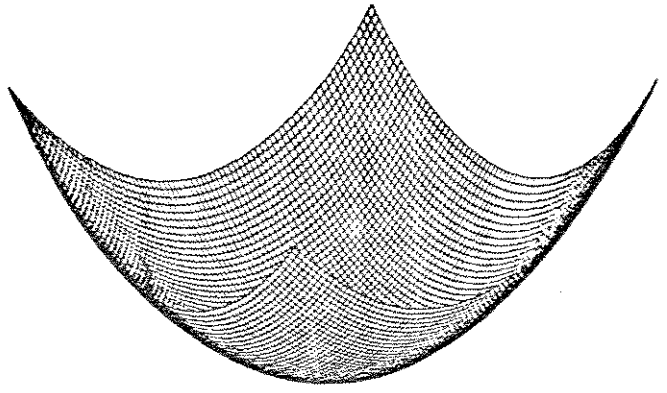
Caso a função de ativação do Adaline seja a sigmóide definida pela eq. (3.5), a aproximação estocástica [31] conduz ao algoritmo do gradiente estocástico sigmoidal (equação (3.16)):

$$\tilde{e}(k) = d(k) - sigm(s(k)) \qquad (3.15a)$$

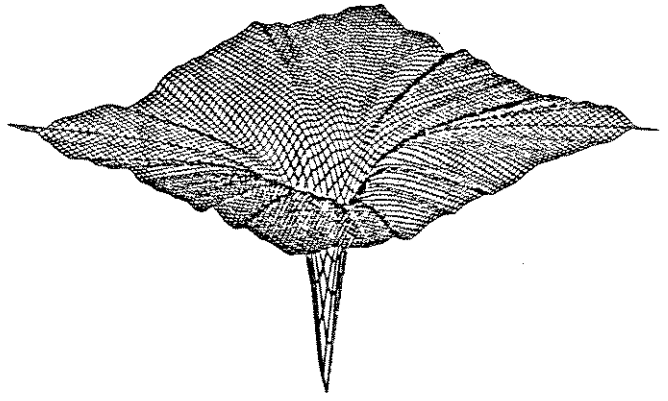
$$sigm'(s_A(k)) \triangleq \frac{d\ sigm(s_A(k))}{d\ s_A(k)} \qquad (3.15b)$$

$$W_A(k+1) = W_A(k) + \mu.\tilde{e}_A(k).sigm'(s_A(k)).X_A(k) \qquad (3.16)$$

As figs. 3.4(a)-(c) apresentam as superfícies de erro quadático médio associadas ao treinamento de um neurônio Adaline de duas entradas com, respectivamente, função de ativação linear (através do algoritmo LMS - eq. (3.11)), sigmoidal (LMS sigmoidal - eq. (3.16)) e sinal (algoritmo de Rosenblatt -eq. (3.8b)) [31]. O neurônio foi treinado para a classificação de diversos tipos de estímulos externos. Enquanto a superfície da fig.3.4(a) possui um único mínimo global, as demais são não-convexas e apresentam mínimos locais, o que está associado à função de ativação não-linear utilizada. Particularmente, constata-se que o algoritmo de Rosenblatt é o mais lento de todos, devido aos vários planos e mínimos locais de sua superfície de erro quadrático (comparativamente às demais), que interferem no procedimento de adaptação.



(a)



(b)

(c)

Figura 3.4 : Comparação das superfícies de erro quadrático médio para um Adaline de duas entradas (N = 2) [31].
(a) função de ativação linear;
(b) função de ativação sigmoidal;
(c) função de ativação sinal.

3.3 PERCEPTRON MULTI-CAMADAS.

A fig. 3.5 apresenta a rede neural Perceptron multi-camadas, definida pela interligação de vários neurônios Perceptron com função de ativação sigmoidal (eq. (3.5)), organizados em camadas. A estrutura da fig. 3.5 é denominada estática ou "feedforward", pois existem conexões entre os neurônios da camada l e todos os outros das camadas l+1 e l-1, não ocorrendo realimentações e nem interligações entre nós situados na mesma camada. Caso exista algum tipo de realimentação da saída do sistema, o Perceptron multi-camadas é denominado "dinâmico".

A rede da fig. 3.5 apresenta três camadas, já que suas entradas não são consideradas como camadas na presente tese. Além disso, por hipótese, há apenas um nó na última camada, cuja função de ativação é linear, conforme será justificado posteriormente. Denominam-se "NEURÔNIOS ESCONDIDOS" todos os nós da rede, exceto o de saída.

O aprendizado do Perceptron multi-camadas é realizado, em geral, através do "Algoritmo de Treinamento por Retropropagação", regra de gradiente sistematizada em [4], que representa uma espécie de generalização do algoritmo LMS sigmoidal da eq. (3.16). Seu objetivo é minimizar o erro quadrático médio associado à saída y(k) da rede neural (que corresponde à saída do único nó da última camada), através de uma aproximação estocástica. A funções de custo J e J(k) envolvidas podem ser expressas por:

$$e(k) = d(k) - y(k) \tag{3.17}$$

$$J \triangleq E[e(k)^2] \tag{3.18a}$$

$$J(k) = e(k)^2 \tag{3.18b}$$

Onde d(k) é o sinal de referência provido externamente.

O algoritmo de retropropagação realiza a adaptação de todos os neurônios da rede a cada iteração. Isto exige o cálculo do gradiente associado a cada nó, o que é realizado de forma indireta

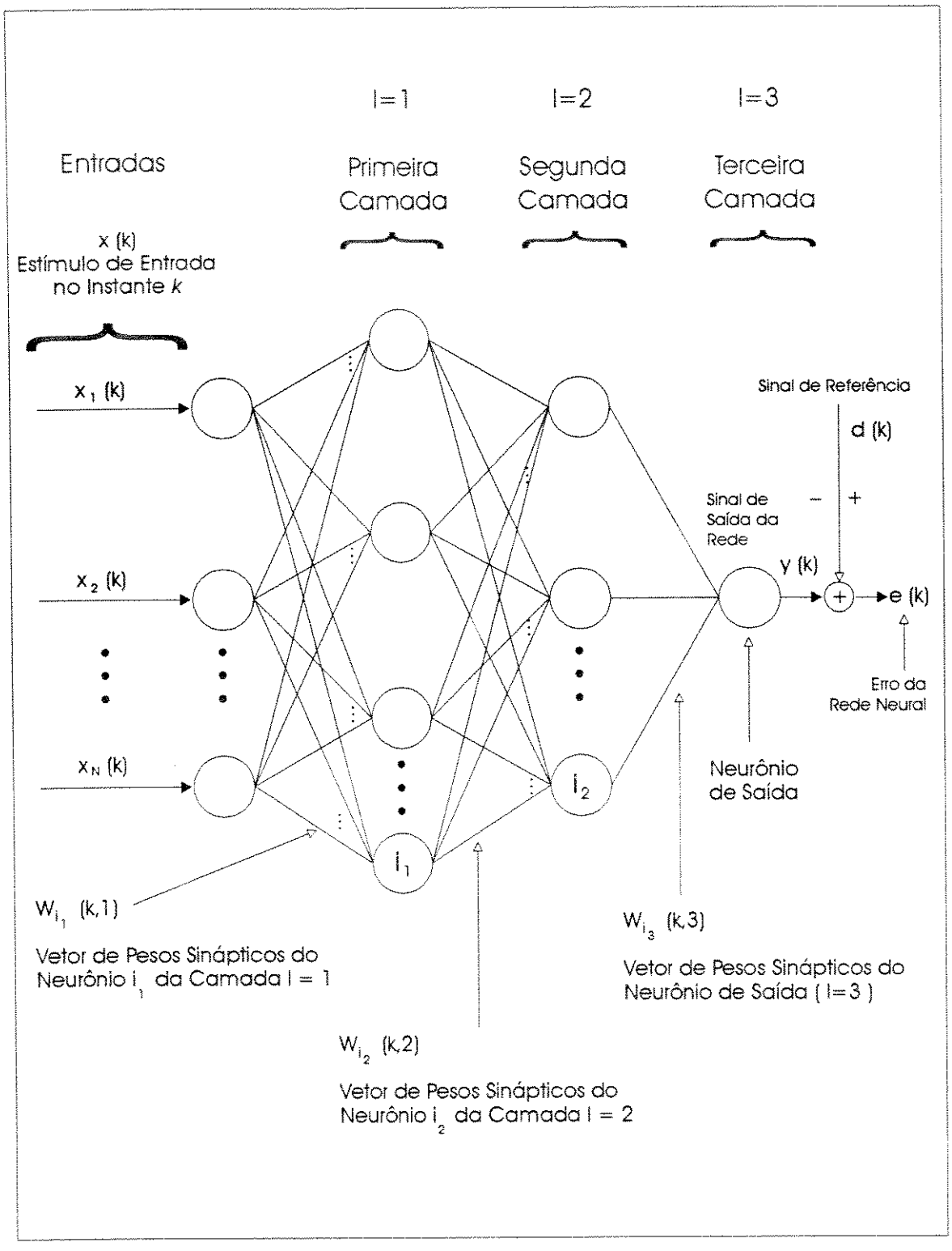


Figura 3.5 : Rede neural Perceptron multi-camadas [4].

através da regra da cadeia [4]:

$$\hat{\nabla} J_i(k,l) = - \frac{\partial J(k)}{\partial W_i(k,l)} = - \frac{\partial e^2(k)}{\partial s_i(k,l)} \cdot \frac{\partial s_i(k,l)}{\partial W_i(k,l)} \qquad (3.19)$$

$$\delta_i(k,l) \triangleq \frac{\partial e^2(k)}{\partial s_i(k,l)} \qquad (3.20a)$$

$$\hat{\nabla} J_i(k,l) = - \delta_i(k,l) . X_i(k,l) \qquad (3.20b)$$

Onde:

$a_i(k,l)$ = grandeza a associada ao neurônio i, situado na camada l de um Perceptron multi-camadas, no instante k. a pode representar qualquer uma das grandezas w_{ij}, f_i, s_i, θ_i e y_i definidas na seção 3.1.

$\hat{\nabla} J_i(k,l)$ = estimativa estocástica do gradiente associado ao neurônio i da camada l, no instante k.

$\delta_i(k,l)$ = sensitividade do erro quadrático relativamente à saída linear do nó i.

Cada iteração do treinamento através do algoritmo de retropropagação ocorre em duas fases, que realizam respectivamente a filtragem do padrão de aprendizado e a adaptação dos pesos sinápticos. Na primeira fase, um padrão $X(k)$ é apresentado à entrada do sistema e é processado pela rede, gerando a saída y(k). Calcula-se então o erro associado ao neurônio de saída. Na segunda fase, este erro é "retropropagado" a partir da última camada em direção à primeira, quando se calcula recursivamente a grandeza $\delta_i(k,l)$ associada a cada neurônio. Em seguida, a adaptação de todos os pesos sinápticos da rede é realizada com base nos valores $\delta_i(k,l)$ calculados (utilizados para a estimação do gradiente do respectivo nó).

As equações para a segunda fase do algoritmo de retropropagação são as seguintes [4]:

Para l = L (L é o *índice associado à última camada da rede)*

$$\delta_l(k,L) = (d(k) - y(k)).f'(s_l(k,L)) = e(k).f'(s_l(k,L)) \qquad (3.21a)$$

Para l = L-1, L-2, ..., 1:

$$\delta_i(k,l) = f'(s_i(k,l)) \cdot \sum_{j=1}^{N_{l+1}} \delta_j(k,l+1).w_{ij}(k) \qquad (3.21b)$$

$$\mathbf{W}_i(k+1,l) = \mathbf{W}_i(k,l) + \mu_p.\delta_i(k,l).\mathbf{X}_i(k,l) \qquad (3.22)$$

Onde f(.) é a função de ativação, N_l representa a quantidade de neurônios da camada l e μ_p é o passo de adaptação do algoritmo. Além disso, define-se:

$$f'(s_i(k,l)) = df(s_i(k,l)) \,/\, ds_i(k,l)$$

A adaptação expressa pela eq. (3.22) utiliza a aproximação da eq. (3.20b). A recursão da eq. (3.21b) foi deduzida em [4].

A quantidade de cálculos é a mesma tanto na primeira quanto na segunda fase de treinamento. Inicialmente, atribuem-se valores aleatórios e pequenos para os pesos sinápticos. O desempenho do algoritmo é influenciado pelo passo de adaptação μ_p (igual para todos os nós da rede em geral) e pela inicialização dos pesos sinápticos (o que está associado aos mínimos locais da superfície de erro quadrático médio). Constata-se experimentalmente que, em geral, o treinamento do Perceptron multi-camadas através do algoritmo de retropropagação e cada vez mais lento à medida que a função de ativação sigmoidal aproxima-se da função sinal ($\beta \to \infty$, eq. (3.6)) [1,31]. Esta constatação experimental está em conformidade com o comportamento dinâmico do Adaline, acima discutido quando da apresentação das curvas de erro quadrático médio das figs. 3.4(b)-(c).

3.4 - CAPACIDADE DE CLASSIFICAÇÃO DAS ESTRUTURAS.

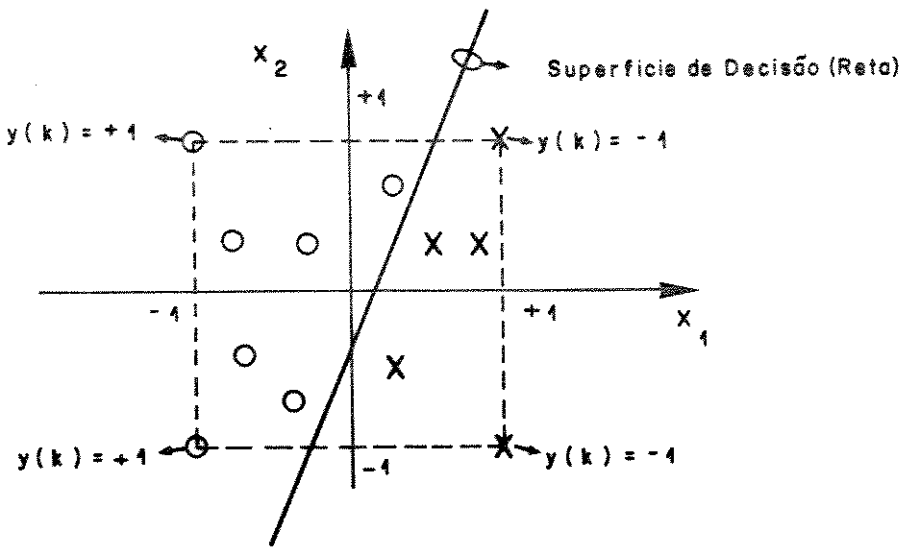
A operação do neurônio Perceptron pode ser analisada de duas formas diferentes. Como porta lógica digital (já que sua saída é um sinal binário), é capaz de implementar uma quantidade limitada de funções lógicas, incluindo as operações básicas da álgebra booleana AND, OR e COMPLEMENT. Como classificador, o neurônio particiona o espaço vetorial N-dimensional dos padrões de entrada $X(k)$ em dois subespaços, associados biunivocamente a y={1, -1}, através da formação de uma superfície de decisão. Portanto, realiza um mapeamento $\mathbb{R}^N \rightarrow \mathbb{R}^2$ que propicia o reconhecimento de padrões em duas classes, identificadas com o respectivo subespaço. Demonstra-se [42] que a convergência de treinamento do Perceptron através do algoritmo de Rosenblatt impõe aos estímulos externos a condição de que devam ser linearmente separáveis.

Padrões linearmente separáveis são aqueles cuja superfície de decisão é um hiperplano em $\mathbb{R}^N$, ou uma reta em $\mathbb{R}^2$. Esta última situação está ilustrada pela fig. 3.6(a), que representa os padrões de entrada por pontos num espaço bidimensional e a superfície de decisão por uma reta. Para esta figura, define-se:

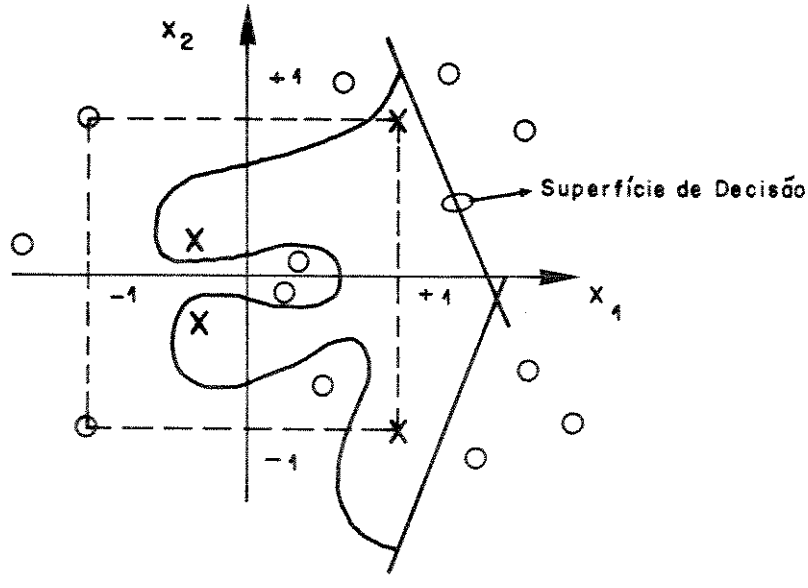
$$X(k)^T \triangleq [x_1(k)\ x_2(k)]$$

O símbolo o representa um padrão de entrada que gera saída y(k) positiva (+1), enquanto que o símbolo x representa um estímulo externo que gera saída negativa (-1). Portanto, o lado esquerdo da reta corresponde aos padrões associados à classe especificada pela saída +1; e o direito, aos padrões pertencentes à classe -1.

Considerando-se que o Perceptron, treinado pelo algoritmo de Rosenblatt, classifica corretamente apenas padrões linearmente separáveis, isto justifica porque este neurônio não consegue implementar todas as funções booleanas, visto que algumas estão associadas a padrões não-linearmente separáveis (cuja superfície de decisão não corresponde a um hiperplano em $\mathbb{R}^N$ ou a uma reta em $\mathbb{R}^2$ - fig. 3.6(b) -). Como exemplo clássico, pode-se citar a função EXCLUSIVE-OR.



(a)



(b)

Figura 3.6 : Tipos de padrões de entrada bidimensionais (N = 2).
(a) Linearmente separáveis;
(b) Não-linearmente separáveis.

Demonstra-se [43] que a rede Perceptron multi-camadas com duas camadas é capaz de formar qualquer superfície de decisão arbitrariamente complexa, ou seja, classifica qualquer conjunto de padrões não-linearmente separáves. Consequentemente, implementa qualquer função lógica booleana. Tal capacidade desta arquitetura do Perceptron multi-camadas está diretamente associada à quantidade de neurônios da camada intermediária [1,32], a qual pode ser extremamente elevada devido às características do estímulo externo e da aplicação, de forma a inviabilizar sua utilização ou acarretar escolha de outra estrutura (por exemplo, uma rede com três camadas, com quantidade menor de nós) [1].

Todas as demonstrações referidas nos dois últimos parágrafos exigem apenas que a função de ativação da rede seja monotônica crescente, contínua e limitada. Não é obrigatória a presença de uma não-linearidade no neurônio de saída da rede para a validade dos resultados, o que justifica a hipótese assumida no início desta seção.

Não obstante a grande potencialidade do Perceptron multi-camadas classificador, sua operação prática enfrenta três problemas principais.

O projeto da rede é realizado empiricamente, arbitrando-se de início uma determinada arquitetura (com base nas características da aplicação) que é então simulada. A configuração final decorre por tentativa e erro, não existindo métodos gerais para se especificar a quantidade de nós necessários à estrutura. Sabe-se, porém, que a capacidade do Perceptron multi-camadas na formação de mapeamentos complexos é proporcional à quantidade de neurônios escondidos [1], cuja ordem de grandeza é em geral a mesma da quantidade de estímulos de treinamento.

O aprendizado do Perceptron multi-camadas através do algoritmo de retropropagação é complexo, lento e dependente das condições iniciais. Isto está associado aos mínimos locais da

superfície de erro quadrático médio [1], que possui profundas depressões e vários planos, acarretando gradiente excessivamente elevado ou reduzido. Além disso, o treinamento deve ser capaz de conferir à rede boa GENERALIZAÇÃO (vide propriedade coletiva no.6, seção 2.5), a qual depende de três fatores: da sistemática de apresentação dos padrões de treinamento, de sua quantidade e de sua representatividade estatística relativamente ao universo de estímulos a serem processados pela rede. Embora existam estudos teóricos a respeito do segundo fator, tanto a sistemática de apresentação como a escolha dos padrões de aprendizado são procedimentos essencialmente empíricos. A generalização assume papel fundamental no caso de aplicações em que se dispõe de uma quantidade limitada de estímulos de treinamento, e pode ser incrementada através da eliminação de pesos sinápticos de magnitude reduzida [44].

A implementação física do Perceptron multi-camadas é problemática, em função da dificuldade de se construir dispositivos eletrônicos estáveis que gerem a função sigmóide (bem como sua derivada) com a precisão necessária para os cálculos envolvidos na operação desta rede.

Com o objetivo de diminuir o tempo de treinamento do Perceptron multi-camadas e de incrementar a propriedade coletiva de representações invariantes, realiza-se atualmente um pré-processamento dos padrões de entrada. Por exemplo, ao invés de se apresentar diretamente à rede os "pixels" de uma imagem ou amostras de um sinal, estes são pré-processados de forma que os dados de entrada consistam em amplitudes do espectro de frequências, difones ou coeficientes LPC [10]. Isto ajuda a rede criar mais rapidamente a representação interna distribuída do estímulo, otimizando o aprendizado. É importante ressaltar que esta sistemática possui forte motivação biológica, em termos de percepção pelo sistema nervoso [5]. Sabe-se, por exemplo, que estímulos sonoros são pré-processados pelos ouvidos externos e médio (detecção de "pitches" via cóclea ou por mapas tonotópicos

[35]) antes de serem reconhecidos no centro auditivo do cortex (o qual pode ser associado a uma rede neural).

3.5 - REDES NEURAIS NÃO-SUPERVISIONADAS: MAPAS DE KOHONEN.

ASPECTOS GERAIS

Kohonen [5,33] introduziu uma rede neural denominada "Mapa Auto-Organizativo", mostrada na fig. 3.7. Consiste no agrupamento de vários neurônios posicionados espacialmente de acordo com um arranjo geométrico pré-definido (por exemplo, quadrangular - fig. 3.7(a) - ou circular - fig. 3.7(b)), e que não são interligados fisicamente entre si por pesos sinápticos. Cada estímulo externo $\mathbf{X}(k)$ é aplicado simultaneamente à entrada de todos os nós. Cada neurônio do mapa é denominado "Unidade Adaptativa Básica de Memória" ou simplesmente "neurônio de Kohonen" [5], consistindo em um Perceptron com função de ativação linear. Apresentam-se abaixo as equações que definem a relação de entrada-saída de um nó de Kohonen de ordem N (vide fig. 3.8).

$$\mathbf{X}(k)^T \triangleq [x_1(k)\ x_2(k)\ \ldots\ x_N(k)] \tag{3.23a}$$

$$M(k) = \{\mathbf{M}_1(k),\ \mathbf{M}_2(k),\ \ldots,\ \mathbf{M}_M(k)\} \tag{3.23b}$$

$$\mathbf{M}_i(k)^T \triangleq [m_{i1}(k)\ m_{i2}(k)\ \ldots\ m_{iN}(k)];\ i = 1,\ 2,\ \ldots,\ M \tag{3.23c}$$

$$\eta_i(k) = \mathbf{X}(k)^T.\mathbf{M}_i(k) \tag{3.23d}$$

Onde

N : quantidade de conexões externas.

M(k) : conjunto dos vetores de pesos sinápticos do mapa no instante k.

M : quantidade de neurônios do mapa auto-organizativo.

$\mathbf{M}_i(k)$: vetor de pesos sinápticos associado ao nó i.

$\eta_i(k)$: ATIVIDADE NEURAL DE SAÍDA (ou simplesmente sinal de saída) do nó i.

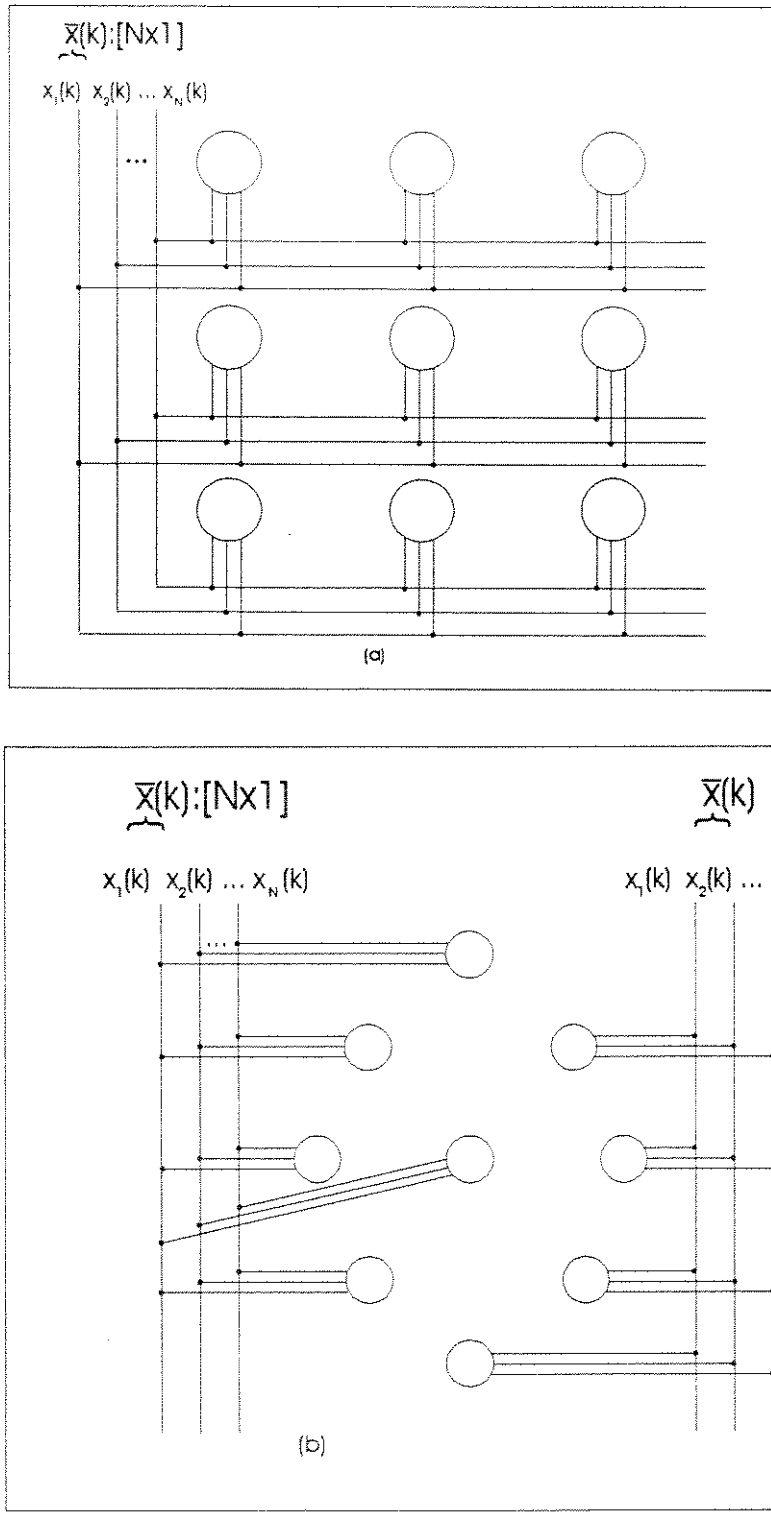


Figura 3.7 : Arquitetura da rede neural mapa auto-organizativo [5].
(a) Topologia quadrangular;
(b) Topologia circular.

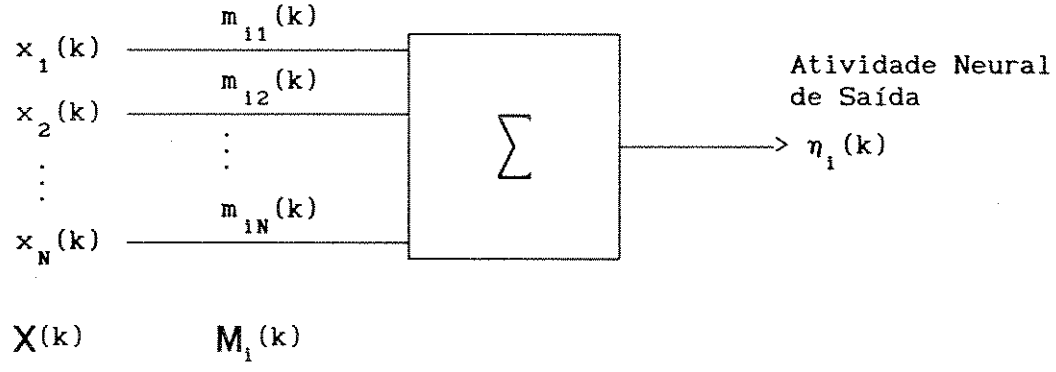


Figura 3.8 : Neurônio de Kohonen ou Unidade Adaptativa Básica de Memória (UABM) - ordem N [5].

As diferenças fundamentais entre a unidade adaptativa básica de memória e o Perceptron residem nos princípios matemáticos e biológicos inspiradores e no tipo de treinamento. Conforme análise a ser conduzida na presente seção, enquanto o nó de Kohonen representa um modelo matemático linear associado aos mapas cerebrais de características - seção 2.3 -, treinado de forma não-supervisionada; o Perceptron constitui uma extensão do modelo não-linear proposto em [2,20], cujo aprendizado é supervisionado.

O mapa auto-organizativo de Kohonen possui três características principais:

1) Apresenta, em regime permanente, a propriedade coletiva de Codificação Espacial (seção 2.5, propriedade coletiva 7).

Cada estímulo externo $X(k)$, apresentado ao mapa de Kohonen, induz atividade neural de saída em apenas um grupo específico de nós da rede, ou seja, apenas tais neurônios apresentam sinal de saída com amplitude significativa. Cada um destes grupos, ativado se e somente se o padrão externo possuir determinadas características físicas, é denominado "CAMPO DE RECEPÇÃO" para tais características.

Portanto, um mapa auto-organizativo de Kohonen, em regime permanente, é um conjunto de campos de recepção estáveis. Isto significa que, considerando-se um sistema de coordenadas x-y posicionado no centro geométrico do mapa, pode-se estabelecer uma relação biunívoca entre a presença de atividade neural de saída em uma determinada localidade da rede (correspondente à ativação de um determinado campo de recepção) e a presença de uma característica física específica no estímulo externo. Isto está exemplificado no mapa de Kohonen da fig. 3.9, de topologia circular, que supõe $X(k)$ como pixels de uma imagem. Um campo de recepção corresponde a um setor circular, e sua ativação indica a cor associada ao padrão de entrada.

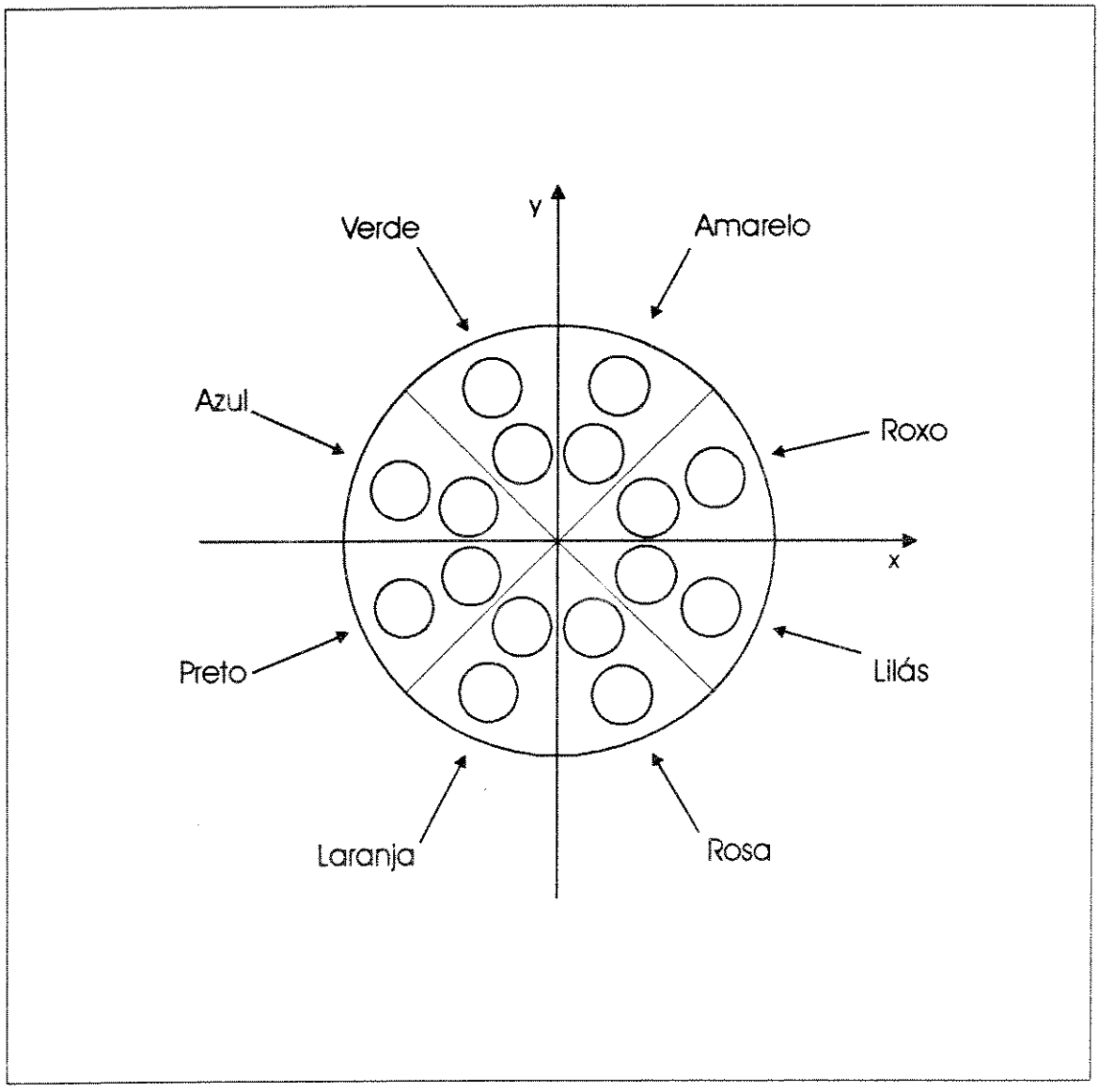


Figura 3.9 : Propriedade coletiva de codificação espacial do mapa auto-organizativo de Kohonen.

2) A adaptação dos pesos sinápticos é não-supervisionada, através de uma técnica denominada "APRENDIZAGEM COMPETITIVA" (a ser analisada na próxima subseção). Neste caso, o objetivo da adaptação NÃO é minimizar uma função de custo do erro (visto que ele não pode ser definido para esta rede neural), mas sim mapear a estrutura estatística do estímulo externo no conjunto de pesos sinápticos do sistema.

3) A amplitude da saída de cada neurônio da rede, para cada iteração do aprendizado, NÃO é obtida por uma operação matemática realizada entre o estímulo que sensibiliza suas entradas e o vetor de pesos sinápticos, mas sim fixada externamente de acordo com uma determinada regra pré-estabelecida. Tal regra está associada à formação dos campos de recepção do mapa e ao princípio biológico de "INTERAÇÃO CELULAR MÚLTIPLA", a ser analisado nesta seção.

O mapa de Kohonen possui duas aplicações principais, decorrentes da sistemática de treinamento. A primeira corresponde à estimação da função densidade de probabilidade do estímulo externo (através do cálculo da função densidade de probabilidade do conjunto de pesos sinápticos da rede, como resultado assintótico da aplicação da técnica de aprendizagem competitiva), e a segunda à separação ou classificação dos padrões de entrada em uma quantidade finita de classes (através da propriedade de codificação espacial). Portanto, deve-se ressaltar que, em contraposição às redes neurais analisadas até o presente momento, a informação associada ao mapa de Kohonen concentra-se nos valores dos pesos sinápticos e na presença (ou ausência) de atividade neural em determinados grupos de neurônios, e NÃO na magnitude dos sinais de saída ou na relação entrada-saída propriamente dita da estrutura.

A seguir, a técnica de aprendizagem competitiva e o princípio de interação celular múltipla, princípios básicos do aprendizado do mapa auto-organizativo de Kohonen, serão brevemente analisados.

Seja:

$$\mathbf{X}(k) \triangleq [x_1(k)\ x_2(k)\ \ldots\ x_N(k)] \tag{3.24a}$$

$$R(k) = \{\mathbf{R}_1(k), \ldots, \mathbf{R}_M(k)\} \tag{3.24b}$$

$$\mathbf{R}_i(k) \triangleq [r_{i1}(k)\ r_{i2}(k)\ \ldots\ r_{iN}(k)];\ i = 1, \ldots, M \tag{3.24c}$$

Onde $\mathbf{X}(k)$ representa um processo estocástico formado por N variáveis aleatórias $x_i(k)$, R(k) um conjunto de M vetores de referência no instante k e $\mathbf{R}_i(k)$ o i-ésimo elemento de R(k). Seja então estabelecido um critério de distância entre $\mathbf{X}(k)$ e $\mathbf{R}_i(k)$ (por exemplo, a norma euclidiana entre os dois vetores), denotado por $d(\mathbf{X}(k),\mathbf{R}_i(k))$. Denomina-se $\mathbf{R}_{i,max}(k)$ como o "vetor de referência sintonizado ao máximo a $\mathbf{X}(k)$", para cada instante k, como o vetor de R(k) tal que $d(\mathbf{X}(k),\mathbf{R}_{i,max}(k))$ seja mínima [33]. Isto está expresso matematicamente pela expressão que se segue.

$$\mathbf{R}_{i,max}(k) \triangleq \left\{\mathbf{R}_a(k) \in R(k) \mid \right.$$

$$\left. \mid d(\mathbf{X}(k),\mathbf{R}_a(k)) \leq d(\mathbf{X}(k),\mathbf{R}_i(k))\ \forall i\right\} \tag{3.25a}$$

$$i = 1, 2, \ldots, M \tag{3.25b}$$

$$i,max = a \tag{3.25c}$$

A aprendizagem competitiva representa um procedimento adaptativo aplicado ao conjunto de vetores de referência R(k), que objetiva minimizar uma função de custo envolvendo $d(\mathbf{X}(k),\mathbf{R}_{i,max}(k))$ [33]. Cada adaptação é realizada apenas sobre $\mathbf{R}_{i,max}(k)$, sendo que os demais vetores $\mathbf{R}_i(k)$ não são modificados. Supõe-se, em geral, inicialização aleatória de R(k). Em regime permanente, o processo estocástico de entrada é "mapeado" no conjunto de vetores de referência, de forma que estes evidenciem características intrínsecas de $\mathbf{X}(k)$.

A aprendizagem competitiva pode ser realizada através do método de "Quantização Vetorial" [45], técnica desenvolvida para

aplicação ao processamento digital de voz. Neste caso, define-se como critério de distância a norma euclidiana, e como função de custo, o valor quadrático médio da norma euclidiana, expressos respectivamente pelas seguintes equações:

$$d(\mathbf{X}(k), \mathbf{R}_i(k)) = \|\mathbf{X}(k) - \mathbf{R}_i(k)\| \tag{3.26}$$

$$J = E\left[\|\mathbf{X}(k) - \mathbf{R}_{i,max}(k)\|^2\right] \tag{3.27}$$

Onde:

$$\mathbf{R}_{i,max}(k) = \left\{\mathbf{R}_a(k) \in R(k) \;|\; \|\mathbf{X}(k) - \mathbf{R}_a(k)\| \leq \|\mathbf{X}(k) - \mathbf{R}_i(k)\| \;\forall i\right\} \tag{3.28a}$$

$$i = 1, 2, \ldots, M \tag{3.28b}$$

$$i,max = a \tag{3.28c}$$

Demonstra-se [33,46] que os valores ótimos teóricos de $\mathbf{R}_i(k)$, estabelecidos pela minimização da eq. (3.27), são tais que:

$$p(R^*) \cong [p(\mathbf{X}(k))] \tag{3.29a}$$

$$10^2 \leq N \leq 10^3 \tag{3.29b}$$

$p(R^*)$: função densidade de probabilidade do conjunto de vetores de referência ótimos R^*.

$p(\mathbf{X}(k))$: função densidade de probabilidade do processo estocástico $\mathbf{X}(k)$.

A desigualdade da eq. (3.29b) corresponde a uma condição mínima para a aproximação da eq. (3.29a), válida para a maioria das aplicações da quantização vetorial ao processamento digital de voz [33].

Portanto, a estimação da função de densidade de probabilidade dos valores ótimos do conjunto de vetores de referência R(k), definidos pela minimização do valor quadrático médio da norma euclidiana (eq. (3.27)), propicia uma aproximação da função

densidade de probabilidade do processo estocástico de entrada.

Não é possível obter os valores ótimos teóricos de $R_i(k)$ diretamente a partir da eq. (3.27), mesmo para um processo aleatório $X(k)$ estacionário [33]. Portanto, deve-se definir um procedimento adaptativo, que minimiza uma função de custo estocástica J(k) através do método "steepest-descent" [33], expresso pelas seguintes equações:

$$J(k) = \|X(k) - R_{i,max}(k)\|^2 \qquad (3.30)$$

$$R_{i,max}(k+1) = R_{i,max}(k) + \alpha(k).\left[X(k) - R_{i,max}(k) \right] \qquad (3.31a)$$

$$R_i(k+1) = R_i(k);\ \text{se } i \neq i,max \qquad (3.31b)$$

$$0 < \alpha(k) < 1;\ \forall\ k \qquad (3.31c)$$

As eqs. (3.31a-c) definem uma espécie de "algoritmo do gradiente estocástico" para aprendizagem competitiva, onde $\alpha(k)$ é uma sequência de valores escalares monotonicamente decrescente, que pode ser associada a um passo de adaptação. Demonstra-se [33] que este procedimento converge assintoticamente para os valores ótimos de $R_i(k)$ definidos pela minimização da eq. (3.27).

INTERAÇÃO CELULAR MÚLTIPLA, REALIMENTAÇÃO LATERAL MÚLTIPLA e FORMAÇÃO DE CAMPOS DE RECEPÇÃO NOS MAPAS DE KOHONEN.

Conforme discutido na seção 2.3, um mapa cerebral de características é uma região particular do cortex do encéfalo de animais vertebrados, ativada somente quando determinadas operações de percepção são realizadas. Por sua vez, a atividade neural de um mapa cerebral de características está organizada espacialmente em campos de recepção. Um campo de recepção é definido como um grupo de células nervosas que apresentam atividade neural de saída apenas quando um sinal sensorial de entrada (associado a uma operação de percepção processada pelo mapa cerebral de características considerado) possuir determinadas características

físicas [5,33].

A interação celular múltipla corresponde ao mecanismo de interação existente entre todas as células nervosas de um mapa cerebral de características. Através de uma complexa rede de sinapses, que interliga todos os neurônios do mapa, as células influenciam-se mutuamente (inibindo ou ativando a si mesmas ou as demais), de forma que o estímulo externo induza atividade neural em apenas um campo de recepção.

A referência [5] apresenta um modelo neurofisiológico de interação celular múltipla, aproximadamente válido para cada um dos neurônios do cortex de animais mamíferos. O tipo de sinapse (excitatória ou inibitória) é uma função da distância entre o neurônio considerado e as demais células nervosas, sendo que a sinapse excitatória de máxima eficiência corresponde a uma conexão de realimentação.

Como consequência deste mecanismo, a amplitude da atividade neural de saída (denotada pela variável $\underline{I}$) dos neurônios componentes de um mapa cerebral de características pode ser aproximada pelas figuras 3.10. Um ponto (d,I) destes gráficos representa o valor de $\underline{I}$ associado a uma célula nervosa situada a uma distância radial $\underline{d}$, relativamente ao centro geométrico de um campo de recepção ativado (delimitado por uma circunferência de raio r). A fig. 3.10(a) apresenta o caso ideal enquanto a fig. 3.10(b), uma aproximação por uma função contínua (uma gaussiana).

O mapa auto-organizativo de Kohonen corresponde, na realidade, a um mapa cerebral de características simplificado. Ambos os sistemas apresentam estrutura subdividida em campos de recepção e a propriedade coletiva de codificação espacial. Entretanto, para o mapa de Kohonen, os campos de recepção (indicados por circunferências na fig. 3.11) possuem em geral uma topologia fixa, sendo que os neurônios componentes apresentam-se simetricamente distribuídos em torno de um centro geométrico, onde se localiza um nó denominado "neurônio máximo" (vide fig. 3.11).

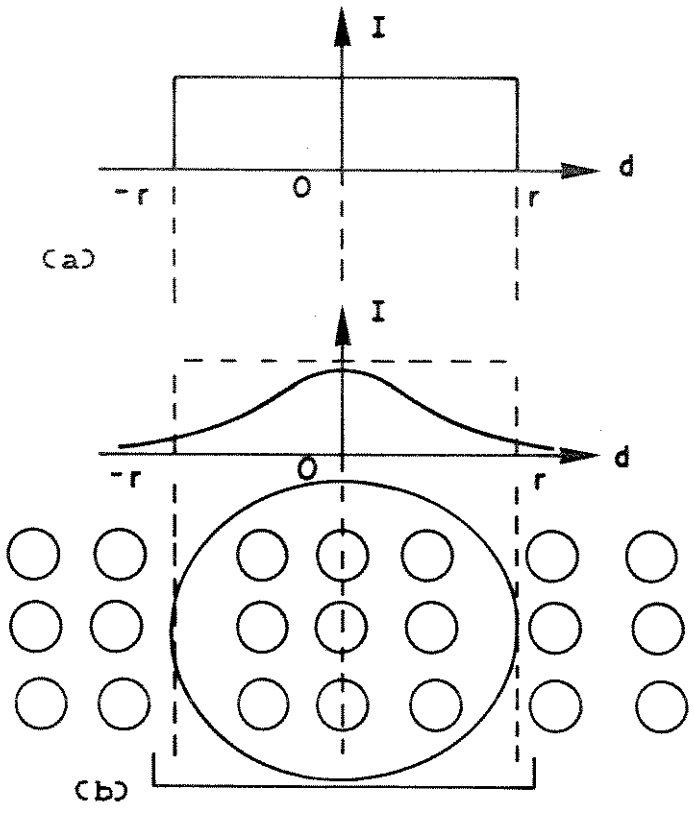


Figura 3.10 : Padrão de atividade neural de saída dos neurônios de um mapa cerebral de características [5].
(a) Caso ideal;
(b) Aproximação por função contínua (gaussiana).

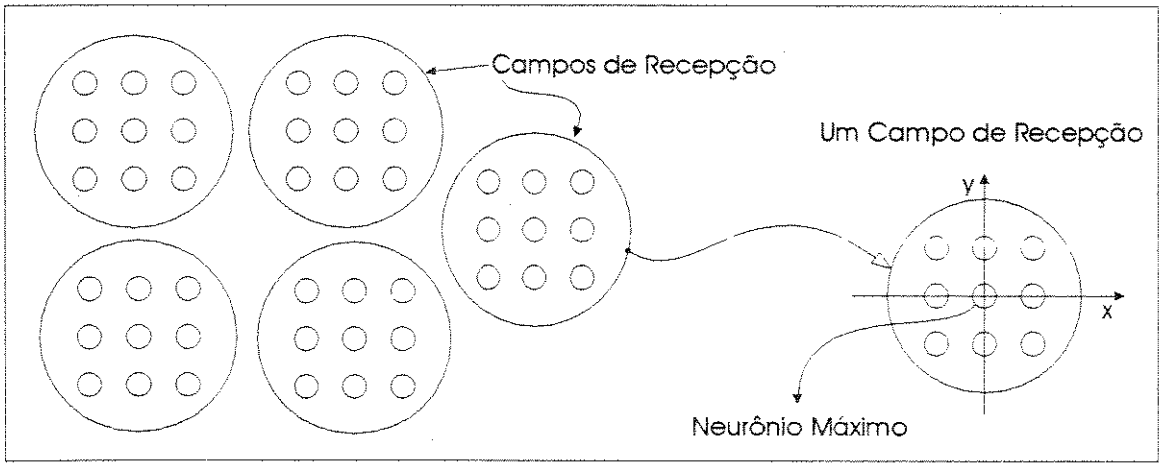


Figura 3.11: Mapa auto-organizativo de Kohonen como conjunto de campos de recepção.

Por analogia ao mapa cerebral, a amplitude da atividade neural de saída de um mapa de Kohonen, em regime permanente, deve ser a mais próxima possível ao padrão apresentado pelas figuras 3.10(a) ou 3.10(b). Com este objetivo em vista, o treinamento do mapa auto-organizativo emprega um procedimento denominado REALIMENTAÇÃO LATERAL MÚLTIPLA, uma espécie de interação celular múltipla simplificada, analisado a seguir.

A formação de um campo de recepção do mapa de Kohonen consiste, a cada iteração k do aprendizado, em três operações fundamentais.

O1) Identificar o centro de atividade neural para o estímulo de entrada atual $\mathbf{X}(k)$. Ou seja, identificar o neurônio do mapa que melhor represente as características intrínsecas de $\mathbf{X}(k)$, sendo que tal neurônio é indexado pelo subscrito i,max.

Em termos de aprendizagem competitiva, o centro de atividade neural corresponde ao nó do mapa cujo vetor de pesos sinápticos $\mathbf{M}_{i,max}(k)$ está "sintonizado ao máximo" ao estímulo externo $\mathbf{X}(k)$, de acordo com as eqs. (3.25a-c) (considerando-se que os vetores de referência $\mathbf{R}_i(k)$ sejam aqui representados por $\mathbf{M}_i(k)$).

O2) Intensificar a atividade neural de saída em torno do nó i,max. Isto equivale a fixar a amplitude dos sinais de saída dos neurônios do mapa de forma que apenas i,max e um determinado grupo de neurônios vizinhos a ele mesmo (situados no interior de uma região geométrica com topologia pré-estabelecida, denotada por $N_c(k)$ - por exemplo, uma circunferência-) apresentem atividade neural de saída com amplitude elevada, comparativamente aos demais nos do sistema.

Este procedimento corresponde a uma espécie de aproximação do padrão de atividade neural de saída do mapa cerebral de características, estabelecido pela função I(d) das figuras 3.10(a)-(b) (neste caso, deve-se notar que o centro geométrico do "campo de recepção", situado em d = 0, corresponde ao nó i,max).

A região geométrica $N_c(k)$ do mapa de Kohonen pode ser considerada como um "campo de recepção" primitivo, ainda em fase de estabilização.

O3) Reduzir o espaço geométrico de $N_c(k)$ até um limite mínimo pré-estabelecido. No caso de uma circunferência, isto é equivalente a reduzir seu raio. Assim, na próxima iteração do treinamento em que o índice $i_{,max}$ considerado representar o centro de atividade neural, a região geométrica $N_c(k+1)$ conterá uma menor quantidade de nós.

As figuras 3.12(a)-(b) apresentam a evolução temporal das operações O2 e O3 do procedimento de realimentação lateral múltipla, para o caso de um centro de atividade neural fixo. Supõe-se uma região geométrica circular $N_c(k)$, de raio r(k). A ordenada I(k,d) representa a amplitude do sinal de saída dos nós do mapa de Kohonen, situados a uma distância d do centro, para o intante k. O padrão de atividade neural de saída é fixado, para cada iteração, pela função I(k,d) (operação O2); enquanto que a redução de r(k) é estabelecida por variações temporais em I(k,d) (operação O3). A fig. 3.12(b) ilustra uma função I(k,d) do tipo gaussiana de variância decrescente no tempo. $N_c(T)$ e r(T) definem um campo de recepção, onde T é a quantidade total de iterações do treinamento.

As figuras 3.13(a)-(c) apresentam a evolução do procedimento de realimentação lateral múltipla para um mapa de Kohonen de topologia quadrangular, onde os pontos representam centros de atividade neural e as circunferências, as regiões $N_c(k)$. Durante as fases iniciais do transitório (fig. 3.13(a)), a operação O1 define múltiplos centros de atividade neural, de forma que os campos de recepção interagem entre si por superposição ou intersecção. Com o decorrer do processo, entretanto, campos associados a grupos de estímulos externos com características comuns tendem a se aproximar (fig. 3.13(b)), reduzindo a quantidade de centros de atividade neural e propiciando o

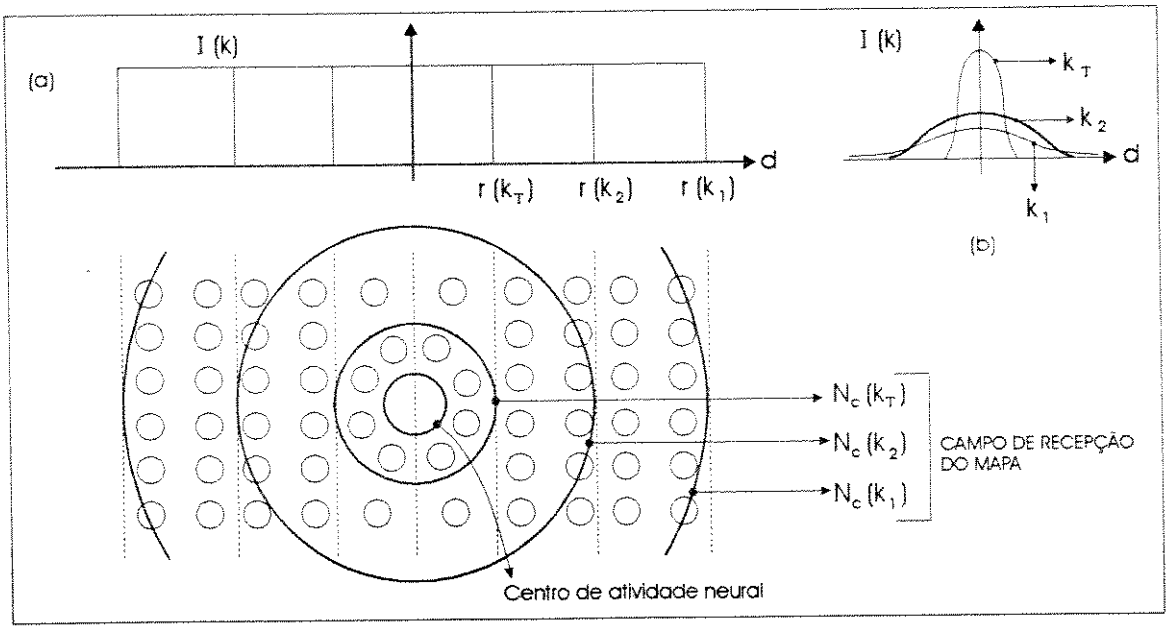


Figura 3.12: Evolução temporal das operações 02 e 03 do procedimento de realimentação lateral múltipla para um centro de atividade neural fixo.
T: quantidade de iterações de treinamento; $k_1 < k_2 < k_T$.
(a) Padrão de atividade neural de saída ideal;
(b) Padrão de atividade neural de saída gaussiano.

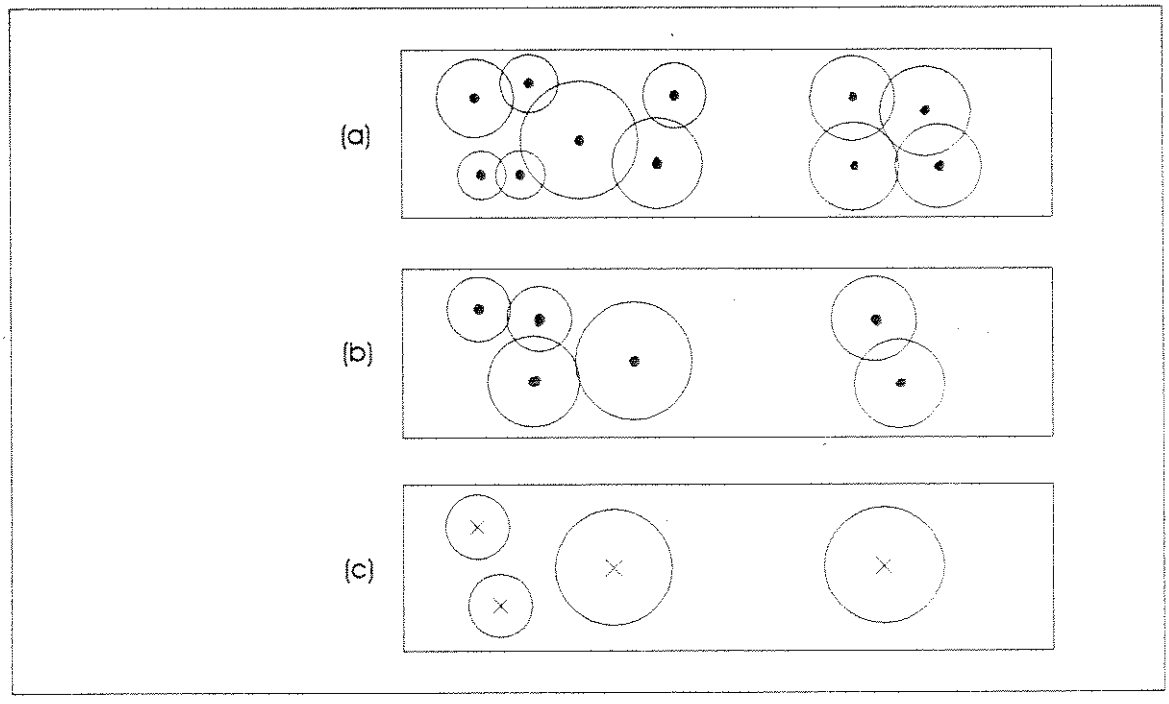


Figura 3.13: Dinâmica da formação de campos de recepção do mapa de Kohonen através do procedimento de realimentação lateral múltipla.
(a) Transitório I;
(b) Transitório II;
(c) Regime Permanente.

estabelecimento de campos de recepção estáveis em regime permanente (ou seja, distintos, conforme a fig. 3.13(c), onde as cruzes representam os centros geométricos dos campos).

Finalmente, deve-se lembrar que o procedimento de realimentação lateral múltipla é um mecanismo eficiente e alternativo para simplificar a implementação da interação celular múltipla. Isto porque o procedimento apresentado substitui o treinamento de um mapa Kohonen de complexa topologia, para a qual todos os neurônios seriam interconectados (o que deveria também incluir conexões de realimentação de elevada magnitude). Portanto, a realimentação lateral múltipla define completamente o processamento paralelo distribuído do mapa de Kohonen.

APRENDIZADO

O treinamento do mapa auto-organizativo de Kohonen corresponde à utilização conjunta do procedimento de realimentação lateral múltipla e da aprendizagem competitiva, que propiciam, respectivamente, a formação de campos de recepção estáveis e o mapeamento da estrutura estatística dos estímulos externos no conjunto de pesos sinápticos da rede, de forma simultânea.

Apresenta-se a seguir um exemplo de algoritmo de treinamento [5], onde se considera a adaptação dos pesos sinápticos do mapa através da quantização vetorial. O conjunto de vetores de referência R(k) e suas M componentes $R_i(k)$ (eqs. (3.24b-c)) correspondem, respectivamente, a M(k) e $M_i(k)$ (eqs. (3.23b-c)) de um mapa de Kohonen com M neurônios, cada um dos quais com N entradas externas. A região geométrica $N_c(k)$ é considerada uma circunferência de raio r(k). Realizam-se T iterações de treinamento.

ALGORITMO DE TREINAMENTO DO MAPA AUTO-ORGANIZATIVO [5]

Para k = 1 a T

Dado $\mathbf{X}(k)$:

* O1) Identificação do Centro de Atividade Neural (nó i,max).

$$\mathbf{M}_{i,max}(k) = \left\{ \mathbf{M}_a(k) \in M(k) \mid \| \mathbf{X}(k) - \mathbf{M}_a(k) \| \leq \| \mathbf{X}(k) - \mathbf{M}_i(k) \| \ \forall i \right\} \quad (3.32a)$$

$$i = 1, 2, \ldots, M \quad (3.32b)$$

$$i,max = a \quad (3.32c)$$

* O2) Intensificar atividade neural de saída em torno do nó i,max.

$$\eta_i(k) = 1 \text{ se } i \in N_c(k) \quad (3.33a)$$

$$\eta_i(k) = 0 \text{ se } i \notin N_c(k) \quad (3.33b)$$

* Adaptar neurônios contidos em $N_c(k)$ por quantização vetorial.

$$\mathbf{M}_i(k+1) = \mathbf{M}_i(k) + \alpha(k).\left[\mathbf{X}(k) - \mathbf{M}_i(k) \right]; \text{ se } i \in N_c(k) \quad (3.34a)$$

$$\mathbf{M}_i(k+1) = \mathbf{M}_i(k); \text{ se } i \notin N_c(k) \quad (3.34b)$$

$$0 < \alpha(k) < 1; \ \forall k \quad (3.34c)$$

* O3) Diminuir espaço geométrico de $N_c(k)$ (raio $r(k)$).

$$r(k+1) = g[r(k)] \mid r(k+1) < r(k) \quad (3.35)$$

$$10.P \leq T \leq 100.P \quad (3.36)$$

P : quantidade de estímulos externos $\mathbf{X}(k)$ para aprendizado.

T : quantidade de iterações de treinamento.

M : quantidade de neurônios do mapa auto-organizativo.

$N_c(k)$: região geométrica em torno do centro de atividade neural.

$r(k)$: raio da região geométrica (circunferência, neste caso).

g[] : operação matemática realizada sobre r(k) para diminuí-lo. Por exemplo, segundo uma evolução decrescente do seguinte tipo: $g(r(k)) = r(k)/k$.

O critério de identificação do centro de atividade neural, expresso pelas eqs. (3.32), corresponde exatamente à expressão estabelecida pelas eqs. (3.28). O padrão de atividade neural de saída (eqs. (3.33)) é estabelecido pela função I(k,d) da figura 3.12(a).

O aprendizado competitivo das eqs. (3.34) representa uma variante do "algoritmo do gradiente estocástico para a quantização vetorial" (eqs.(3.31)). Para o treinamento auto-organizativo, não somente $\mathbf{M}_{i,max}(k)$ é adaptado, como também todos os outros vetores de referência $\mathbf{M}_i(k)$ compreendidos pela região $N_c(k)$. Desta forma, em regime permanente, cada neurônio de um campo de recepção NÃO está "sintonizado ao máximo" a apenas um determinado $\mathbf{X}(k)$, mas sim a um conjunto de estímulos externos com características semelhantes. Este fato possui duas consequências principais. Primeiro, cada campo de recepção pode ser biunivocamente associado a uma determinada "classe" dos padrões de entrada. Segundo, o resultado expresso pelas eqs. (3.29) (demonstrado matematicamente) não é mais válido, sendo substituído pela seguinte proposição, apresentada em [5]:

PROPOSIÇÃO 3.1

"Seja X(k) um processo aleatório estacionário com função densidade de probabilidade p(X(k)) e f(.) uma função monotônica crescente e contínua qualquer. O treinamento de um mapa de Kohonen através do algoritmo das eqs (3.32-3.35) é tal que, em regime permanente, a função de densidade de probabilidade do conjunto de vetores de pesos sinápticos (p(M(k)) verifica a seguinte identidade:

$$\lim_{k\to\infty} p(M(k)) \cong f(p(\mathbf{X}(k)) \qquad (3.37)"$$

A proposição é demonstrada matematicamente apenas para casos simples (mapas de pequena dimensão), porém foi comprovada experimentalmente pela simulação de vários casos mais complexos [5].

Conclui-se, portanto, que o aprendizado auto-organizativo do mapa de Kohonen não apresenta a mesma precisão de representação estatística que a aprendizagem competitiva, embora seja capaz de separar e classificar os padrões de entrada em classes distintas, por mais complexos que sejam os estímulos externos $\mathbf{X}(k)$.

Tanto a convergência do algoritmo apresentado quanto a validade da proposição 3.1 independem da forma da região geométrica $N_c(k)$ e da sistemática de apresentação dos estímulos $\mathbf{X}(k)$ ao mapa. Por outro lado, a estabilidade do algoritmo depende da variação temporal de $\alpha(k)$ e de $r(k)$ (ambos necessariamente decrescentes), ajustados empiricamente à aplicação considerada. Os vetores $\mathbf{M}_i(k)$ devem ser inicializados aleatoriamente, porém de forma que sejam distintos entre si. A quantidade de iterações de treinamento é empírica, não possuindo limite fixo.

Finalmente, deve-se ressaltar um aspecto muito importante do treinamento do mapa auto-organizativo. Constata-se experimentalmente [33] que sua capacidade de classificação, em regime permanente, pode ser melhorada se os neurônios forem retreinados de forma supervisionada. Isto está associado ao fato de que nem sempre a precisão da estimativa da estrutura estatística do estímulo externo pelo mapa de Kohonen é compatível com as exigências da aplicação.

COMPARAÇÃO A REDES NEURAIS SUPERVISIONADAS

As principais diferenças entre o mapa auto-organizativo de Kohonen e o Perceptron multi-camadas são:

a) O processamento paralelo distribuído do mapa é estabelecido

pelo procedimento de realimentação lateral múltipla, podendo ser "controlado", de certa forma, externamente. Os neurônios não são interconectados. Para o Perceptron, o processamento paralelo depende das conexões sinápticas existentes entre os nós da rede.

b) O mapa apresenta a propriedade coletiva de codificação espacial, que implica, simultaneamente, numa representação interna das características do estímulo externo de forma local (organizada espacialmente em campos de recepção) e de forma distribuída (através do conjunto de pesos sinápticos, que assume significado estatístico preciso). Consequentemente, o arranjo geométrico dos nós do mapa é fundamental e o estímulo externo é conectado a todas unidades constituintes. Em contrapartida, a representação interna do Perceptron é distribuída, porém seus pesos sinápticos não podem ser associados diretamente a um determinado tipo de representação matemática, já conhecida ou bem estabelecida, das características do estímulo externo (em comparação ao conjunto de nós do mapa de Kohonen, que representam os padrões de aprendizado por sua função densidade de probabilidade).

c) O nó de Kohonen é linear (embora o mapa não o seja, visto que a amplitude dos sinais de saída é "quantizada" pelo padrão de atividade neural de saída I(k,d) - figuras 3.12 -), enquanto o Perceptron é não-linear.

d) O aprendizado do mapa é auto-organizativo e objetiva estimar a estrutura estatística dos estímulos externos. A cada iteração, apenas alguns vetores de pesos sinápticos são modificados ("adaptação seletiva"). O treinamento do Perceptron é supervisionado, objetiva minimizar um erro quadrático médio e todos os neurônios da rede são adaptados a cada iteração.

e) A informação do mapa concentra-se nos valores do conjunto de pesos sinápticos e na presença (ou ausência) de atividade neural de saída nos campos de recepção. A informação do Perceptron concentra-se na amplitude do sinal de saída da rede.

Baseado em recentes resultados experimentais [33], Kohonen afirma que a magnitude do tempo de treinamento de um mapa auto-organizativo com 1000 neurônios é de algumas ordens de grandeza menor que aquela de muitas redes supervisionadas. Deve-se notar que, segundo Widrow [31], em 1990 a literatura reportava Perceptrons multi-camadas de no máximo 150 neurônios como suficientes para aplicações a problemas extremamente complexos (por exemplo, controle multivariável). Os mapas auto-organizativos parecem, portanto, superar o incoveniente da complexidade de treinamento associada às redes supervisionadas, à custa de maior quantidade de neurônios. Isto pode estar associado, de certa forma, ao processamento paralelo distribuído "controlado" e à "adaptação seletiva" do mapa de Kohonen.

3.6 - REDE NEURAL FUNÇÃO RADIAL DE BASE (RBF).

ESTRUTURA, APRENDIZADO E ASPECTOS PRÁTICOS.

A rede neural função radial de base (RBF) [16] corresponde a uma estrutura híbrida de duas camadas de neurônios, que utiliza tanto campos de recepção auto-organizativos (cujos pesos sinápticos em regime podem ser estatisticamente caracterizados) quanto o processamento paralelo distribuído supervisionado, agrupando assim as vantagens das duas metodologias.

A fig. 3.14 apresenta a rede RBF. Cada nó escondido i da primeira camada é um neurônio de Kohonen com uma função de ativação não-linear do tipo "função radial de base" [16] (analisada no próximo parágrafo), e representa um campo de recepção (indexado por i), o qual está biunivocamente associado a um determinado conjunto de estímulos de treinamento $X_{j,i}$ com características semelhantes. A segunda camada é composta por Perceptrons lineares.

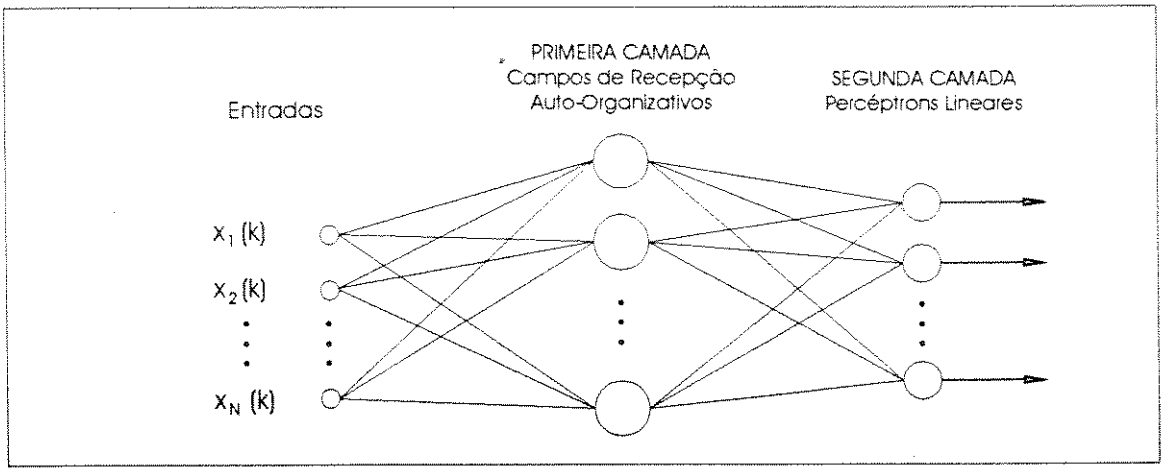


Figura 3.14: Rede neural função radial de base [7].

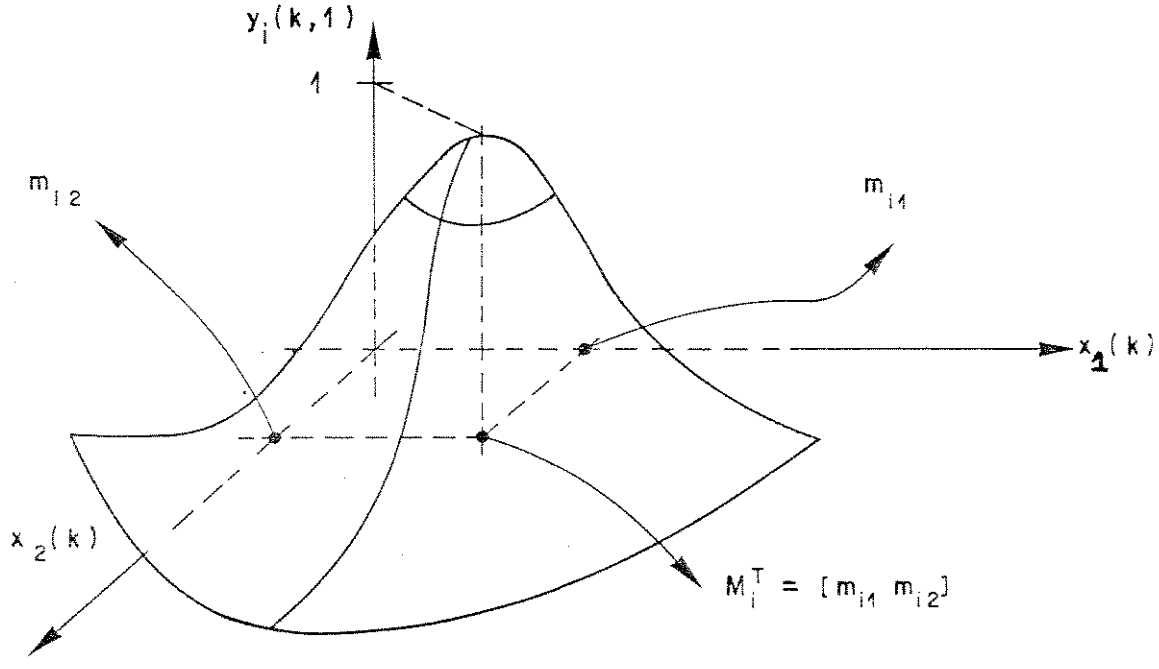


Figura 3.15: Sinal de saída do neurônio escondido i de uma rede neural função radial de base, com função de ativação gaussiana.
Padrões bidimensionais: $X(k)^T = [x_1(k)\ x_2(k)]$.

Uma função radial de base corresponde a uma função matemática não-linear multivariável que apresenta simetria radial. Como exemplo, pode-se citar a gaussiana, que estabelece a seguinte relação de entrada-saída para os neurônios da primeira camada da rede neural RBF em regime permanente (supostos com N pesos sinápticos):

$$y_i(k,1) = \exp\left[- \|\mathbf{X}(k) - \mathbf{M}_i\|^2 / 2.\sigma_i^2 \right]; \; i = 1, 2, \ldots, N_1 \qquad (3.38a)$$

$$\sigma_i^2 = (1/P_i) \cdot \sum_{j=1}^{P_i} \|\mathbf{X}_{j,i} - \mathbf{M}_i\|^2 \qquad (3.38b)$$

Onde

$y_i(k,1)$: saída do i-ésimo neurônio escondido da primeira camada.

$\mathbf{X}(k)$: estímulo de entrada atual.

$\mathbf{M}_i$: vetor de pesos sinápticos do i-ésimo nó escondido em regime permanente (ou parâmetro CENTRO DA GAUSSIANA), dimensão [Nx1].

N_1 : quantidade de nós da primeira camada (l = 1).

$\mathbf{X}_{j,i}$: j-ésimo estímulo externo de treinamento associado ao campo de recepção i .

P_i : quantidade de estímulos externos de treinamento $\mathbf{X}_{j,i}$.

σ_i^2 : variância da diferença entre os padrões de aprendizado $\mathbf{X}_{j,i}$ representados pelo i-ésimo campo de recepção e o centro da gaussiana $\mathbf{M}_i$, em regime permanente (ou parâmetro de NORMALIZAÇÃO).

A fig. 3.15 ilustra a saída $y_i(k,1)$ de um neurônio da primeira camada em regime permanente, para o caso de padrões de entrada bidimensionais (N=2). A atividade neural de saída do campo de recepção i é máxima no ponto determinado por $\mathbf{M}_i$, decrescendo conforme a variância especificada por σ_i^2. Quanto menor o valor desta última grandeza, menor o espalhamento dos dados de entrada relativamente ao campo de recepção e melhor a representação estatística destes. Deve-se notar que a saída de um nó escondido

da rede função radial de base, em regime permanente, é contínua e limitada no intervalo [0,1], ao contrário da saída do neurônio de Kohonen, em geral quantizada pelo padrão de atividade neural estabelecido pelo procedimento de realimentação lateral múltipla.

O aprendizado da rede neural função radial de base é geralmente realizado em três fases.

a) Fase 1: Formação dos campos de recepção da primeira camada através do estabelecimento de estimativas iniciais para os parâmetros σ_i^2 e $\mathbf{M}_i$ de cada nó escondido, com base na aprendizagem competitiva. Corresponde, portanto, a um treinamento auto-organizativo.

O método de quantização vetorial (eqs. (3.28), (3.30) e (3.31), seção 3.5, onde os vetores de referência $\mathbf{R}_i(k)$ são aqui representados pelos vetores de pesos sinápticos $\mathbf{M}_i(k)$) pode ser diretamente utilizado, visto que cada campo de recepção corresponde a um único neurônio da primeira camada (ao contrário do mapa de Kohonen, onde cada campo envolve vários nós). Além disso, deve-se notar que a formação dos campos de recepção não utiliza o procedimento de realimentação lateral múltipla (seção 3.5), pois o sinal de saída dos nós escondidos é determinado pela função de ativação radial de base empregada.

b) Fase 2: Treinamento dos neurônios Perceptron da segunda camada de forma supervisionada, através do algoritmo do gradiente estocástico convencional (eqs. (3.10) e (3.11)).

c) Fase 3: Retreinamento das duas camadas, de forma supervisionada, através do algoritmo de retropropagação (eqs. (3.21) e (3.22), seção 3.3). Esta fase é facultativa, e seu objetivo é refinar a estimação da estrutura estatística do estímulo externo pelos campos de recepção (através do ajuste das conexões sinápticas escondidas, de forma análoga ao retreinamento do mapa auto-organizativo de Kohonen) e de harmonizar a operação conjunta das duas camadas.

Em termos de classificação de padrões, demonstra-se matematicamente que a rede neural função radial de base é capaz de aproximar, de forma arbitrariamente próxima, qualquer mapeamento não-linear contínuo de entrada-saída [47]. Tal capacidade é matematicamente equivalente àquela de um Perceptron multi-camadas de classificar qualquer conjunto de padrões de entrada não-linearmente separáveis [43].

Por outro lado, contata-se experimentalmente que a velocidade de treinamento das redes funções radiais de base é em geral maior que aquela do Perceptron multi-camadas [1,48]. Isto pode ser explicado pela associação das vantagens do aprendizado auto-organizativo e supervisionado. Na primeira fase, os neurônios escondidos são treinados diretamente, com objetivo de capturar a estrutura estatística da entrada, o que agiliza a formação da representação interna distribuída da rede. Isto propicia treinamento supervisionado mais rápido, visto que a complexidade e a lentidão do algoritmo de retropropagação estão associadas ao aprendizado do nós escondidos. Além disso, deve-se destacar que a função de ativação destes neurônios da rede função radial de base não é pré-estabelecida e nem igual para todos os nós escondidos (como no caso do Perceptron multi-camadas), porém armazena informação estatística precisa sobre o estímulo externo.

ASPECTOS BIOLÓGICOS

Além de representar uma definição matemática sutil e conveniente, a rede função radial de base está intrinsecamente fundamentada pela organização operacional do sistema nervoso vertebrado.

Resultados experimentais da neurofisiologia sobre o sistema visual humano [49-51] evidenciam a existência de uma hierarquia no funcionamento de seus neurônios associados. O estudo realizado em [49] caracteriza o primeiro estágio de processamento de estímulos visuais externos pela codificação espacial da informação, através

dos mapas de características visuais (como por exemplo, os mapas de linhas e de cores, seção 2.3), implementados por neurônios da retina ocular, os quais são sensibilizados diretamente pelos estímulos externos. Por outro lado, [50] apresenta resultados experimentais que evidenciam a existência de um segundo nível de processamento dos padrões visuais. Consiste em neurônios ativados indiretamente por estímulos do meio-ambiente (pois localizam-se no cortex, distantes da retina), capazes de responder a sinais externos complexos, como por exemplo, objetos com diferentes orientações espaciais e características de alto nível (faces humanas ou animais). Tais propriedades de representação invariante e de compressão de informação (vide seção 2.5) destes neurônios sugerem o possível modelamento do segundo nível de processamento de padrões visuais por uma rede neural multi-camadas. Isto foi realizado por [51], cujos resultados mostraram-se coerentes com a neurofisiologia.

Tal hierarquia de processamento de informação também se manifesta na organização do sistema nervoso. Ele realiza, em primeira instância, uma percepção do meio-ambiente e das variáveis internas do organismo através de estruturas neurais denominadas, respectivamente, receptores externos e proprioceptores [19]. Em uma segunda etapa, toda esta informação é transmitida pelos nervos ao sistema nervoso central, que se encarrega do tratamento dos dados e da geração de decisões, tarefas realizadas notavelmente pelo cortex cerebral.

Os dois parágrafos anteriores justificam, sob o ponto de vista da neurofisiologia, os comentários de Kohonen [5,33] a respeito do processamento de informação pelo sistema nervoso: "A representação interna de informação no cérebro, nos níveis de processamento mais baixos, ocorre de forma espacialmente organizada. (...) Nos níveis mais elevados, a operação do cérebro é baseada em conceitos abstratos, operações simbólicas e processos decisórios, baseados nas primeiras etapas de processamento. Nestes níveis, os mapas cerebrais são desordenados (...)". Note-se que

esta "desordem" da codificação espacial do segundo nível poderia ser interpretada como representação interna distribuída, ou seja, aquela associada ao processamento paralelo distribuído intrínseco ao Perceptron multi-camadas.

Os comentários de Kohonen sugerem, portanto, uma clara analogia entre o processamento de informação pelo sistema nervoso (tanto a nível de percepção visual quanto a nível de organização sistêmica) e por redes neurais funções radiais de base. A primeira etapa de processamento, percepção, corresponde à função do campo receptivo auto-organizativo da primeira camada da rede. A segunda etapa, análise e decisão baseadas no processamento multivariável simbólico, pode ser associada ao processamento paralelo distribuído da segunda camada da rede, formada por neurônios Perceptron lineares, cuja representação interna do estímulo de entrada está baseada em "símbolos". ("Símbolo", aqui, significa "padrão de atividade neural", capaz de representar simultaneamente diversas estruturas conceituais ou grandezas físicas - vide propriedade coletiva (1), seção 2.5 - ."Símbolo", portanto, evidencia a propriedade de compressão de informação, característica fundamental das redes neurais).

Deve-se notar que as duas etapas de processamento complementam-se. De um lado, os símbolos, utilizados para a representação interna do estímulo do meio-ambiente durante a segunda etapa, são formados a partir da informação estatística adquirida pela percepção. (Deve-se notar que tal informação representa, na realidade, uma estimativa de pouca precisão matemática, tendo em vista o volume de variáveis processadas pelo sistema nervoso, os ruídos ambientais e a exigência de resposta em tempo real). Por outro, a generalidade e abstração dos próprios símbolos (reflexo da propriedade de compressão de informação) possibilitam, de certa forma, superar a ineficiência da descrição precisa do estímulo externo do meio-ambiente pela percepção, o que pode ser justificado pelo fato do sistema nervoso ser capaz de gerar complexas decisões e ações em tempo real, garantindo a sobrevivência do organismo.

Neste contexto, pode-se então justificar a necessidade da terceira fase de treinamento da rede função radial de base, bem como a melhoria do desempenho do mapa auto-organizativo de Kohonen após retreinamento supervisionado.

A analogia entre o processamento de informação pelo sistema nervoso e pela rede neural função radial de base sugere que esta arquitetura corresponde a um modelo matemático mais próximo à realidade biológica do sistema nervoso vertebrado que o Perceptron multi-camadas ou o mapa de Kohonen isolados. Tal afirmação também está de acordo com os menores tempos de treinamento da rede função radial de base comparativamente ao Perceptron multi-camadas, o que aproxima seu desempenho dinâmico ao do sistema nervoso (capaz de gerar complexas decisões em tempo real).

3.7 - APLICAÇÕES AO PROCESSAMENTO DE SINAIS

O processamento de sinais corresponde, historicamente, a uma das primeiras aplicações de redes neurais [3]. Compreende os seguintes domínios:

- classificação e reconhecimento de padrões (principalmente para o caso de padrões com distribuição probabilística não-gaussiana ou arbitrária [30]);

- sinais de vídeo (compressão [39] e invariância do desempenho de classificação relativamente à translação, rotação e mudança de escala de imagens [30]);

- processamento de sinais geofísicos [52];

- filtragem adaptativa (equalização eficiente de canais de comunicação não-lineares e com diversos tipos de ruído [25,37,38]; além da maior velocidade de convergência e menor erro quadrático médio de regime que em filtros tradicionais IIR [53]).

Em termos de processamento de sinais de voz, deve-se destacar o sistema NETTalk [27] para a conversão texto-fala como uma das primeiras implementações práticas de redes neurais que chamou a atenção da comunidade científica internacional. Consiste em um Perceptron multi-camadas classificador e em um sintetizador digital de voz. Realizando-se uma varredura contínua do texto, a rede neural gera uma sequência coerente de parâmetros fonéticos para o controle do sintetizador, que realiza a conversão texto-fala propriamente dita. Deve-se notar que a tarefa desempenhada pela rede é notavelmente cognitiva, pois a pronúncia de palavras segue regras gerais baseadas na fonética e no contexto linguístico do texto, com inúmeras exceções e variantes dependentes da sentença. Ao invés de programar o sistema para responder corretamente para cada caso (o que implica em quantidade excessiva de memória física e estudo cansativo das inúmeras situações possíveis), a rede neural aprende as regras gerais e suas exceções através de exemplos.

O Perceptron utilizado no NETTalk possui 106 neurônios distribuídos em três camadas, com função de ativação sigmoidal. A entrada da rede é um vetor de sete caracteres de um trecho de texto, e suas 26 saídas correspondem a características de articulação fonética, das quais apenas uma permanece ativa para uma determinada entrada. O treinamento utilizou um conjunto de 1024 palavras, apresentado 50 vezes à rede.

Observou-se que o aprendizado da pronúncia de um texto pelo NETTalk ocorre de maneira similar ao processo infantil de aquisição de fala. Nas fases iniciais do treinamento da rede, observaram-se balbúcios na saída do sintetizador. A seguir, após o aprendizado de um conjunto limitado de regras gerais, os balbúcios evoluíram para sílabas e, posteriormente, para frases com erros sequenciais (análogo a uma criança sem conhecimento total para pronunciar corretamente). Finalmente, aprendidas as exceções, a voz produzida é clara e precisa, inclusive para palavras que não foram nem mesmo incluídas no treinamento.

Finalmente, as perspectivas de aplicação para as principais arquiteturas de redes neurais estudadas neste capítulo são apresentadas.

1) Perceptron multi-camadas

Pode ser considerado como um "filtro não-linear" genérico, capaz de desempenhar qualquer tipo de aplicação, sem campos específicos. Contudo, devido à sua capacidade de processamento de dados simbólicos, deve-se destacar que o Perceptron multi-camadas representa uma escolha natural para uma conexão entre a inteligência artificial (e mesmo a neurofisiologia ou a psicologia cognitiva) e o processamento de sinais, com o objetivo de implementar sistemas inteligentes.

Representa uma "rede neural de referência" para comparações, visto que é a arquitetura estudada há mais longo tempo de todas. Comparativamente ao mapa de Kohonen, tende a apresentar estrutura com menor quantidade de neurônios. Tal característica é importante para aplicações que utilizam extrema quantidade de dados de entrada multivariáveis, por exemplo, processamento de sinais geofísicos e de imagens.

2) Mapa auto-organizativo de Kohonen

Aplicado sobretudo à separação em classes e posterior classificação de padrões com estrutura estatística desconhecida, nos campos de processamento de voz (síntese) e de imagem (compressão, visão computacional). Pode também ser utilizado como detector de características (por exemplo, detecção de senóides em ruído branco ou de sinais em comunicação digital), para posterior geração de decisões em sistemas especialistas. Também pode ser utilizado na neurofisiologia para modelamento matemático do mecanismo de formação de mapas cerebrais de características.

3) Rede função radial de base

Aplicada sobretudo a tarefas que exijam compressão de informação seguida por um pós-processamento qualquer, por exemplo, compressão de imagens e classificação de padrões em visão computacional. Visto que implementa diretamente a quantização vetorial com precisão estatística elevada, também pode ser utilizada para análise de sinais (modelamento paramétrico, por momentos estatísticos de ordem superior, por "wavelets"). Cumpre lembrar que esta estrutura é a mais jovem de todas as redes neurais. Consequentemente, seu projeto e treinamento é o mais empírico de todos.

3.8 - CONCLUSÃO

Estudaram-se neste capítulo diversos modelos de redes neurais, comentando-se suas principais aplicações ao processamento de sinais. Sob o ponto de vista do processamento paralelo distribuído, pode-se identificar dois tipos básicos de redes:

- O primeiro bloco corresponde às arquiteturas não-lineares baseadas no processamento paralelo distribuído simples: Adaline, Perceptron e Perceptron multi-camadas, cuja análise matemática baseia-se na teoria de classificação de padrões. O treinamento é realizado de forma supervisionada, através de algoritmos de gradiente. A força e a fraqueza das estruturas deste bloco residem na não-linearidade e nos neurônios escondidos; se por um lado estes dois elementos são os responsáveis pela capacidade de formação de superfícies de decisão arbitrariamente complexas, por outro, os mesmos elementos acarretam, respectivamente, desempenho dependente da inicialização dos coeficientes (associado aos mínimos locais da superfície de erro quadrático médio) e complexidade de estimação do gradiente.

- O segundo bloco corresponde ao mapa auto-organizativo de Kohonen, que representa um mapa cerebral de características simplificado. É formado por neurônios lineares dispostos geometricamente de acordo com uma determinada topologia, cujo processamento paralelo distribuído não ocorre por conexões sinápticas, porém é estabelecido de acordo com a interação celular múltipla. O treinamento é não-supervisionado, com base na técnica de aprendizagem competitiva (que objetiva mapear a estrutura estatística do estímulo externo no conjunto de pesos sinápticos) e no procedimento de realimentação lateral múltipla (responsável pela formação dos campos de recepção do mapa, de forma que cada grupo de nós associado a um campo represente um "detector" de características intrínsecas do estímulo de entrada). Apenas alguns neurônios do mapa são modificados a cada iteração do aprendizado ("adaptação seletiva").

A rede função radial de base representa o ponto de encontro destes dois blocos. A utilização conjunta do processamento paralelo distribuído simples (na camada de saída) e da auto-organização (para os nós da primeira camada, sendo que um campo de recepção envolve apenas um neurônio escondido) em uma única estrutura multi-camadas propicia uma rede com capacidades semelhantes àquelas do Perceptron multi-camadas, porém com tempos de treinamento menores. É importante ressaltar o fato de que as características estatísticas do estímulo externo estão armazenadas não somente nos pesos sinápticos da primeira camada, como também nos parâmetros da função de ativação dos nós escondidos.

Evidenciou-se como os conceitos biológicos associados ao cortex, aos mapas cerebrais de características (por exemplo, neste caso, a interação celular múltipla) e à hierarquia de processamento de informação pelo sistema nervoso de animais vertebrados (a nível de organização sistêmica e a nível de percepção visual) estão intrinsecamente relacionados às redes neurais supervisionadas, ao mapa auto-organizativo de Kohonen e à

rede função radial de base, respectivamente. Através destes elementos, foi possível justificar o melhor desempenho dinâmico do Perceptron multi-camadas treinados com padrões pré-processados, a necessidade da terceira fase de treinamento da rede função radial de base e o seu aprendizado mais rápido comparativamente ao Perceptron multi-camadas. Tais elementos biológicos constituem, portanto, ferramentas fundamentais na análise de redes neurais.

Apresentados os princípios básicos (Capítulo 2) e os modelos matemáticos principais (Capítulo 3), comparam-se conceitos, algoritmos e estruturas de redes neurais e da filtragem adaptativa no capítulo seguinte.

CAPÍTULO 4

REDES NEURAIS E FILTRAGEM ADAPTATIVA: ANALOGIAS E DIFERENÇAS

Comparam-se neste capítulo as redes neurais e os filtros adaptativos, com objetivo de evidenciar a intrínseca interrelação existente entre os dois campos. Inicialmente, as diferenças fundamentais entre os dois sistemas são analisadas sob a ótica do treinamento supervisionado. Em seguida, apresentam-se algumas analogias matemáticas entre os dois campos, em termos de estruturas e de algoritmos de treinamento. Tais analogias envolvem, de um lado, filtros não-lineares, lineares e espaciais adaptativos, além do algoritmo do gradiente estocástico e dos algoritmos adaptativos supervisionados aplicados à predição linear; de outro, o neurônio Perceptron, o Perceptron multi-camadas e o neurônio de Kohonen, além do algoritmo de treinamento por retropropagação e as equações de aprendizado auto-organizativo [5]. Finalmente, discutem-se os pesos sinápticos ótimos de um neurônio Perceptron com base na equação de Wiener-Hopf [54].

4.1 - DIFERENÇAS NO CONTEXTO DO TREINAMENTO SUPERVISIONADO.

O aprendizado supervisionado tradicional de redes neurais, aplicadas à classificação de padrões, supõe em geral quantidade limitada de estímulos de treinamento e é dividido em três fases distintas (vide seção 2.2): adaptação, teste e utilização. Para a filtragem adaptativa, entretanto, não existe separação clara destas três etapas. O aprendizado é contínuo, onde se supõe que o sinal de referência pode ser sempre acessado, em contrapartida ao treinamento não-contínuo de redes neurais. Por outro lado, deve-se ressalvar que algumas aplicações da filtragem adaptativa limitam a quantidade de amostras a serem tratadas pelo filtro, devido à

rígida exigência de processamento em tempo real. Como exemplo, pode-se citar os sistemas de codificação de voz por predição linear [55].

As figuras 4.1 e 4.2 apresentam, respectivamente, uma rede neural e um filtro adaptativo para a comparação que se segue. Os subscritos $_a$ e $_n$ denotam grandezas associadas respectivamente ao filtro e à rede. O sinal d(k) é dito de referência, enquanto y(k) é a saída e e(k) o erro do sistema. A fig. 4.2(a) apresenta o filtro com entrada e saída escalares, enquanto a fig.4.2(b) utiliza uma notação vetorial para esta mesma entrada. O conjunto de coeficientes do filtro adaptativo é representado por $\mathbf{H}(k)$.

Um conjunto composto por P pares $\mathbf{X}(k)/d(k)$ é apresentado à rede neural (fig. 4.1) durante o aprendizado. Deve-se notar que k simplesmente indexa cada um destes pares, não representando o papel da variável tempo. O vetor $\mathbf{X}(k)$ é composto por N valores $x_i(k)$, que não são necessariamente correlacionados entre si, sob o ponto de vista matemático ou mesmo estatístico, podendo assumir diferentes sentidos físicos. Por exemplo, se $\mathbf{X}(k)$ for definido como o vetor estado clínico de um paciente, $x_1(k)$ poderia representar a frequência dos batimentos cardíacos [Hz], $x_2(k)$ a temperatura interna [$^{\circ}$C], $x_3(k)$ um valor binário associado ao sexo do paciente, etc.

No caso do filtro adaptativo com entrada vetorial (fig. 4.2(b)), cada vetor de treinamento $\mathbf{X}(k)$ é composto por N variáveis que representam, em geral, amostras adjacentes de uma série temporal, possuindo mesmo significado físico e mesma unidade de medida. O índice k representa o tempo e cresce indefinidamente. Desta forma, $\mathbf{X}(k)$ e $\mathbf{X}(k+1)$ podem ser expressos respectivamente pelas seguintes equações:

$$\mathbf{X}(k)^T = [x_1(k)\ x_2(k)\ \ldots\ x_N(k)] = [x(k)\ x(k-1)\ \ldots\ x(k-N)] \quad (4.1a)$$

$$\mathbf{X}(k+1)^T = [x_1(k+1)\ldots x_N(k+1)] = [x(k+1)\ x(k)\ \ldots\ x(k-N+1)] \quad (4.1b)$$

$$x_i(k+1) = \begin{cases} x(k+1) & ;\ i = 1; \quad \forall k \\ x_{i-1}(k) & ;\ i = 2, \ldots, N; \quad \forall k \end{cases} \quad (4.1c)$$

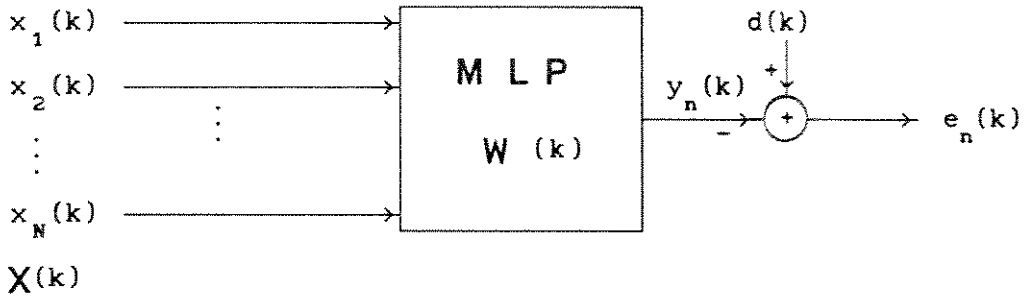
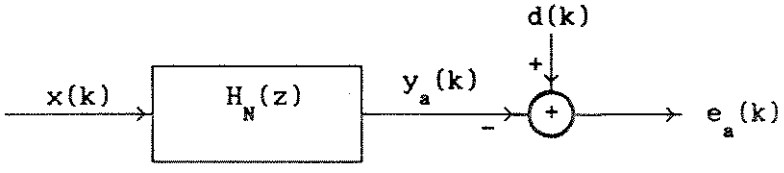
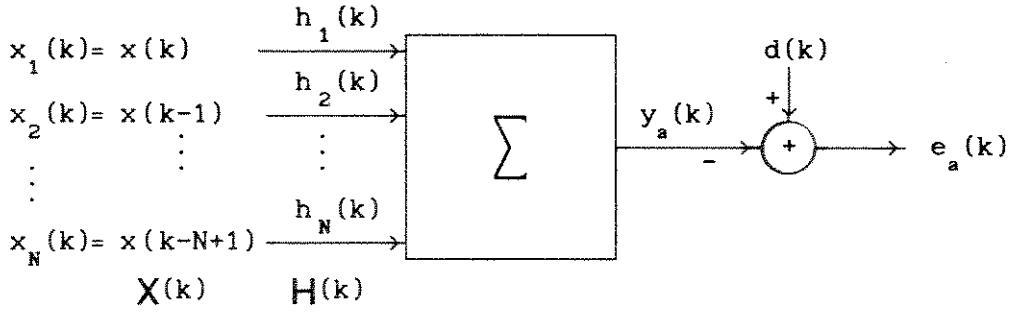


Figura 4.1 : Perceptron multi-camadas.



(a)



(b)

Figura 4.2: Filtro adaptativo - ordem N.
(a) Entrada Serial x(k);
(b) Entrada Vetorial X(k).

Assim, x(k+1) denota a nova amostra do sinal temporal apresentada ao filtro no instante k+1. As equações (4.1c) evidenciam a propriedade do deslocamento temporal associada aos padrões de treinamento $\mathbf{X}(k)$ da filtragem adaptativa.

Os estímulos de aprendizado da rede neural são geralmente conhecidos antes da realização do treinamento, correspondendo a um subconjunto representativo das características estatísticas do universo de padrões a serem processados pela rede. Portanto, os padrões de aprendizado são escolhidos de acordo com uma determinada metodologia, e a sequência de apresentação destes à rede não segue qualquer critério de ordenação. Deve-se notar que, em geral, alguns estímulos são reapresentados várias vezes durante o treinamento. Em contrapartida, o aprendizado da filtragem adaptativa não supõe um sinal de entrada previamente conhecido, e a ordem de apresentação das amostras é imposta pelas características físicas do processo estocástico filtrado.

Portanto, no contexto da classificação de padrões, a operação de redes neurais é não-adaptativa e não-recursiva (pois a apresentação de um novo vetor $\mathbf{X}(k)$ não provoca necessariamente alterações em $\mathbf{W}(k)$). O treinamento destes sistemas pode ser caracterizado pela função de custo J_n, definida pelas seguintes equações:

$$J_n = E[e_n(k)^2] \tag{4.2a}$$

$$e_n(k) = d(k) - y_n(k); \quad k = 1, 2, \ldots, T_n \tag{4.2b}$$

T_n : quantidade total de iterações da fase de adaptação da rede.

Uma aproximação prática comum para J_n, por exemplo, é:

$$J_n \cong (1/T_n) \cdot \sum_{m=1}^{T_n} e_n^2(m) \tag{4.2c}$$

$$P_n \leq T_n \tag{4.2d}$$

P_n : quantidade total de estímulos de aprendizado.

Se $P_n < T_n$, então ocorre apresentação repetida de determinados padrões, em geral os mais representativos do universo de estímulos externos, cuja caracterização estatística é normalmente suposta estacionária.

O aprendizado contínuo e recursivo de filtros adaptativos (onde cada estímulo externo $\chi(k)$ apresentado ao sistema modifica $H(k)$) objetiva minimizar a seguinte função de custo:

$$J_a = E[e_a(k)^2] \qquad (4.3a)$$

$$e_a(k) = d(k) - y_a(k); \quad k = 1, 2, \ldots, T_a \qquad (4.3b)$$

T_a : quantidade total de iterações de adaptação do filtro.

Uma aproximação prática comum para J_a, por exemplo, é:

$$J_a(k) = (1/N_c) \cdot \sum_{m=k-N_c+1}^{k} e_a^2(m) \qquad (4.3c)$$

$$T_a > N_c \qquad (4.3d)$$

$$P_a = T_a \qquad (4.3e)$$

N_c : comprimento da janela de dados.

P_a : quantidade total de amostras de x(k) utilizadas para aprendizado.

A eq. (4.3c) denota a aprendizagem permanente de uma série temporal (que pode ser não-estacionária), cujo erro quadrático $J_a(k)$ é estimado para cada iteração k utilizando-se uma janela de dados de comprimento N_c, correspondente a um intervalo de tempo pequeno em relação à ordem de grandeza do tempo de estacionariedade do sinal.

Os critérios de minimização expressos pelas funções de custo J_n (eq. (4.2c)), para a rede neural; e $J_a(k)$ (eq. (4.3c)), para o filtro adaptativo, tendem a ser assintoticamente equivalentes sob as seguintes condições [22,23]:

- O conjunto de vetores de aprendizado $\mathbf{X}(k)$ deve ser representativo das características estatísticas do universo de estímulos a serem processados pela rede neural;

- O sinal x(k) deve ser estacionário no sentido amplo (o que é equivalente a exigir que o conjunto $\mathbf{X}(k)$ possua características estatísticas estacionárias);

- A quantidade total de iterações da fase de adaptação da rede neural deve ser suficientemente elevada para a correta estimação da estatística dos padrões de treinamento $\mathbf{X}(k)$ (ou seja, $T_n \to \infty$).

A principal preocupação do treinamento de redes neurais reside em propiciar boa GENERALIZAÇÃO ao sistema (capacidade de tratar padrões diferentes daqueles de treinamento, quando os pesos sinápticos são mantidos constantes - propriedade coletiva 6, seção 2.5 - , ou seja, capacidade da rede em processar não-estacionariedades ou ruídos impostos aos estímulos de aprendizado sem modificar seu pesos sinápticos), item mais importante do que a própria velocidade de convergência do processo. Isto decorre das características dos estímulos de aprendizado, estruturalmente complexos (e disponíveis em quantidade limitada), bem como da sistemática não-adaptativa de operação. Neste caso, a metodologia de seleção dos vetores $\mathbf{X}(k)$ de aprendizado, sua quantidade e sua sistemática de apresentação à rede assumem papel fundamental.

Em contrapartida, a principal preocupação do treinamento de filtros adaptativos concentra-se na ADAPTATIVIDADE do sistema, definida como sendo a capacidade do filtro em acompanhar as variações da estrutura estatística do sinal de entrada. Portanto, objetiva-se fundamentalmente alcançar máxima velocidade de

convergência de aprendizado. Tais características do treinamento de filtros adaptativos estão associadas à natureza do sinal de entrada, suposto, em sua forma mais geral, um processo estocástico não-estacionário (matematicamente mais simples que os dados de entrada das aplicações clássicas de redes neurais, cujos vetores de aprendizado não possuem a propriedade do deslocamento temporal), do qual se dispõe um conjunto ilimitado de amostras. Isto exige, portanto, aprendizado contínuo. Consequentemente, a generalização de filtros adaptativos é reduzida.

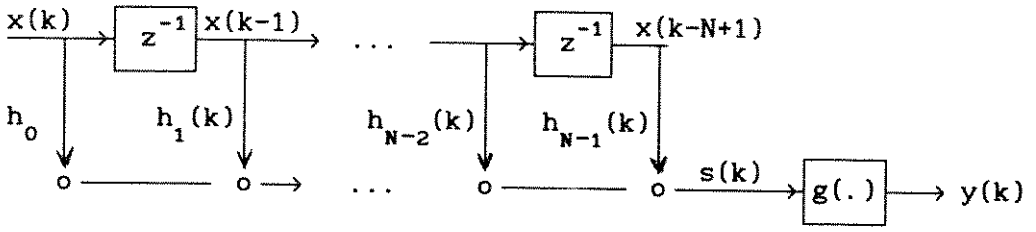
Cumpre destacar que, mesmo para o caso de um estímulo externo com caracterização estatística precisa (por exemplo, um sinal descrito por um modelo paramétrico), não se pode determinar teoricamente os pesos sinápticos ótimos de regime de uma rede neural supervisionada multi-camadas. Portanto, ao contrário dos coeficientes $H(k)$ de um filtro adaptativo, os pesos w_{ij} de um Perceptron multi-camadas não possuem significado matemático preciso, nem isolados e nem em grupo. (Entretanto, conforme será discutido na seção 4.4, demonstra-se [54] que os pesos sinápticos ótimos teóricos de um neurônio Perceptron correspondem à solução da equação de Wiener-Hopf [10,17,18] multiplicada por uma constante, sob determinadas restrições impostas ao estímulo externo).

Finalmente, deve-se destacar que não se pode provar matematicamente a convergência dos algoritmos de treinamento de redes neurais e não se tem conhecimento da existência de suas versões rápidas (como por exemplo, análogas aos mínimos quadrados rápidos [10, 17, 18] em filtragem adaptativa).

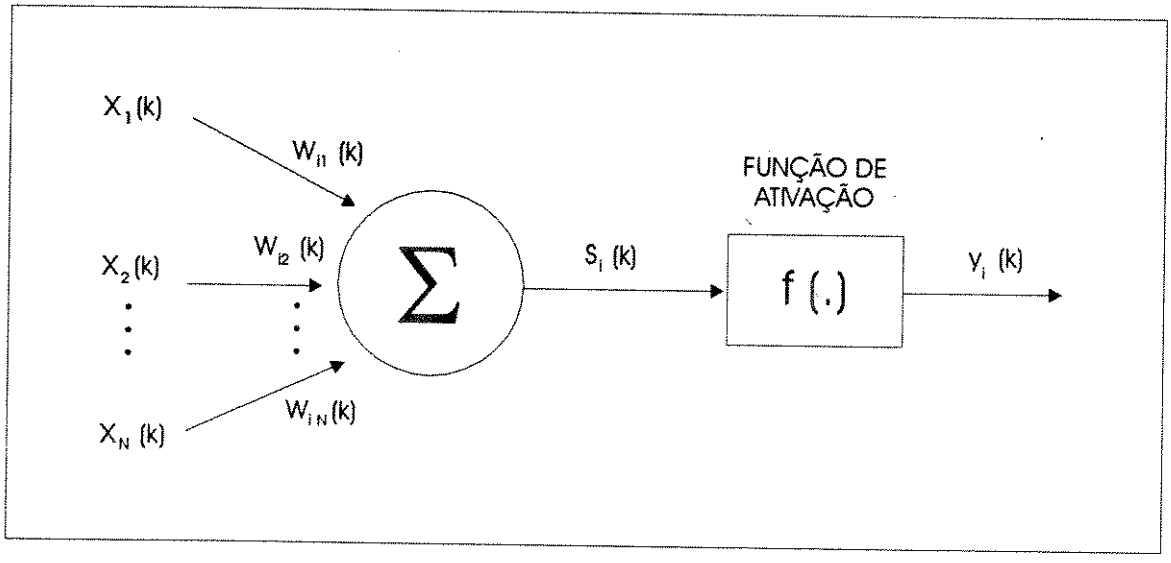
4.2 - ANALOGIAS ESTRUTURAIS

A1) Filtros Adaptativos; Perceptron e Perceptron multi-camadas.

As figuras 4.3(a)-(b) apresentam um filtro adaptativo não-linear de ordem N e um neurônio Perceptron i (com N pesos



(a)



(b)

Figura 4.3 - Analogia A1.

(a) Filtro não-linear adaptativo - ordem N.

(b) Neurônio Perceptron [15].

sinápticos), cujas relações de entrada-saída são respectivamente expressas por:

$$y(k) = g(s(k)) = g\left(\sum_{j=0}^{N-1} h_j(k).x(k-j) \right) \qquad (4.4a)$$

$$y_i(k) = f(s_i(k)) = f\left(\sum_{j=1}^{N} w_{ij}(k).x_j(k) \right) \qquad (4.4b)$$

As funções f(.) e g(.) são não-lineares por hipótese. Se estas forem iguais, conclui-se então que o Perceptron representa simplesmente um filtro adaptativo não-linear, estabelecendo-se uma clara correspondência entre os coeficientes w_{ij} e h_j, bem como entre as entradas $x_j(k)$ e $x(k-j)$. Tal analogia pode ser estendida para o caso de um filtro adaptativo recursivo não-linear [23], caso uma linha de atrasadores providenciar uma realimentação da saída $y_i(k)$, de modo que o estímulo externo $\mathbf{X}(k)$ do Perceptron seja composto por entradas $x_j(k)$ e por saídas atrasadas $y_i(k-j)$.

Portanto, um neurônio Perceptron pode representar simultaneamente filtros adaptativos FIR e IIR não-lineares. Deve-se notar que ambas as saídas y(k) e $y_i(k)$ correspondem ao mapeamento por uma função não-linear da correlação estatística entre o estímulo de entrada $\mathbf{X}(k)$ (agrupamento de amostras adjacentes do sinal x(k) para o caso do filtro adaptativo) e a informação especificada pelos coeficientes da estrutura considerada, conforme expresso pelas eqs. (4.4).

Tal analogia pode ser estendida para o caso do Perceptron Multi-camadas. Considerando-se que esta rede neural é uma associação de neurônios Perceptron e que é capaz de formar superfícies de decisão arbitrariamente complexas (vide seção 3.4), conclui-se que o Perceptron multi-camadas (ou MLP) implementa filtros FIR não-lineares (caso do MLP estático) e IIR não-lineares (caso do Perceptron multi-camadas dinâmico) de grande potencialidade. De fato, [23] apresenta uma arquitetura de um Perceptron multi-camadas dinâmico e demonstra a equivalência matemática de suas equações em relação ao modelo paramétrico

NARMAX do controle adaptativo.

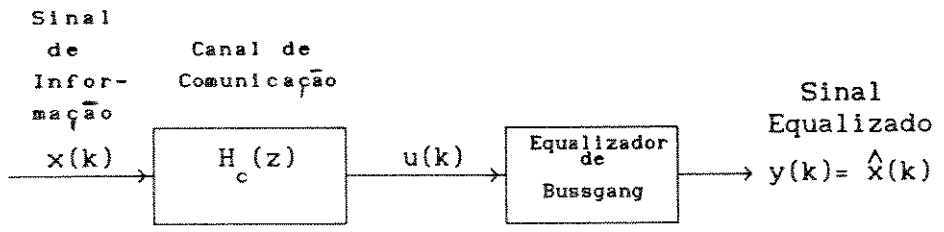
Em termos de analogias operacionais, tanto o treinamento do Perceptron e do Perceptron multi-camadas quanto o aprendizado de filtros adaptativos não-lineares são extremamente dependentes das condições iniciais (fato associado à não-convexidade das superfícies de erro quadrático médio). Em geral, os Perceptrons multi-camadas estáticos correspondem a estruturas maiores e mais complexas que suas respectivas redes dinâmicas equivalentes, utilizadas para desempenhar a mesma aplicação. Esta constatação prática é análoga ao fato de filtros adaptativos FIR de ordem infinita serem capazes de representar a mesma função de transferência de qualquer filtro IIR [10,17].

Caso a função de ativação do Perceptron for linear, o neurônio pode representar tanto um filtro adaptativo transversal quanto um IIR linear [23]. O mesmo pode ser afirmado para um Perceptron multi-camadas linear [22].

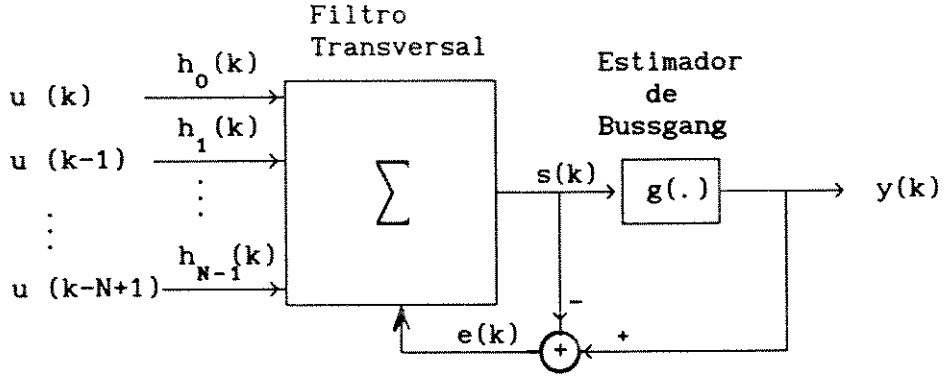
A2) Equalizador Adaptativo de Bussgang e Neurônio Perceptron.

A figuras 4.4(a)-(b) apresentam um filtro adaptativo não-linear, aplicado para a equalização cega de sinais digitais [10], denominado Equalizador Adaptativo de Bussgang. Sua entrada u(k) representa um sinal de informação x(k) transmitido por um canal de comunicação, cuja função de transferência $H_c(z)$ é de fase não-mínima por hipótese. O sinal u(k) é inicialmente processado por um filtro transversal $\mathbf{H}(k)$ e, em seguida, mapeado por uma função não-linear g(.) (em geral um estimador não-linear e sem memória [10], referido aqui como "Estimador de Bussgang") para gerar o sinal equalizado y(k) (ou $\hat{x}(k)$). Supõe-se que o sinal de referência externo não seja disponível.

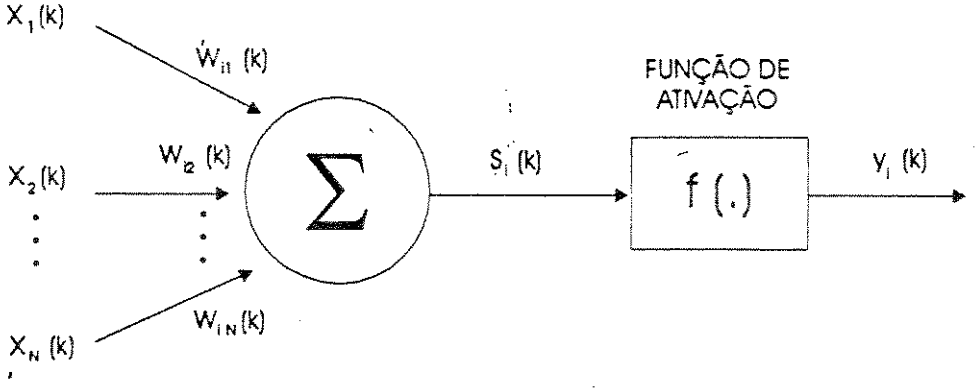
O equalizador de Bussgang objetiva, portanto, inverter a função de transferência do canal ($H_c(z)$) de forma adaptativa e não-supervisionada. O treinamento do filtro transversal é



(a)



(b)



(c)

Figura 4.4 : Analogia A2.

(a) Modelo do sistema de comunicação;

(b) Estrutura interna do equalizador de Bussgang [10];

(c) Neurônio Perceptron [15].

realizado através do algoritmo do gradiente estocástico, o qual emprega um "sinal de erro" e(k) estimado a partir da diferença entre o sinal equalizado ($\hat{x}(k)$) e a própria saída do filtro $H(k)$ (s(k)).

O neurônio Perceptron (fig.4.4(b)) e o equalizador adaptativo de Bussgang (fig.4.4(a)) são respectivamente caracterizados pelas seguintes equações:

$$y_i(k) = f(s_i(k)) = f\left(\sum_{j=1}^{N} w_{ij}(k).x_j(k) \right) \qquad (4.5a)$$

$$y(k) = g(s(k)) = g\left(\sum_{j=0}^{N-1} h_j(k).u(k-j) \right) \qquad (4.5b)$$

Onde N representa tanto a ordem do filtro transversal $H(k)$ quanto a quantidade de pesos sinápticos do Perceptron. Existe clara correspondência entre os coeficientes, as saídas lineares e as entradas das duas estruturas. Além disso, o estimador de Bussgang parece estar associado à função de ativação. De fato, g(.), na sua forma mais geral, é uma função não-linear exatamente como f(.) do neurônio. Deve-se notar que o estimador de Bussgang pode ser deduzido matematicamente e depende da relação sinal-ruído convolucional do sistema de comunicação [10]. Para valores reduzidos deste parâmetro, em particular, g(.) representa uma sigmóide, ou seja, uma função de ativação [10].

Entretanto, deve-se evidenciar que o treinamento do equalizador de Bussgang é auto-organizativo. Esta característica contrasta com o aprendizado supervisionado do Perceptron.

Outras analogias operacionais entre o Perceptron multi-camadas supervisionado e o equalizador cego de Bussgang podem ser destacadas:

- Tanto os vetores de treinamento $X(k)$ desta rede neural quanto o sinal x(k), a entrada do equalizador, são caracterizados por modelos estatísticos não-gaussianos [10,31];

- A função de custo a ser minimizada recursivamente é não-quadrática em relação aos parâmetros;

- A convergência inicial do algoritmo de treinamento é extremamente lenta, sendo que a velocidade de aprendizado aumenta somente quando se atinge erro quadrático médio abaixo de um determinado limiar [10].

A3) Filtro de Pilha e Neurônio Perceptron.

Filtro de pilha é um filtro digital não-linear adaptativo que, em regime permanente, gera saídas correspondentes a funções booleanas de suas entradas, as quais são quantizadas aos níveis {0,1} por hipótese. A fig. 4.5(a) apresenta um filtro de pilha [56], descrito pelas seguintes equações:

$$Y(k) = G(\mathbf{X}(k)) \tag{4.6}$$

$$\mathbf{X}(k)^T = [x_1(k) \ \ldots \ x_N(k)] \tag{4.7a}$$

$$\mathbf{H}(k)^T = [h_1(k) \ \ldots \ h_N(k)] \tag{4.7b}$$

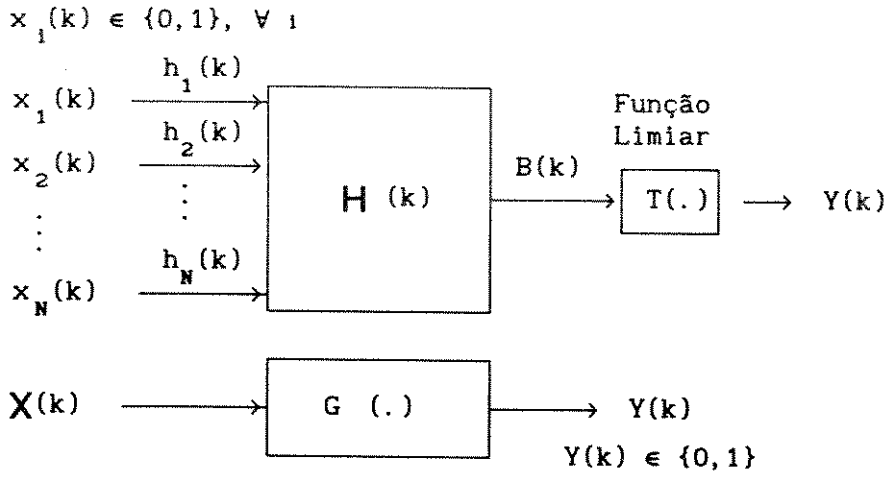
$$B(k) = \mathbf{X}(k)^T . \mathbf{H}(k) \tag{4.7c}$$

$$Y(k) = T(B(k)) = T\left(\sum_{j=1}^{N} x_j(k).h_j(k) \right) \tag{4.8a}$$

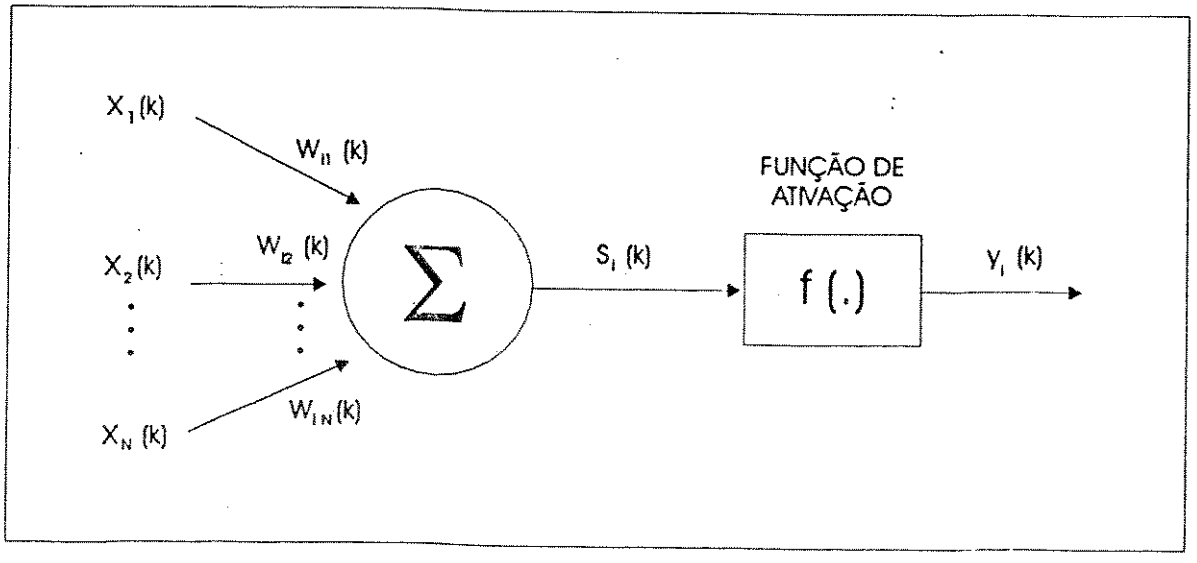
$$T(B(k)) = \begin{cases} 1 \text{ se } B(k) \geq \theta \\ 0 \text{ se } B(k) < \theta \end{cases} \tag{4.8b}$$

Onde:

$Y(k)$ = saída do filtro de pilha, booleana por definição ({0,1}).
$G(.)$ = relação de entrada-saída do filtro.
$\mathbf{X}(k)$ = vetor de dados de entrada (formado por valores booleanos por hipótese).
$\mathbf{H}(k)$ = vetor de coeficientes do filtro, comprimento N.
$B(k)$ = saída linear do filtro de pilha.
$T(\underline{a})$ = função de limiar do argumento $\underline{a}$.
θ = limiar mínimo de $T(\underline{a})$.



(a)



(b)

Figura 4.5 : Analogia A3.
(a) Filtro de pilha - ordem N [25,56];
(b) Neurônio Perceptron [15].

Apresentam-se logo abaixo exemplos de funções booleanas G(.), onde se supõe que N = 3 e que as operações lógicas AND, OR e COMPLEMENT sejam respectivamente representadas pelos símbolos ".", "+" e "-" :

$$G1(.) = x_1(k) + x_2(k).x_3(k) \tag{4.9a}$$

$$G2(.) = x_1(k).x_2(k) + \overline{x_3(k)} \tag{4.9b}$$

$$G3(.) = x_1(k) + \overline{x_2(k).x_3(k)} \tag{4.9c}$$

G1(.) é denominada função booleana positiva, visto que não envolve a operação lógica COMPLEMENT. Deve-se evidenciar que, por hipótese, filtros de pilha implementam apenas funções booleanas positivas em regime permanente, o que impõe as seguintes restrições matemáticas aos parâmetros do filtro [56]:

$$1)\ \theta \geq 0 \tag{4.10a}$$

$$2)\ h_i(k) \geq 0\ ;\ \text{para}\ \forall\ i = 1, \ldots, N\ \text{e para}\ \forall\ k \tag{4.10b}$$

As equações de um neurônio Perceptron de N entradas com função de ativação sinal (fig. 4.5(b)) são:

$$y_i(k) = f(s_i(k)) = f\left(\sum_{j=1}^{N} w_{ij}(k).x_j(k) \right) \tag{4.11a}$$

$$f(s_i(k)) = \begin{cases} 1 \text{ se } s_i(k) \geq 0 \\ 0 \text{ se } s_i(k) < 0 \end{cases} \tag{4.11b}$$

Portanto, desde que todos os pesos sinápticos w_{ij} do neurônio sejam positivos para qualquer instante k considerado (o que é equivalente à restrição expressa pela eq. (4.10b)), o neurônio Perceptron possui as mesmas equações que o filtro de pilha. Pode-se estabelecer, portanto, uma clara correspondência entre a função de limiar T(.) e a função de ativação f(.).

A4) Filtro Mediano e Perceptron Multi-Camadas.

Filtro mediano de ordem N é um filtro adaptativo não-linear cuja entrada é um sinal x(k) quantizado a N+1 níveis por hipótese. Corresponde à operação paralela de N filtros de pilha, cada qual processando um sinal binário $x_i(k)$ (quantizado aos níveis {0,1}), que é obtido de x(k) através de uma transformação matemática denominada "DECOMPOSIÇÃO DE LIMIAR" [56]. Tal decomposição define cada sequência booleana $x_i(k)$ através da seguinte equação:

$$x_i(k) = T_i(x(k)) = \begin{cases} 1 \text{ se } x(k) \geq i \\ 0 \text{ se } x(k) < i \end{cases} ; \; i = 1, 2, \ldots, N \qquad (4.12a)$$

Onde $T_i(.)$ é um caso particular da função de limiar definida pela eq. (4.8b), tal que o argumento a é representado pelo sinal x(k) e cujo limiar mínimo θ é dado por:

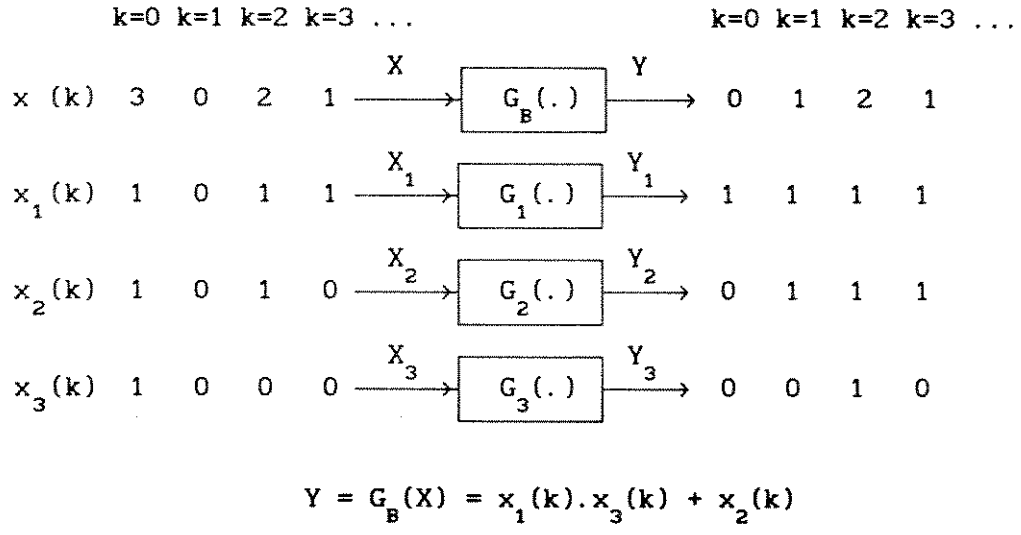
$$\theta = i; \; i = 1, 2, \ldots, N \qquad (4.12b)$$

A saída do filtro mediano é obtida por uma operação matemática que envolve todas as saídas dos filtros de pilha.

A fig. 4.6(a) apresenta um filtro mediano de ordem N=3 em regime permanente, caracterizado pela relação de entrada-saída $G_B(.)$, que processa entrada x(k) quantizada a 4 níveis ({0,1,2,3}). Corresponde à operação paralela de três filtros de pilha $G_1(.)$, $G_2(.)$, $G_3(.)$, que processam respectivamente as entradas $x_1(k)$, $x_2(k)$, $x_3(k)$, obtidas pela decomposição de limiar de x(k). A sequência de valores apresenta a evolução temporal das entradas e das saídas da estrutura. Deve-se notar que, ao contrário das relações de entrada-saída $G_i(.)$ dos filtros de pilha, $G_B(.)$ do filtro mediano não se restringe às funções booleanas positivas. De fato, para o caso da fig. 4.6(a) tem-se:

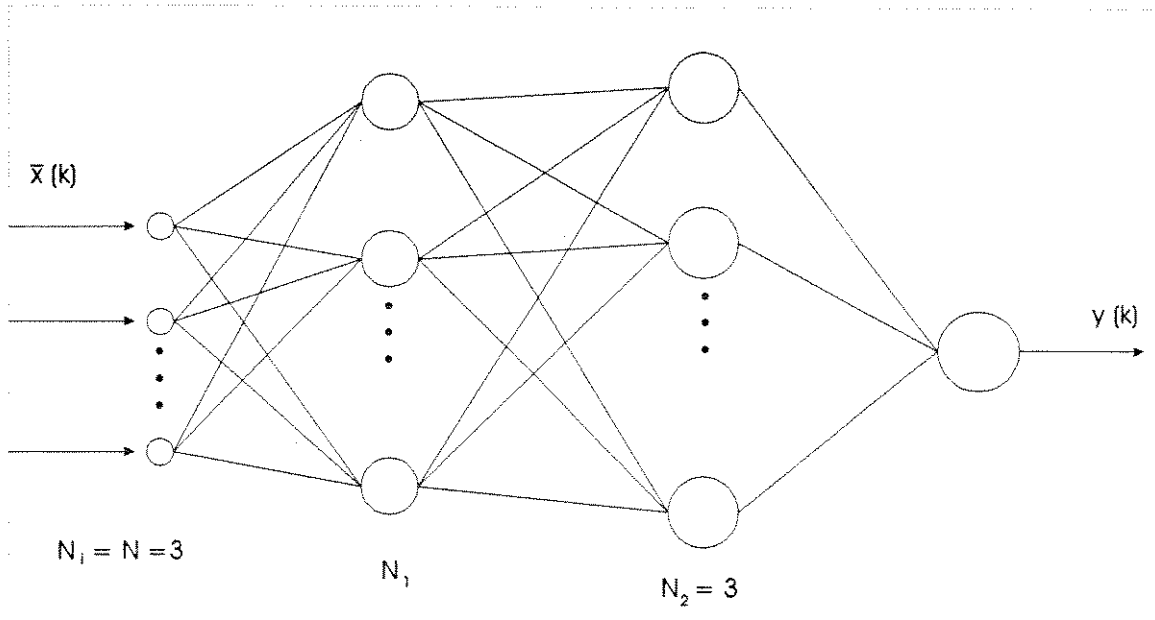
$$Y = G_B(X) = x_1(k).x_3(k) + x_2(k)$$

Onde "+" e "." denotam aqui operações aritméticas.



$$Y = G_B(X) = x_1(k).x_3(k) + x_2(k)$$

(a)



(b)

Figura 4.6 : Analogia A4.

(a) Filtro mediano de ordem N = 3 [25];

(b) Rede neural Perceptron multi-camadas $N_1 - N_1 - 3 - 1$ [57].

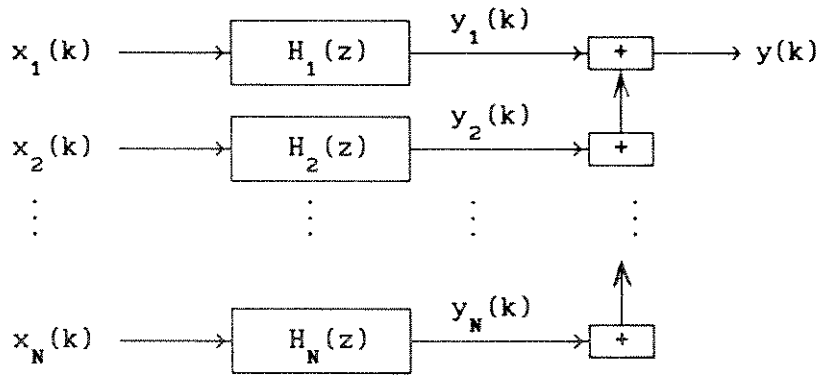
O interesse principal na utilização do filtro mediano reside na sua não-linearidade, que lhe confere capacidades superiores àquelas de estruturas lineares para determinados tipos de aplicação. Como exemplo, pode-se citar a equalização de imagens perturbadas por ruído não-gaussiano e não-aditivo [25,56]. A adaptividade do filtro é necessária para sua operação, visto que o projeto teórico envolve resolução de equações não-lineares e grande volume de variáveis.

Seja então o Perceptron multi-camadas N_1-N_1-3-1 da fig. 4.6(b), que possui N_1 entradas, N_1 neurônios na primeira camada, 3 neurônios na segunda e um nó de saída [57]. Impondo-se que cada uma das N_1 componentes dos vetores de treinamento $X(k)$ da rede sejam quantizadas a N_{1+1} níveis, e, supondo-se ainda a função de ativação sinal e pesos sinápticos positivos para todos os neurônios da segunda camada (para qualquer instante k considerado), pode-se então estabelecer uma analogia entre o filtro mediano da fig. 4.6(a) (N=3) e o Perceptron multi-camadas da fig. 4.6(b) (N_1=3). Neste contexto, as entradas e os nós da primeira camada realizam a decomposição de limiar sobre os padrões de entrada, enquanto os neurônios da segunda camada podem ser diretamente associados aos respectivos filtros de pilha da fig. 4.6(a).

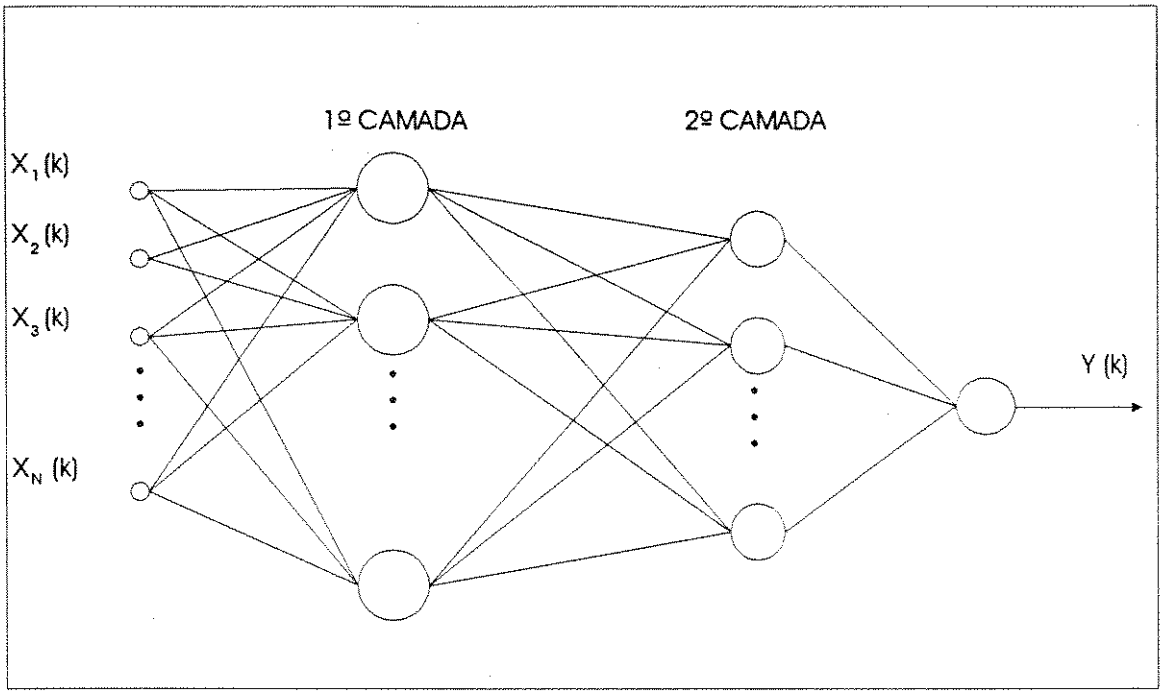
A5) Filtro Adaptativo Espacial e Perceptron Multi-Camadas

Estas estruturas estão respectivamente mostradas nas figuras 4.7(a)-(b). De forma geral, tanto os dados de entrada como os de saída, para ambos os casos, são multidimensionais. A comparação das figuras 4.7(a)-(b) sugere que o conjunto de filtros adaptativos (que compõem o filtro espacial) desempenha papel análogo a uma camada de neurônios do Perceptron multi-camadas.

A6) Filtro Adaptativo Transversal e Neurônio de Kohonen (ou Unidade Adaptativa Básica de Memória [5]).



(a)



(b)

Figura 4.7 : Analogia A5.

(a) Filtro adaptativo espacial - ordem N [17];

(b) Perceptron multi-camadas.

As equações de filtragem destas estruturas, ambas possuindo N coeficientes por hipótese, correspondem respectivamente a:

$$y(k) = \sum_{j=0}^{N-1} h_j(k).x(k-j) \qquad (4.13a)$$

$$\eta_i(k) = \sum_{j=1}^{N} m_{ij}(k).x_j(k) \qquad (4.13b)$$

O neurônio de Kohonen representa, portanto, um filtro adaptativo treinado de forma auto-organizativa, sendo que uma equivalência entre os coeficientes $h_j(k)$ e $m_{ij}(k)$ pode ser estabelecida. O comportamento assintótico deste neurônio, quando treinado por um conjunto limitado de padrões externos e através de um determinado algoritmo auto-organizativo [5] (a ser apresentado na analogia A8 da seção 4.3, eq. (4.18a)), é análogo ao regime permanente de um filtro de erro de predição linear adaptativo, empregado por exemplo em sistemas ADPCM (modulação por código de pulso adaptativa diferencial) ou em análise LPC (codificação de voz pelo método da predição linear) [55]. Para tais aplicações, o filtro realiza identificação do modelo paramétrico de uma sequência de quadros. Um quadro é definido como um segmento do sinal de entrada (por exemplo, um sinal de voz), composto por um conjunto limitado de amostras que possuem, geralmente, características estatísticas estacionárias.

Em regime permanente [5], a saída do neurônio de Kohonen apresenta a propriedade de ser nula para estímulos de entrada que representem uma combinação linear daqueles de treinamento, ou elevada caso contrário. (Diz-se, então, que o nó representa um "Detector de Novidades" [5]). Analogamente, a amplitude do erro de predição linear do filtro adaptativo em análise LPC tende a ser pequena em "regime permanente" (o qual corresponde à apresentação das últimas amostras de um quadro ao sistema), visto que os dados de entrada, neste instante, ainda estão associados ao modelo paramétrico identificado pelo filtro. Entretanto, o erro de predição pode assumir valores elevados quando as primeiras

amostras do quadro seguinte são apresentadas, já que a não-estacionariedade do sinal de voz pode induzir diferenças entre dois quadros consecutivos, em termos de modelamento paramétrico.

4.3 - ANALOGIAS ENTRE ALGORITMOS

A7) Algoritmo do Gradiente Estocástico (LMS) e de Treinamento por Retropropagação (vide seção 3.3).

As equações destes dois algoritmos são respectivamente apresentadas logo abaixo, onde se supõe o treinamento de um filtro transversal adaptativo $\underline{i}$ e o aprendizado de um neurônio $\underline{i}$ da camada $\underline{l}$ de um Perceptron multi-camadas não-linear.

Para o filtro adaptativo:

$$\mathbf{H}_i(k+1) = \mathbf{H}_i(k) + \mu . e(k) . \mathbf{X}_i(k) \qquad (4.14a)$$

Para o neurônio Perceptron (vide seção 3.3):

$$\mathbf{W}_i(k+1,l) = \mathbf{W}_i(k,l) + \mu_p . \delta_i(k,l) . \mathbf{X}_i(k,l) \qquad (4.14b)$$

Onde se define [4]:

$$\delta_i(k,l) \triangleq \frac{\partial\, e^2(k)}{\partial\, s_i(k,l)} \qquad (4.15a)$$

$$\delta_i(k,l) = f'(s_i(k,l)) \cdot \sum_{j=1}^{N_{l+1}} \delta_j(k,l+1) . w_{ij}(k) \qquad (4.15b)$$

A analogia entre as equações (4.14a-b) é evidente. Em particular, $\delta_i(k,l)$ pode ser associado a $e(k)$, embora esta primeira grandeza não seja exatamente um erro. Na realidade, de acordo com a definição da eq. (4.15a), $\delta_i(k,l)$ corresponde à sensitividade do erro quadrático relativamente ao sinal de saída linear do neurônio ($s_i(k,l)$). Deve-se notar que $\delta_i(k,l)$ pode ser calculado pela eq. (4.15b), onde se considera as interações entre o sinal $s_i(k,l)$ e todos os demais nós da rede que o processam.

Portanto, a adaptação de um neurônio i qualquer, situado na camada l de um Perceptron multi-camadas, deve levar em consideração dois elementos (que representam, na realidade, dois tipos distintos de influência sobre o erro quadrático do sistema):

- O prório nó i, considerado isoladamente;

- O processamento paralelo distribuído, resultante da interação entre o neurônio i e todos os demais nós das camadas l+1, l+2, , ..., até à última.

Em contrapartida, a adaptação do filtro adaptativo considera somente a influência de uma única estrutura (ou seja, o próprio filtro) no erro quadrático do sistema, caso em que se pode identificar precisamente o agente causador do erro de saída e(k).

Deve-se notar que, para o neurônio da última camada de um Perceptron multi-camadas qualquer, as eqs. (4.14a-b) igualam-se, pois o nó tem acesso a um sinal de referência (o qual não pode ser definido diretamente para os neurônios escondidos).

A lentidão de aprendizado através do algoritmo de retropropagação pode ser justificada, de certa forma, pela complexidade computacional do cálculo de $\delta_i(k,l)$ e pelo fato deste representar uma espécie de "aproximação indireta", através da regra da cadeia, do erro associado ao i-ésimo neurônio escondido da camada l (e não o erro exato, como no caso do algoritmo do gradiente estocástico).

A8) Algoritmos Adaptativos para a Predição Linear e as Equações de Aprendizado Auto-Organizativo [5].

O treinamento não-supervisionado do neurônio de Kohonen (ou Unidade Adaptativa Básica de Memória, seção 3.5) pode ser expresso, de forma genérica, pela seguinte equação [5]:

$$\mathbf{M}_i(k+1) = \mathbf{M}_i(k) + \phi(.).\mathbf{X}(k) - \nu(.).\mathbf{M}_i(k) \qquad (4.16)$$

Esta expressão é denominada "Equação Generalizada de Adaptação Auto-Organizativa" [5], onde $\mathbf{M}_i(k)$ é o vetor de pesos sinápticos do neurônio de Kohonen e $\mathbf{X}(k)$, o vetor de dados de entrada. Os escalares $\phi(.)$ e $\nu(.)$ representam funções quaisquer de $\mathbf{M}_i(k)$, $\mathbf{X}(k)$ e de $\eta_i(k)$ (saída do nó de Kohonen). Estudou-se em [5] a convergência dos pesos sinápticos deste neurônio quando treinado por vários casos da eq. (4.16), que correspondem a diferentes escolhas para $\phi(.)$ e $\nu(.)$. Merecem destaque as seguintes situações:

$$\Delta\mathbf{M}_i(k) \triangleq \mathbf{M}_i(k+1) - \mathbf{M}_i(k) \qquad (4.17a)$$

$$\eta_i(k) = \mathbf{M}_i(k)^T.\mathbf{X}(k) \qquad (4.17b)$$

$$\eta_i(k) = \mathbf{X}(k)^T.\mathbf{M}_i(k) \qquad (4.17c)$$

Caso 1: $\Delta\mathbf{M}_i(k) = -\alpha.\eta_i(k).\mathbf{X}(k)$ (4.18a)

Caso 2: $\Delta\mathbf{M}_i(k) = -\alpha.\eta_i(k).\mathbf{X}(k) - \beta.\mathbf{M}_i(k)$ (4.18b)

Caso 3: $\Delta\mathbf{M}_i(k) = \alpha.\mathbf{X}(k).[\eta_i(k)] - \beta.\mathbf{M}_i(k).[\eta_i(k)^2]$ (4.18c)

$\alpha > 0$ (4.19a)

$\beta > 0$, (4.19b)

onde α e β são constantes positivas por hipótese. Os vetores $\mathbf{M}_i(k)$ e $\mathbf{X}(k)$ possuem dimensão N.

Kohonen [5] analisa matematicamente a convergência da unidade adaptativa básica de memória treinada por estes algoritmos. Demonstra os seguintes teoremas para a adaptação, respectivamente, com o caso 3 (eq. (4.18c)) e com o caso 1 (eq. (4.18a)).

TEOREMA 4.1 [5]

"Seja $\mathbf{X}(k)$ um processo estocástico estacionário, tal que $\mathbf{Q}_{MAX}$ represente o autovetor de sua matriz de autocorrelação $\mathbb{R}_x$ associado ao autovalor de máxima magnitude λ_{MAX}, de acordo com as seguintes equações:

$$\mathbb{R}_x \triangleq E[\mathbf{X}(k).\mathbf{X}(k)^T] \qquad (4.20a)$$

$$\mathbb{R}_x.\mathbf{Q}_{MAX} = \lambda_{MAX}.\mathbf{Q}_{MAX} \qquad (4.20b)$$

A apresentação contínua de $\mathbf{X}(k)$ ao neurônio de Kohonen, adaptado de acordo com a eq. (4.18c) - ou caso 3 -, é tal que em regime permanente tem-se:

$$\lim_{k\to\infty} \mathbf{M}_1(k) = (\alpha/\beta)^{1/2}.(\mathbf{Q}_{max}/\|\mathbf{Q}_{max}\|) \qquad (4.21a)$$

$$\mathbf{Q}_{max}^T.\mathbf{M}_1(k=0) \neq 0 \qquad (4.21b)$$

Onde a eq. (4.21b) representa a condição necessária para a validade do resultado da eq. (4.21a)."

TEOREMA 4.2

"Seja um conjunto limitado de vetores $\mathbf{X}(k)$, com características estatísticas estacionárias. Se $\|\mathbf{X}(k)\|_{MAX}$ representa a máxima norma deste conjunto, o treinamento contínuo do neurônio de Kohonen através da eq. (4.18a) - ou Caso 1 - converge se e somente se:

$$\alpha \leq 2/(\|\mathbf{X}(k)\|^2_{MAX}) \qquad (4.22a)"$$

As hipóteses do teorema 4.2 não são alteradas se considerarmos que o conjunto de vetores $\mathbf{X}(k)$ seja formado pelo agrupamento de amostras adjacentes de um processo aleatório estacionário x(k). Neste caso, então:

$$\|\mathbf{X}(k)\|^2_{MAX} \cong N \;.\; E[x(k)^2] = N.\sigma_x^2$$

E a condição da eq. (4.22a) pode ser reescrita como:

$$\alpha \leq 2/(N.\sigma_x^2) \qquad (4.22b)$$

Resumindo, o aprendizado do neurônio de Kohonen através do

caso 3 (eq. (4.18c)) propicia uma representação estatística do estímulo externo $\mathbf{X}(k)$ em termos do autovetor máximo $\mathbf{Q}_{MAX}$ (eq. (4.21a)), sob a condição da eq. (4.21b). O treinamento com o caso 1 (eq. (4.18a)) converge sob a restrição imposta pelas eqs. (4.22a-b), sendo que o comportamento assintótico do nó de Kohonen, neste caso, corresponde ao de um sistema "detector de novidades" (definido no terceiro parágrafo da analogia A6, seção 4.2).

Seja um filtro adaptativo transversal de ordem N, aplicado à predição linear de um processo estocástico x(k) (que corresponde, alternativamente, ao processamento de uma entrada vetorial $\mathbf{X}(k)$ formada a partir do agrupamento de amostras de x(k) - vide fig. 4.2(b) -). Apresentam-se logo abaixo alguns algoritmos adaptativos convencionais para o treinamento do filtro, cuja saída (erro de predição linear progressivo de ordem N) e cujos coeficientes são respectivamente denotados por f(k) e $\mathbf{A}(k)$.

$$d(k) = 0 \tag{4.23a}$$

$$\Delta\mathbf{A}(k) = \mathbf{A}(k+1) - \mathbf{A}(k) \tag{4.23b}$$

Algoritmo do Gradiente Estocástico:

$$\Delta\mathbf{A}(k) = - \mu . f(k) . \mathbf{X}(k) \tag{4.24a}$$

Algoritmo do Gradiente Estocástico com termo de Momento W:

$$\Delta\mathbf{A}(k) = - \mu . f(k) . \mathbf{X}(k) + W^{-k} . \mathbf{A}(k) \tag{4.24b}$$

Algoritmo de Newton Estocástico:

$$\Delta\mathbf{A}(k) = \mu . \hat{\mathbb{R}}_x(k)^{-1} . [\hat{\mathbb{R}}_x(k) . \mathbf{A}(k)] \tag{4.24c}$$

$$0.9 \leq W \leq 0.99 \tag{4.25a}$$

$$\hat{\mathbb{R}}_x(k) \cong \mathbf{X}(k) . \mathbf{X}(k)^T \tag{4.25b}$$

Onde d(k) é o sinal de referência (considerado nulo) e W, o "fator de esquecimento" [17] ou "termo de momento" [18].

Para evidenciar a semelhança existente entre as eqs. (4.24a-c) e as eqs. (4.18a-c), utiliza-se a analogia A6 da seção 4.2. As grandezas $\mathbf{A}(k)$, $\Delta\mathbf{A}(k)$ e $f(k)$ do filtro de erro de predição linear correspondem respectivamente a $\mathbf{M}_i(k)$, $\Delta\mathbf{M}_i(k)$ e $\eta_i(k)$ do neurônio de Kohonen.

Neste contexto, as eqs. (4.24a-b) e (4.18a-b) são respectivamente análogas, e a comparação entre ambas sugere que β pode ser associado ao fator de esquecimento variável W^{-k}, enquanto que α representa um passo de adaptação constante. De fato, a condição de convergência da eq. (4.22b) corresponde exatamente à restrição tradicionalmente imposta ao passo de adaptação μ para a convergência do algoritmo do gradiente estocástico (vide eq. (3.12)).

A analogia entre as eqs. (4.24c) e (4.18c) não é evidente. Para ressaltá-la, a eq. (4.18c) é reescrita considerando-se que a saída do neurônio de Kohonen, $\eta_i(k)$, é um escalar. Usando-se as eqs. (4.17b-c) e reagrupando o produto dos vetores $\mathbf{M}_i(k)$ e $\mathbf{X}(k)$, tem-se:

$$\Delta\mathbf{M}_i(k) = \alpha.\mathbf{X}(k).[\mathbf{X}(k)^T.\mathbf{M}_i(k)] - \beta.\mathbf{M}_i(k).[\mathbf{M}_i(k)^T.\mathbf{X}(k)].[\mathbf{X}(k)^T.\mathbf{M}_i(k)]$$

$$\Delta\mathbf{M}_i(k) = \alpha.\left\{ [\mathbf{X}(k).\mathbf{X}(k)^T].\mathbf{M}_i(k) \right\} - \beta.\mathbf{M}_i(k).\mathbf{M}_i(k)^T.\left\{[\mathbf{X}(k).\mathbf{X}(k)^T].\mathbf{M}_i(k)\right\}$$

Finalmente, a última equação é reescrita relembrando-se a aproximação da eq. (4.25b):

$$\Delta\mathbf{M}_i(k) = \left[\alpha.\mathbb{I}_N - \beta.\mathbf{M}_i(k).\mathbf{M}_i(k)^T\right] . \left\{\hat{\mathbb{R}}_x(k).\mathbf{M}_i(k)\right\} \qquad (4.26a)$$

$$\Delta\mathbf{M}_i(k) = \alpha.\mathbb{F}(k).\left\{\hat{\mathbb{R}}_x(k).\mathbf{M}_i(k)\right\} \qquad (4.26b)$$

$$\mathbb{F}(k) = \left[\mathbb{I}_N - (\beta/\alpha).\mathbf{M}_i(k).\mathbf{M}_i(k)^T\right] \qquad (4.26c)$$

Onde $\mathbb{I}_N$ é a matriz identidade de dimensão N x N. A eq. (4.26b), forma alternativa do caso 3 (eq. (4.18c)), é portanto aproximadamente análoga à eq. (4.24c). Deve-se notar que, embora $\mathbb{F}(k)$ (eqs. (4.26b-c)) não seja exatamente igual a $\mathbb{R}_x(k)^{-1}$ da eq. (4.24c), ambas as matrizes são simétricas.

Desta forma, demonstrou-se que as equações de aprendizado auto-organizativo (4.18a-c) são análogas aos respectivos algoritmos adaptativos supervisionados das eqs. (4.24a-c), aplicados à predição linear.

Particularmente, a saída do neurônio de Kohonen, treinado pelo caso 1 (eq. (4.18a)), tende a zero em regime permanente, mas não representa necessariamente um processo inovação, como no caso do erro de predição linear associado ao filtro adaptativo, treinado pelo algoritmo do gradiente estocástico. Se o aprendizado do neurônio de Kohonen ocorre através do caso 3 (eq. (4.18c)), seus coeficientes representam o processo estocástico estacionário de entrada através do autovetor máximo $\mathbf{Q}_{MAX}$ (eq. (4.21a)). Em contrapartida, os coeficientes a_1 do filtro de erro de predição linear, treinado pelo algoritmo de Newton estocástico (eq. (4.24c)), convergem para a representação paramétrica auto-regressiva do sinal de entrada x(k).

A discussão do parágrafo anterior sugere que o neurônio de Kohonen, treinado pelos algoritmos neurais das eqs. (4.18a-c), pode ser considerado como um "filtro de erro de predição linear auto-organizativo", que utiliza um conjunto finito de padrões de aprendizado (ou segmento limitado de amostras de uma série temporal) e cujos pesos sinápticos, em regime permanente, armazenam conhecimento sobre o estímulo externo de forma possivelmente interrelacionada à representação paramétrica tradicional. Estas hipóteses foram analisadas experimentalmente em [57], onde um neurônio de Kohonen de ordem N = 4, treinado pela eq. (4.18a) - caso 1 -, foi simulado para uma entrada x(k) do tipo processo aleatório AR de ordem 1. Constatou-se que as oscilações dos coeficientes do nó durante o transitório apresentam amplitudes

menores e são menos dependentes do passo de adaptação α (para uma ampla faixa de variação da magnitude deste parâmetro do algoritmo) relativamente ao filtro de erro de predição linear de ordem N = 4, treinado pelo algoritmo do gradiente estocástico.

4.4 - A EQUAÇÃO DE WIENER-HOPF E OS PESOS SINÁPTICOS ÓTIMOS DO NEURÔNIO PERCEPTRON.

Visto que um neurônio Perceptron com função de ativação linear corresponde a um filtro adaptativo transversal (analogia A1, seção 4.2), conclui-se que os pesos sinápticos ótimos $\mathbf{W}^*$ deste nó que minimizam a função de custo expressa pela eq. (4.3a) correspondem exatamente à solução de Wiener-Hopf:

$$\mathbf{W}^* = \mathbb{R}_x^{-1} \cdot \mathbf{R}_{xd} \tag{4.27a}$$

$$\mathbb{R}_x = E[\mathbf{X}(k) . \mathbf{X}(k)^T] \tag{4.27b}$$

$$\mathbf{R}_{xd} = E[\mathbf{X}(k) . d(k)] \tag{4.27c}$$

Onde $\mathbf{X}(k)$ é o vetor de dados de entrada e d(k) o sinal de referência. Supõe-se que $\mathbf{X}(k)$ e d(k) sejam conjuntamente estacionários e de média nula.

Seja agora o neurônio Perceptron mostrado pela fig. 4.8. De uma forma geral, o aprendizado desta estrutura objetiva minimizar a seguinte função de custo:

$$J = E[\ v\{s(k), d(k)\}\ ] \tag{4.28a}$$

$$v\{a\} \leq \exp\{a\}\ \forall\ a \in \mathbb{R}, \tag{4.28b}$$

onde $v\{.\}$ corresponde a uma função não-linear qualquer, porém limitada por uma exponencial crescente (eq. (4.28b)), e que leva em consideração a função de ativação do neurônio. Deve-se notar que as funções de custo das eqs. (4.2a), (4.2c), (4.3a), (4.3c) correspondem a casos específicos da eq. (4.28a).

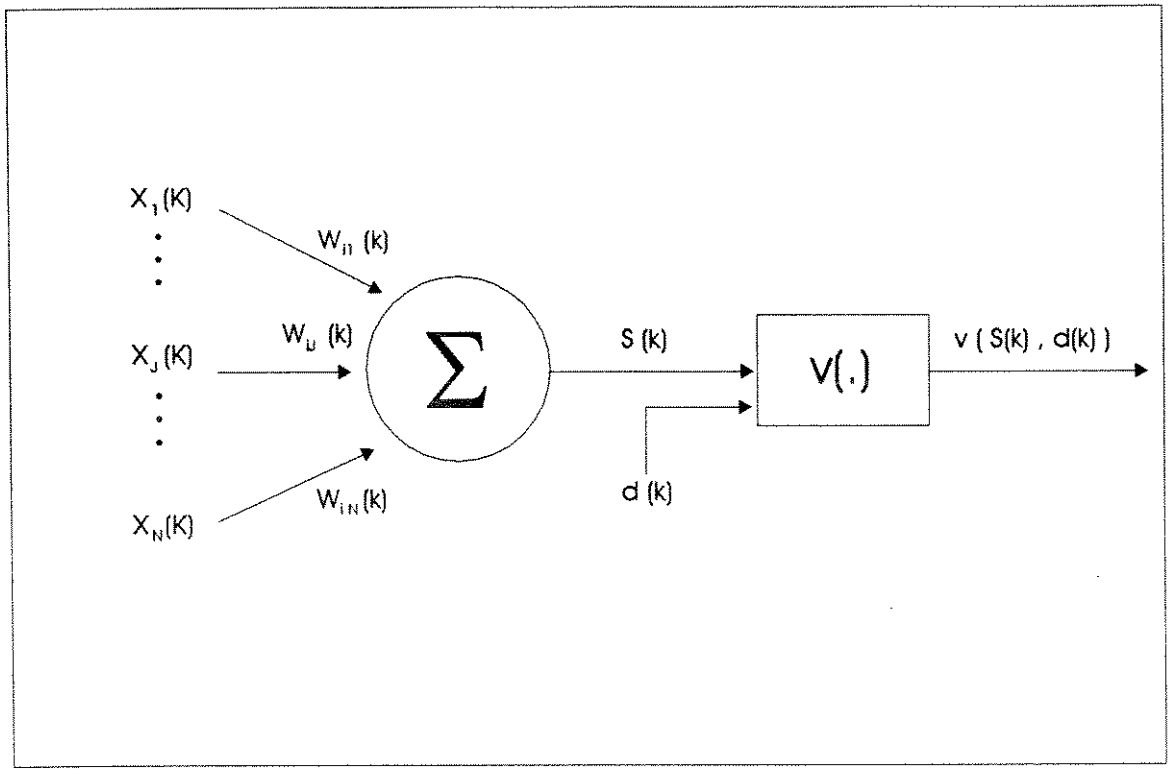


Figura 4.8 : Estrutura do neurônio Perceptron para a definição da função de custo generalizada J - eq. (4.28a) - [54].

A referência [54] demonstra que, se $X(k)$ e $d(k)$ representarem processos aleatórios conjuntamente estacionários e gaussianos, o vetor de pesos sinápticos ótimos do Perceptron que minimiza a função de custo J (eq. (4.28a)), se existir, é sempre colinear à solução de Wiener-Hopf:

$$W^* = \alpha . R_x^{-1} . R_{xd} \qquad (4.29)$$

Onde α é uma constante que depende de $v\{s(k),d(k)\}$ e da função de densidade de probabilidade conjunta de s(k) e de d(k), especificada em [54]. Os autores apresentam uma expressão que corresponde a uma condição suficiente para a validade da eq. (4.29), que também depende de $v\{s(k),d(k)\}$, da função de densidade de probabilidade conjunta, de α e das matrizes R_x^{-1} e R_{xd}.

Em particular, para o modelo do Perceptron estabelecido por Rosenblatt [15] (eqs. (3.7a-b), seção 3.2), [54] demonstra que os pesos sinápticos ótimos da eq. (4.29) sempre existem, sendo que α pode assumir qualquer valor positivo e não-nulo.

4.5 - CONCLUSÃO

As principais analogias e diferenças entre redes neurais e a filtragem adaptativa foram analisadas neste capítulo.

Em termos do treinamento supervisionado e da aplicação ao reconhecimento de padrões, redes neurais são caracterizadas, em regime permanente, por sua elevada GENERALIZAÇÃO. Portanto, são capazes de classificar corretamente estímulos externos extremamente complexos, o que é alcançado através de operação não-adaptativa (fases de adaptação, teste e utilização) e de um conjunto limitado de vetores de treinamento. Por outro lado, filtros adaptativos correspondem a sistemas de elevada ADAPTATIVIDADE. Ou seja, são capazes de acompanhar eficientemente as variações temporais da estrutura estatística do sinal de

entrada (que é geralmente uma série temporal), através de aprendizado contínuo. Possuem, portanto, generalização reduzida.

Neste contexto, o treinamento neural implica em rigorosa seleção dos vetores de aprendizado (que devem ser representativos das características estatísticas dos estímulos externos, em geral supostas estacionárias), bem como determinação da quantidade e da sistemática de apresentação à rede neural de tais vetores. Esta preocupação não ocorre para o aprendizado de filtros adaptativos. Para uma rede neural, os pesos sinápticos w_{ij} não possuem significado matemático, e seus valores ótimos não podem ser teoricamente calculados. Entretanto, demonstra-se que, sob determinadas restrições, a equação de Wiener-Hopf representa o valor ótimo para os pesos sinápticos de um neurônio Perceptron [54].

Sumarizam-se a seguir as principais analogias apresentadas neste capítulo.

Analogias envolvendo redes neurais supervisionadas:

I) Filtro adaptativo não-linear ou linear, IIR ou FIR ⇔ neurônio Perceptron e Perceptron multi-camadas com função de ativação não-linear ou linear, dinâmicos ou estáticos [22,23].

II) Filtro de pilha ⇔ neurônio Perceptron (função de ativação sinal e pesos sinápticos sempre positivos) [25,56].

Filtro mediano ⇔ Perceptron multi-camadas (função de ativação sinal).

III) Filtro espacial adaptativo ⇔ Perceptron multi-camadas.

IV) Algoritmo do gradiente estocástico ⇔ algoritmo de treinamento por retropropagação .

Analogias envolvendo redes neurais não-supervisionadas:

V) Filtro adaptativo transversal ⇔ neurônio de Kohonen.

VI) Equalizador adaptativo de Bussgang ⇔ neurônio Perceptron.

VII) Algoritmos adaptativos supervisionados aplicados ao treinamento de um filtro de erro de predição linear transversal ⇔ equações de aprendizado auto-organizativo do neurônio de Kohonen.

Deve-se destacar que a função de ativação está associada à definição de função de limiar (item II) [25] e ao estimador de Bussgang (item VI) [10].

Considerando-se a quantidade de analogias estabelecidas e, em particular, aquelas entre filtros adaptativos e os neurônios Perceptron e de Kohonen (unidades básicas componentes de todas as arquiteturas neurais existentes), conclui-se que o formalismo matemático associado à filtragem adaptativa pode estar de certa forma relacionado ao processamento de informação neural. Entretanto, deve-se ressaltar que todos os algoritmos e estruturas da filtragem adaptativa referenciados foram desenvolvidos a partir de teorias (como processos estocásticos, teoria de otimização, de comunicações, filtragem linear de Wiener, etc [10,17,18]) desvinculadas dos conceitos biológicos que fundamentam a definição de redes neurais, pelo menos aparentemente.

Em conclusão, demonstrou-se neste capítulo como o formalismo matemático que embasa a filtragem adaptativa permite relacioná-la a redes neurais, ou seja, como referenciar estruturas neurais a partir de filtros adaptativos. No próximo capítulo, será realizado o processo inverso, através do estudo da filtragem adaptativa com base nos princípios que fundamentam redes neurais.

CAPÍTULO 5

A FILTRAGEM ADAPTATIVA NO CONTEXTO DE REDES NEURAIS

Analisam-se neste capítulo conceitos, estruturas e algoritmos associados à filtragem adaptativa sob o ponto de vista dos princípios básicos que fundamentam a definição de redes neurais, com o objetivo de evidenciar a interrelação espontânea existente entre os dois campos. Inicialmente, compara-se a operação de filtros adaptativos e o algoritmo do gradiente estocástico (aplicado à predição linear) ao funcionamento do sistema nervoso vertebrado e à sinapase de Hebb [20], respectivamente. Em seguida, evidencia-se como alguns filtros adaptativos realizam processamento paralelo distribuído e como o princípio da mínima perturbação fundamenta o algoritmo do gradiente estocástico, o que possibilita analisar as propriedades coletivas emergentes rudimentares destas estruturas e algumas aplicações a tarefas cognitivas simples. A desconvolução cega e a predição linear são comparadas à auto-organização. Finalmente, a filtragem adaptativa é analisada no contexto da psicologia cognitiva.

5.1 - A FILTRAGEM ADAPTATIVA NO CONTEXTO DA NEUROFISIOLOGIA.

O SISTEMA NERVOSO VERTEBRADO E A FILTRAGEM ADAPTATIVA

A característica fundamental do sistema nervoso dos animais vertebrados, que possibilita a execução eficiente da complexa tarefa cognitiva "sobrevivência do organismo", é sua capacidade de interagir com o mundo real. Ele consegue processar rapidamente sinais naturais (tanto aqueles do meio-ambiente quanto os internos à própria estrutura biológica), que consistem geralmente em processos estocásticos multivariáveis, não-estacionários, quase sempre perturbados por ruído. O sistema nervoso deve ser capaz de

complementar e de corrigir informações decorrentes de observações parciais (por exemplo, visão noturna), além de antecipar decisões com base nas condições atuais de sobrevivência (por exemplo, movimento migratório de aves durante a mudança de estação, fuga de predadores) e de economizar os recursos disponíveis para o organismo. (De fato, constata-se experimentalmente [5] que o ciclo respiratório dos animais é estatisticamente quase ótimo, em termos da utilização de energia para desempenho desta função).

Conclui-se, portanto, que os "recursos computacionais" do sistema nervoso vertebrado dedicam-se sobretudo à realização de filtragem estatística, através da eliminação do ruído que afeta os sinais sensoriais, da estimação e da predição de eventos, operações que possibilitam a otimização do uso de recursos e do funcionamento do complexo sistema "organismo". Analogamente, tais operações também são realizadas por filtros adaptativos, os quais objetivam minimizar uma determinada função de custo.

A SINAPSE DE HEBB E O ALGORITMO DO GRADIENTE ESTOCÁSTICO APLICADO À PREDIÇÃO LINEAR

Kohonen [5] propõe uma formulação estocástica para a lei de Hebb. Seja a sinapse existente entre a terminação j do axônio de uma célula nervosa A e o j-ésimo dendrito de um neurônio B, conforme mostrado na fig. 5.1. Define-se, para o instante k:

$y_B(k)$: evento atividade neural de saída elevada (acima de um determinado limiar) da célula B.

$x_{j,A}(k)$: evento excitação externa j, proveniente de A, de amplitude elevada (acima de um determinado limiar).

α_{AB} : coeficiente de plasticidade [20] ("adaptatividade") da sinapse A-B.

$w_{AB}(k)$: eficiência da sinapse A-B.

$\Delta w_{AB}(k)$: variação temporal da eficiência da sinapse A-B.

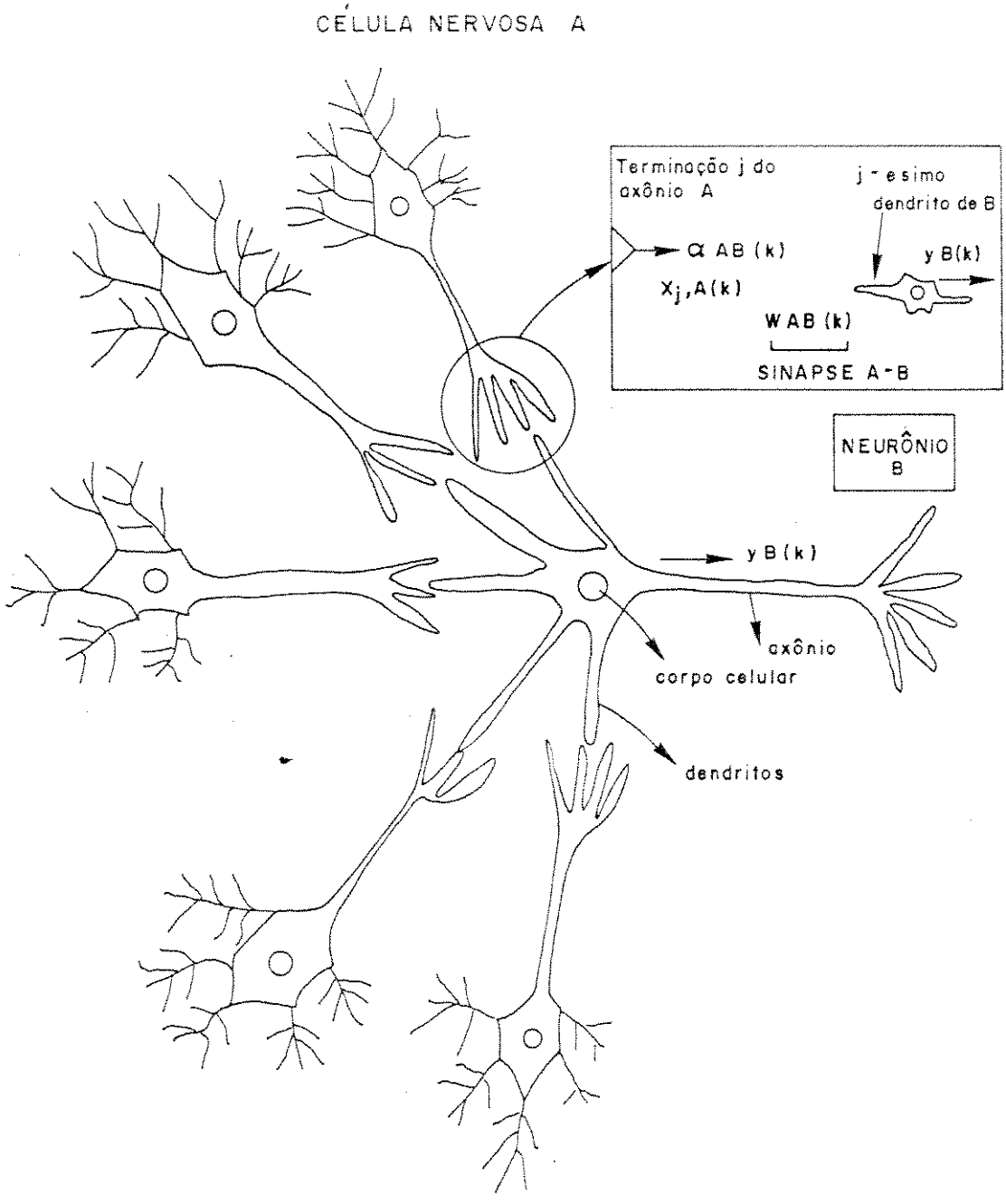


Figura 5.1 : Sinapse de Hebb [20].

$$\Delta w_{AB}(k) = w_{AB}(k+1) - w_{AB}(k) \qquad (5.1)$$

O coeficiente de plasticidade α_{AB} evidencia o quanto a sinapse A-B é susceptível a variações do seu mecanismo bioquímico controlador, ou seja, quanto menor o valor de α_{AB}, mais estável são os parâmetros bioquímicos que regulam esta sinapse. De acordo com a seção 2.3, a eficiência $w_{AB}(k)$ é uma grandeza diretamente proporcional ao grau de influência do sinal de excitação proveniente da célula A sobre a atividade neural de saída de B. Quanto maior a eficiência da sinapse (ou seja, quanto maior $w_{AB}(k)$), maior esta influência em relação aos demais sinais excitatórios que atingem B, provenientes de outras células nervosas; bem como maior a probabilidade de B apresentar atividade neural de saída sempre que for estimulada pela célula A.

A lei de Hebb [20] afirma que (vide seção 2.3) "a sinapse A-B será cada vez mais eficiente (ou seja, $\Delta w_{AB}(k)$ será positivo) à medida que um processo simultâneo de excitação externa intensa, proveniente de A, gerando atividade neural de saída intensa em B, persistir repetidamente" (ou seja, toda vez que o produto $x_{j,A}(k).y_B(k)$ for elevado durante um determinado intervalo de tempo). Consequentemente, a partir das definições anteriores, propõe-se uma possível expressão matemática para a sinapse de Hebb [5]:

$$E[\Delta w_{AB}(k)] = + \alpha_{AB}.E[x_{j,A}(k).y_B(k)] \qquad (5.2)$$

Aplicando-se o operador esperança matemática ao algoritmo do gradiente estocástico, utilizado para o treinamento de um filtro de erro de predição linear (eq. (4.24a)), otém-se a seguinte equação para a adaptação do j-ésimo coeficiente $a_j(k)$ desta estrutura:

$$E[\Delta a_j(k)] = -\mu.E[x_j(k).f(k)] \qquad (5.3)$$

Onde $x_j(k)$ é o j-ésimo elemento do vetor de dados de entrada $\mathbf{X}(k)$ e $f(k)$ a saída do sistema (erro de predição linear).

As equações (5.2) e (5.3) são análogas. Além disso, considerando-se o sistema nervoso vertebrado , o neurônio B recebe da ordem de centenas a milhares de excitações externas. É razoável, portanto, supor que os eventos $y_B(k)$ e $x_{j,A}(k)$ sejam estatisticamente independentes. De acordo com a analogia estabelecida, isto é equivalente a afirmar que o erro de predição linear f(k), em regime permanente, é estatisticamente independente do sinal de entrada x(k) do filtro adaptativo, e portanto, descorrelacionado. Tal afirmação pode ser considerada válida desde que a ordem do filtro adaptativo possibilite f(k) o mais próximo possível a um processo inovação [17].

A referência [57] compara o modelo elétrico geral do neurônio (mais próximo à realidade biológica desta estrutura que o modelo apresentado na seção 3.1) ao quantizador de sistemas de modulação por código de pulso adaptativa diferencial (ADPCM). Esta analogia, conjuntamente com as demais apresentadas nesta seção e na 4.2, evidenciam como a célula nervosa, unidade básica fundamental do sistema nervoso vertebrado, pode ser relacionada a filtros adaptativos.

5.2 - A FILTRAGEM ADAPTATIVA E OS PRINCÍPIOS BÁSICOS DE REDES NEURAIS.

O PROCESSAMENTO PARALELO DISTRIBUÍDO E A FILTRAGEM ADAPTATIVA

O conhecimento armazenado por qualquer estrutura da filtragem adaptativa a respeito do sinal de entrada x(k) é expresso, matematicamente, pelos coeficientes de regime do sistema considerado. Por exemplo, no caso de um filtro transversal, tais coeficientes correspondem às componentes h_i do vetor $\mathbf{H}$ (vide fig. 4.2(b)). Cada segmento de amostras de x(k), com características estatísticas estacionárias, está associado a um modelo paramétrico, ou seja, a um determinado conjunto de coeficientes

h_i. O aprendizado do filtro tranversal consiste em determinar os valores de h_i tais que as saídas convenientes sejam produzidas em função dos dados de entrada e da tarefa desempenhada pelo sistema. Portanto, a representação interna do sinal de entrada por QUALQUER estrutura da filtragem adaptativa é distribuída, analogamente às redes neurais. Entretanto, no contexto da classificação de padrões, enquanto um estímulo de treinamento está biunivocamente associado a um padrão de atividade neural, um mesmo conjunto de coeficientes h_i pode representar diversos segmentos de amostras de um mesmo sinal temporal - ou seja, diversos padrões de aprendizado -, desde que estes possuam a mesma caracterização estatística.

A cascata de filtros adaptativos, em particular, é um exemplo importante de processamento de informação distribuído. Por exemplo, um filtro de ordem 8 (a realização direta) pode ser implementado através da cascata de quatro filtros de ordem 2. O comportamento do sistema em cascata resulta do complexo mecanismo de interações existente entre os diversos filtros componentes da estrutura, mecanismo este que depende da sistemática de adaptação empregada. A tarefa desempenhada pelo sistema é igualmente repartida entre os filtros componentes, sendo que todos estão envolvidos na caracterização estatística do conjunto de amostras do sinal de entrada. Através da interação simultânea de filtros de ordem pequena (em comparação com a respectiva realização direta), é possível alcançar maior eficiência de operação do sistema, em termos de controle da estabilidade da função de transferência do filtro e do rápido acesso a zeros do modelo auto-regressivo do sinal de entrada [58,59], para o caso de predição linear.

O paralelismo é também um conceito intrínseco a determinadas estruturas da filtragem adaptativa, por exemplo, o filtro espacial (fig. 4.7(a), seção 4.2) e o filtro mediano [25,56] (fig. 4.6(a), seção 4.2).

Conclui-se, portanto, que o princípio do processamento paralelo distribuído pode ser associado à filtragem adaptativa,

cujas estruturas, à semelhança de redes neurais, também devem apresentar propriedades coletivas emergentes. Isto será analisado ainda nesta seção.

O PRINCÍPIO DA MÍNIMA PERTURBAÇÃO E A FILTRAGEM ADAPTATIVA

Analisa-se agora, matematicamente, como o princípio da mínima perturbação (seção 2.4) também fundamenta o algoritmo do gradiente estocástico, frequentemente empregado na filtragem adaptativa. Este algoritmo é expresso pelas seguintes equações (vide figuras 4.2(a)-(b)):

$$\mathbf{H}(k+1) = \mathbf{H}(k) + \mu . e_a(k) . \mathbf{X}(k) \qquad (5.4a)$$

$$e_a(k) = d(k) - \mathbf{X}(k)^T . \mathbf{H}(k) \qquad (5.4b)$$

$$\Delta\mathbf{H}(k) = \mathbf{H}(k+1) - \mathbf{H}(k) \qquad (5.4c)$$

$$\mu \leq \mu_{max} = 2/(N.\sigma_x^2) \qquad (5.4d)$$

Onde $\mathbf{H}(k)$ é o vetor de coeficientes h_i de um filtro adaptativo, e $e_a(k)$, o erro associado. A eq. (5.4d) apresenta uma restrição necessária para a convergência deste algoritmo, para o caso de um sinal de entrada suposto estacionário, com distribuição probabilística gaussiana e de potência média σ_x^2 [22]. A grandeza N denota a ordem do filtro transversal.

Calculando-se o diferencial total Δ do erro $e_a(k)$ a partir da eq. (5.4b), para um dado k fixado, e combinando as eqs. (5.4a) e (5.4c), tem-se:

$$\Delta e_a(k) = \Delta\,(d(k) - \mathbf{X}(k)^T . \mathbf{H}(k)) = - \mathbf{X}(k)^T . \Delta\mathbf{H}(k) \qquad (5.5a)$$

$$\Delta\mathbf{H}(k) = \mu . e_a(k) . \mathbf{X}(k) \qquad (5.5b)$$

De acordo com a última equação, $\Delta\mathbf{H}(k)$ é colinear ao vetor $\mathbf{X}(k)$. Consequentemente, isto possibilita máximo decréscimo do erro $e_a(k)$, visto que $\Delta e_a(k)$ resulta do produto escalar entre $\mathbf{X}(k)$ e

$\Delta\underline{H}(k)$ (vide eq. (5.5a)). Por outro lado, a eq. (5.5b) evidencia que a variação nos coeficientes $h_i(k)$ é controlada pela magnitude do passo de adaptação μ, limitada por sua vez pela eq. (5.4d) para garantir a convergência do algoritmo. Portanto, conclui-se que o algoritmo do gradiente estocástico propicia máximo decréscimo do erro ($\Delta e_a(k)$) através da perturbação "controlada" dos coeficientes do sistema no instante k atual ($\Delta\underline{H}(k)$), o que possibilita manter a estrutura de conhecimento anterior do sistema. Isto está em conformidade com o princípio da mínima perturbação.

Deve-se notar que, para o filtro adaptativo aqui considerado, é o próprio passo de adaptação μ que regula o compromisso redução do erro versus perturbação do conhecimento anterior.

EXEMPLOS DE TAREFAS COGNITIVAS SIMPLES REALIZADAS POR FILTROS ADAPTATIVOS

Aplicação 1: Equalização de sinais transmitidos pelo sistema de rádio digital

Esta aplicação envolve sinais e sistemas de difícil caracterização matemática e estatística. Por exemplo, o canal de comunicação atmosférico variante no tempo, os ruídos (muitas vezes não-gaussianos e não-aditivos, como a inteferência de cocanal e o desvanecimento [37]) e o próprio sinal de informação (que consiste em dados ou sinais de voz codificados, em geral não-estacionários). A decisão quanto ao valor digital transmitido deve levar em consideração um conjunto de determinadas condições ambíguas ou imperfeitamente especificadas, tais como as interrelações entre o canal de comunicação, o ruído e o sinal de informação.

Aplicação 2: Codificação de processos estocásticos por predição linear em ambiente ruidoso (por exemplo, análise LPC de sinais de voz em telefonia celular).

A obtenção dos coeficientes LPC exige o cálculo do preditor ótimo para cada quadro do sinal de voz. Isto envolve as seguintes operações:

1) Determinação da ordem do modelo paramétrico auto-regressivo, que deve ser coerente com a estrutura estatística do sinal de entrada;

2) Separação do sinal de voz em quadros, ou seja, conjunto de amostras cujas características estatísticas são supostas estacionárias;

3) Estimação da função de autocorrelação do sinal de entrada;

4) Emprego do algoritmo de Levinson-Durbin [10,17,18].

Deve-se notar que a não-estacionariedade do sinal de voz e do ruído de fundo, bem como os efeitos causados pelas interações existentes entre estes dois sinais, dificultam a realização das operações (1)-(3), podendo implicar em baixo desempenho.

Tanto a aplicação 1 (abreviada aqui como A1) como a aplicação 2 (A2) envolvem as três características fundamentais das tarefas cognitivas (vide seção 2.5):

- geração de decisões para o caso A1 (ou de dados de saída para A2) em tempo real;
- dados de entrada de difícil caracterização matemática;
- quantidade relativamente elevada de cálculo computacional.

Entretanto, deve-se notar que tais tarefas são pouco complexas se comparadas com as aplicações tradicionais de redes neurais, apresentadas nas seções 2.5, 2.6 e 3.7.

A filtragem adaptativa representa uma alternativa viável e eficiente para a execução das aplicações citadas, visto que ela é

capaz de, sob determinadas restrições, manipular sinais não-estacionários ou de caracterização matemática desconhecida. O treinamento contínuo através de um algoritmo adaptativo permite levar em consideração todas as múltiplas condições ambíguas da tarefa cognitiva A1, já que o filtro armazena toda a "história passada" - ou seja, o estado do sistema e o conhecimento sobre o sinal externo - através da recursão. Quanto à aplicação A2, a adaptação permanente do filtro permite acompanhar as variações da estatística do sinal de entrada, evitando a necessidade de divisão do conjunto de amostras do sinal em "blocos de estacionariedade".

AS PROPRIEDADES COLETIVAS EMERGENTES E A FILTRAGEM ADAPTATIVA

Considerando-se que o processamento de informação por filtros adaptativos é distribuído (e, em alguns casos, paralelo), espera-se que estas estruturas apresentem formas rudimentares de propriedades coletivas emergentes, mais simples e menos potentes que aquelas de redes neurais. A seguir, as principais propriedades coletivas da filtragem adaptativa serão analisadas.

1) Compressão de Informação

A compressão de informação de filtros adaptativos baseia-se na caracterização da estrutura estatística dos estímulos externos (ou do sinal de entrada) na representação paramétrica. Esta propriedade capacita filtros adaptativos a:

- Identificar e inverter sistemas com parâmetros variantes no tempo, por exemplo, cancelamento de eco em transmissão via satélite ou equalização de transmissão digital [17,18];

- Analisar sinais não-estacionários, por exemplo, análise de voz por predição linear [10];

- Analisar sinais multidimensionais, por exemplo, filtragem espacial em processamento de sinais geofísicos [10,17];

- Compactar informação, por exemplo, quantização ADPCM de sinais de voz ou codificação de sinais de vídeo [55].

Deve-se notar, entretanto, que tais aplicações quase sempre supõem ruído gaussiano e aditivo [25]. Em particular, para a equalização, técnicas adaptativas clássicas são ineficientes para o caso de canal de comunicação de fase não-mínima ou de relação sinal-ruído reduzida [38].

2) Recuperação espontânea do sistema

Um filtro adaptativo é capaz de retornar ao regime permanente após a modificação aleatória de seus coeficientes, no contexto de aprendizado contínuo através de um sinal de referência externo. Entretanto, este procedimento não garante o mesmo desempenho do sistema para alterações estruturais, como por exemplo, diminuição da ordem do filtro.

Em conclusão, constata-se que redes neurais são especilizadas na execução de tarefas cognitivas, através do processamento paralelo distribuído baseado em uma arquitetura constituída por múltiplos neurônios. Apresentam diversas propriedades coletivas emergentes e geram resultados sub-ótimos do ponto de vista teórico.

Filtros adaptativos são eficientemente aplicados a tarefas que envolvem sinais de entrada não-estacionários e ruidosos (sobretudo nos casos de ruído gaussiano e aditivo), além de sistemas com parâmetros variantes no tempo. O processamento de informação é sempre distribuído, de forma serial (por exemplo, cascata) ou paralela (por exemplo, filtro espacial), realizado através de circuitos digitais dedicados ao processamento de sinais. Apresentam propriedades coletivas emergentes rudimentares, sobretudo compressão de informação, e geram resultados sub-ótimos sob o ponto de vista teórico.

Computadores são especializados na resolução de problemas matemáticos precisamente definidos, através de processamento serial (ou paralelo, em alguns casos) baseado em circuitos digitais extremamente rápidos, se comparados a neurônios biológicos. Não apresentam propriedades coletivas emergentes e geram resultados ótimos sob o ponto de vista teórico.

Conclui-se, portanto, que filtros adaptativos apresentam características "intermediárias" entre computadores e redes neurais.

5.3 - A FILTRAGEM ADAPTATIVA E A AUTO-ORGANIZAÇÃO

O MAPA AUTO-ORGANIZATIVO DE KOHONEN E A DESCONVOULUÇÃO CEGA

Os mapas de Kohonen são capazes de separar e classificar estímulos externos complexos. Isto é possível graças à realimentação lateral múltipla, uma espécie de "não-linearidade" sistêmica (visto que este procedimento quantiza o sinal de saída dos nós - seção 3.5); e à aprendizagem competitiva, responsável pelo mapeamento da função de densidade de probabilidade do estímulo externo no conjunto de pesos sinápticos da rede.

Tais características do mapa de Kohonen referenciam conceitos intrínsecos da desconvolução cega (inversão adaptativa de sistemas sem acesso a um sinal de referência). Por exemplo, a equalização cega em sistemas de transmissão digital corresponde a uma classificação do sinal de entrada do receptor, com o objetivo de decodificar a informação transmitida através de canais de comunicação ruidosos. Isto envolve a utilização de não-linearidades (caso do equalizador de Bussgang [10], seção 4.2) ou de momentos estatísticos de até quarta ordem (caso da desconvolução cega através de algoritmos baseados no "tricepstrum" [10], utilizados para estimar a função de densidade de probabilidade do sinal de entrada do equalizador).

O NEURÔNIO DE KOHONEN E A PREDIÇÃO LINEAR

Conforme estudado na seção 4.3, o treinamento do nó de Kohonen através da equação generalizada de adaptação auto-organizativa (eq. (4.16)) propicia, para o caso 3 (eq. (4.18c)), analisar o estímulo externo em termos dos autovetores de sua matriz de autocorrelação. O treinamento através do caso 1 (eq. (4.18a)) confere ao neurônio a característica de sistema "detector de novidades". Isto significa que o nó de Kohonen apresenta saída não-nula em regime permanente apenas se o estímulo de entrada não representar combinação linear dos padrões de treinamento (ou seja, se o estímulo "não possuir" as mesmas características daqueles de treinamento).

De acordo com a analogia A8 da seção 4.3, os diversos algoritmos da filtragem adaptativa aplicados à predição linear constituem casos especiais da equação generalizada de aprendizado auto-organizativo. Além disso, tais algoritmos possibilitam analisar o sinal de entrada com base no modelamento paramétrico. Particularmente, o filtro de erro de predição linear (utilizado em análise LPC [55]) também apresenta, de forma aproximada, a característica de sistema "detector de novidades" em regime permanente.

Portanto, conclui-se que a predição linear (que corresponde ao treinamento de um filtro adaptativo quando o sinal de referência é considerado nulo) corresponde a um caso especial de auto-organização, a qual também está associada à desconvolução cega.

5.4 - A FILTRAGEM ADAPTATIVA NO CONTEXTO DA PSICOLOGIA COGNITIVA

O estudo do comportamento do ser humano em sociedade constitui um dos principais objetivos da psicologia [36,60]. Em particular, deve-se destacar a análise dos mecanismos psicológicos

responsáveis pelas reações do indivíduo em resposta a estímulos do meio-ambiente (ou estímulos externos), mecanismos estes diretamente envolvidos no processo de aprendizagem.

O cognitivismo de Piaget [21] sustenta que a reação do indivíduo ao estímulo depende de sua estrutura biológica e de seu conhecimento anteriormente armazenado, o qual também se modifica em função do estímulo externo atual.

Seja agora um filtro adaptativo treinado pelo algoritmo do gradiente estocástico, expresso pelas seguintes equações:

$$\mathbf{H}(k+1) = \mathbf{H}(k) + \mu.e(k).\mathbf{X}(k) \qquad (5.8a)$$

$$y_a(k) = \mathbf{X}(k)^T.\mathbf{H}(k) \qquad (5.8b)$$

Se o sinal de entrada $\mathbf{X}(k)$ do filtro adaptativo for associado ao "estímulo externo" de Piaget, então o sinal de saída do sistema $y_a(k)$ (vide fig. 4.2(a)) pode ser considerado como a "reação" do filtro a $\mathbf{X}(k)$. Neste contexto, $\mathbf{H}(k)$ representa "o conhecimento anteriormente armazenado", visto que os coeficientes acumulam toda a informação do sistema acerca do sinal de entrada, através do aprendizado contínuo. Desta forma, a eq. (5.8b) representa simplesmente uma expressão matemática da proposição "a reação depende do conhecimento armazenado", enquanto que a eq. (5.8a) exprime o fato do "conhecimento armazenado modificar-se em função do estímulo atual" ($\mathbf{X}(k)$).

Por outro lado, se "a reação do indivíduo depende do conhecimento anteriormente armazenado" [21], conclui-se que a memória corresponde a uma das bases fundamentais dos processos psicológicos. Conforme discutido na seção 2.3, a memória também representa, de certa forma, a influência do meio social sobre o funcionamento do sistema nervoso de um organismo.

Em termos biológicos [5], a memória caracteriza-se por participar de quase todas as funções cerebrais, de forma espacialmente distribuída. Ela pode ser considerada, portanto,

como um "sistema" que realiza processamento distribuído.

Uma das funções da memória corresponde a compactar, de forma simplificada, a informação decorrente do processamento de estímulos do meio-ambiente. A recuperação dos dados por ela armazenados ocorre pela identificação do padrão de atividade neural que apresentar máxima similaridade ao estímulo externo, o qual está biunivocamente associado à informação desejada. Este mecanismo está em conformidade com a propriedade coletiva de processamento associativo (seção 2.5, propriedade (3)) e é tratado pela psicologia sob a denominação de "paradigma da memória associativa" [61].

Conforme discutido na seção 5.2, o processamento de filtros adaptativos também é distribuído, sendo que a informação decorrente do tratamento da entrada externa pelo filtro está "compactada" no conjunto de coeficientes da estrutura, representativo de um modelo paramétrico. (Em particular, um filtro de erro de predição linear transforma o "estímulo de entrada" complexo em um processo inovação, matematicamente mais simples). Além disso, deve-se notar que a saída de um filtro adaptativo corresponde a uma função do produto escalar entre o vetor de dados de entrada $\mathbf{X}(k)$ e o vetor de coeficientes $\mathbf{H}(k)$ (por exemplo, vide eq. (5.8b)). Se o produto escalar for considerado como uma aproximação da correlação entre $\mathbf{X}(k)$ e $\mathbf{H}(k)$, e interpretando-se a correlação como uma "medida de similaridade", conclui-se que a saída do filtro adaptativo representa uma "estimação" da similaridade entre o estímulo externo e o conhecimento do sistema. Desta forma, analogamente à memória, o processamento desta estrutura também pode ser considerado associativo.

Um filtro adaptativo representa, portanto, uma expressão matemática do conceito de "memória", estrutura cognitiva fundamental para o aprendizado e para a reação frente a estímulos externos. Esta conclusão justifica, de certa forma, porque Kohonen denomina a unidade fundamental dos mapas auto-organizativos

(análoga a um filtro adaptativo transversal, de acordo com a analogia A6 da seção 4.1) como "unidade adaptativa básica de memória" [5].

5.5 - CONCLUSÃO

Analisou-se neste capítulo a filtragem adaptativa a partir dos princípios básicos que fundamentam a definição de redes neurais.

Em termos da neurofisiologia, concluiu-se que a filtragem estatística corresponde a uma das operações realizadas pelo sistema nervoso, associada à sua capacidade de interagir com o mundo real. Além disso, comentou-se como os filtros adaptativos correspondem a modelos simplificados da célula nervosa (unidade básica constituinte do sistema nervoso) e demonstrou-se como o algoritmo do gradiente estocástico aplicado à predição linear representa uma expressão matemática da lei de Hebb (elemento fundamental para a análise da adaptação sináptica biológica). Portanto, pode-se estabelecer analogias entre a filtragem adaptativa e o sistema nervoso vertebrado, tanto a nível microscópico quanto a nível sistêmico.

Filtros adaptativos podem ser considerados redes neurais extremamente rudimentares. Realizam processamento de informação distribuído (e, para alguns casos, também paralelo) e seu treinamento através do algoritmo do gradiente estocástico segue o princípio da mínima perturbação. Consequentemente, apresentam formas rudimentares de propriedades coletivas emergentes (como compressão de informação e recuperação espontânea), que as capacitam a executar tarefas cognitivas simples (por exemplo, equalização), menos complexas que aquelas típicas de redes neurais. Todas estas características justificam, de certa forma, porque filtros adaptativos apresentam baixo desempenho para aplicações que envolvem ruído não-gaussiano ou não-aditivo; e

possibilitam evidenciar suas características intermediáriais entre redes neurais e computadores.

A análise de sinais através da desconvolução cega (que utiliza funções não-lineares ou estimativas da função de densidade de probabilidade, baseadas em momentos de ordem superior) e da predição linear (via modelamento paramétrico) pode ser associada, respectivamente, a formas de auto-organização a nível sistêmico (mapa de Kohonen) e a nível microscópico (unidade básica adaptativa de memória).

Em termos da psicologia cognitiva, ciência pertencente à linha de pesquisa básica do funcionamento do cérebro humano, a filtragem adaptativa pode ser relacionada ao conceito de "conhecimento" psicológico propriamente dito. Além disso, a formulação do filtro adaptativo transversal, treinado pelo algoritmo do gradiente estocástico, pode ser associada ao elemento psicológico "memória", estrutura cognitiva fundamental para análise do mecanismo das reações do indivíduo frente a estímulos externos.

Coincidentemente ou não, redes neurais referenciam espontaneamente conceitos intrínsecos à filtragem adaptativa. Portanto, os princípios básicos de redes neurais parecem estar relacionados, de certa forma, ao processamento de informação pela filtragem adaptativa. Deve-se notar, entretanto, que tais princípios (sinapse de Hebb, processamento paralelo distribuído, auto-organização, conhecimento e memória) foram desenvolvidos de forma desvinculada da teoria de filtragem adaptativa, pelo menos aparentemente.

Demonstrou-se neste capítulo como os princípios básicos que embasam o processamento de informação neural permitem relacionar redes neurais a filtros adaptativos, ou seja, como referenciar estruturas da filtragem adaptativa a partir de redes neurais. No próximo capítulo, a interrelação espontânea e intrínseca entre os dois conceitos será evidenciada, o que sugere uma cooperação útil para os dois campos.

CAPÍTULO 6

PROCESSAMENTO NEURAL-ADAPTATIVO DE SINAIS

Propõe-se neste capítulo um tratamento unificado das técnicas de redes neurais e da filtragem adaptativa, denominado "Processamento Neural-Adaptativo de Sinais" . Incialmente, com base nas discussões apresentadas nos capítulos 4 e 5, evidencia-se a inter-relação intrínseca e espontânea existente entre os dois conceitos. Em seguida, discutem-se as principais vantagens e fraquezas de cada técnica. Constatada a complementariedade destas, define-se o processamento neural-adaptativo de sinais com base na inter-relação espontânea, exemplificando-se a utilidade prática deste novo enfoque através da apresentação de alguns resultados já publicados na literatura [22-25]. Finalmente, uma aplicação do processamento neural-adaptativo de sinais é proposta e alguns resultados já alcançados são discutidos.

6.1 - A INTER-RELAÇÃO INTRÍNSECA E ESPONTÂNEA EXISTENTE ENTRE AS REDES NEURAIS E A FILTRAGEM ADAPTATIVA.

No capítulo 4, partindo-se de elementos associados ao formalismo matemático da filtragem adaptativa, foi possível evidenciar algumas analogias extremamente fortes com as redes neurais, que estabelecem o seguinte elo de ligação:

Filtragem Adaptativa (FA) ——— Formalismo matemático da FA ———→ Redes Neurais (RNs)

No capítulo 5, partindo-se dos princípios básicos do processamento de informação neural (sistema nervoso vertebrado [19], lei de Hebb [20], processamento paralelo distribuído [4],

auto-organização [5], conhecimento e memória [21]), foi também possível evidenciar outras analogias igualmente significativas, que aproximam as redes neurais aos filtros adaptativos, estabelecendo outro elo de ligação no esquema acima apresentado:

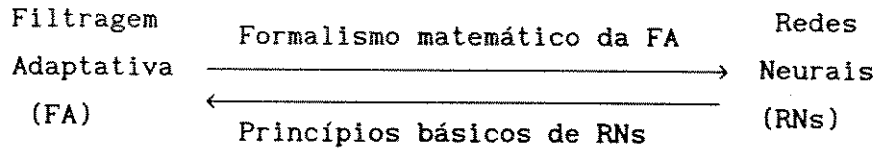


Figura 6.1 : A inter-relação intrínseca e espontânea existente entre as redes neurais e a filtragem adaptativa.

Desta forma, é possível explicitar uma inter-relação entre os dois campos, que é caracterizada pelo seu caráter intrínseco (pois o elo de ligação foi estabelecido com base nos conceitos intrínsecos de cada campo) e espontâneo (visto que, partindo-se de qualquer um dos lados, pode-se atingir o outro).

Tal interrelação está dissimulada, de certa forma, pela aparente distância existente entre o formalismo matemático da filtragem adaptativa e os princípios básicos de redes neurais. Entretanto, conforme discutido no capítulo 1, ambos foram desenvolvidos em paralelo (porém de forma separada) a partir da década de 40 [2,12,13].

Deve-se destacar que a inter-relação entre as redes neurais e a filtragem adaptativa está também expressa, de forma explícita, nas próprias denominações de alguns modelos de neurônios (embora este não representasse o objetivo principal de seus autores). Por exemplo, o neurônio Adaline, abreviatura inglesa que significa "Elemento Linear Adaptativo", é também referenciado como "Elemento Adaptativo Neural" pelos próprios autores [3]. O neurônio de Kohonen constitui outro exemplo, pois é também denominado de Unidade Adaptativa Básica de Memória [5], onde memória representa um princípio básico associado às redes neurais.

6.2 - COMPLEMENTARIEDADE ENTRE AS TÉCNICAS DE REDES NEURAIS E DA FILTRAGEM ADAPTATIVA.

O estudo realizado nos capítulos 2 e 3 evidencia a grande potencialidade de redes neurais na realização de tarefas extremamente complexas (por exemplo, nos domínios da robótica [30] e do reconhecimento de padrões [27]), sendo que sua análise teórica concentra-se basicamente na imposição de limites à capacidade de classificação da estrutura, embora alguns pesquisadores [5,9] dediquem-se também ao desenvolvimento de um formalismo matemático adequado para a análise de redes neurais. Como consequência desta linha de trabalho e da própria juventude da pesquisa nesta área, identificam-se claramente três problemas principais associados à operação de redes neurais:

1) Dificuldade de compreensão matemática dos princípios básicos que fundamentam o processamento de informação neural (auto-organização e processamento paralelo distribuído).

Deve-se notar que tais princípios, bem como suas interrelações com a função de ativação, constituem elementos determinantes do comportamento transitório e do desempenho em regime permanente de redes neurais. A análise matemática da influência dos princípios básicos nestes sistemas abre amplas perspectivas de estudo, e se faz necessária para uma melhor compreensão formal do comportamento de redes neurais e do papel desempenhado pela função de ativação [9].

2) Projeto empírico da arquitetura neural;

3) Aprendizado complexo e lento.

Tanto para algoritmos supervisionados ou auto-organizativos a formação, seleção e sistemática de apresentação de estímulos de treinamento são métodos empíricos, que influenciam diretamente o desempenho do algoritmo, em geral dependente da inicialização dos pesos sinápticos do sistema.

Deve-se notar que o segundo e terceiro problemas básicos são consequências diretas do primeiro. Desta forma, as aplicações industriais de redes neurais ainda estão, atualmente, em fase experimental [11].

A filtragem adaptativa está fundamentada pelas teorias de controle automático, de comunicações e de processos estocásticos, desenvolvidas a a partir dos trabalhos pioneiros de Shannon e de Wiener [11]. As técnicas de processamento adaptativo já são aplicadas industrialmente em diversos campos (por exemplo, telecomunicações, geofísica, engenharia biomédica [10,17]), o que pode ser justificado de certa forma pelo seu sólido embasamento matemático. Entretanto, a filtragem adaptativa possui também algumas limitações. Por exemplo, apresenta baixo desempenho na equalização de canais de comunicação de fase não-mínima ou de relação sinal-ruído reduzida [38], e supõe em geral ruído branco gaussiano e aditivo [25] (principalmente no caso de estruturas lineares).

Conclui-se, portanto, que o processamento adaptativo e as redes neurais correspondem a técnicas com vantagens e fraquezas complementares. Enquanto os filtros adaptativos processam eficientemente sinais não-estacionários a partir de teorias já consolidadas (capacidade associada ao aprendizado contínuo); as redes neurais são capazes de manipular dados de entrada mais genéricos e complexos, perturbados por ruído de caracterização matemática desconhecida (capacidade oriunda das propriedades coletivas associadas ao processamento de informação neural), embora sua análise matemática possa ser considerada ainda em estágio inicial [9], o que acarreta treinamento lento e projeto empírico.

6.3 - PROCESSAMENTO NEURAL-ADAPTATIVO DE SINAIS: UMA POSSÍVEL COOPERAÇÃO.

Visto que as redes neurais e a filtragem adaptativa correspondem a técnicas com características complementares (seção 6.2), e considerando-se a inter-relação intrínseca e espontânea de conceitos (seção 6.1), é natural buscar um tratamento unificado dos dois campos, de forma que ambos possam ser desenvolvidos simultaneamente.

Tal tratamento será denominado, a partir deste ponto, como "Processamento Neural-Adaptativo de Sinais".

Os objetivos fundamentais a serem alcançados consistem em melhor utilizar as propriedades coletivas emergentes de redes neurais e em generalizar a aplicação da filtragem adaptativa para situações mais complexas (por exemplo, para o caso de ruído não-gaussiano ou não-aditivo). Estas metas finais serão alcançadas gradualmente, de acordo com as seguintes etapas:

I) Análise matemática e formalização dos princípios básicos que fundamentam o processamento de informação neural (auto-organização e processamento paralelo distribuído).

II) Como consequência deste embasamento formal, será possível:

a) Analisar a influência dos princípios básicos (isoladamente ou interrelacionados à não-linearidade da função de ativação) na fase transitória de treinamento e no desempenho de regime de redes neurais;

b) Propor algoritmos mais eficientes para o treinamento de filtros adaptativos e de redes neurais;

c) Propor estruturas mais eficientes para determinadas aplicações do processamento de sinais (por exemplo, classificação, equalização, identificação, etc.).

Isto será realizado a partir das analogias associadas à inter-relação intrínseca e espontânea entre os dois campos. Para tanto, redes neurais e estruturas da filtragem adaptativa matematicamente equivalentes, aplicadas ao processamento de sinais, serão simuladas e posteriormente analisadas em conjunto, com base no formalismo matemático da filtragem adaptativa. Desta forma, conclusões alcançadas em um campo poderão ser transferidas para o outro. Nesta tese, as etapas I e IIa-b serão realizadas, conforme a proposta e os resultados a serem analisados no presente capítulo (seção 6.5).

A idéia principal do processamento neural-adaptativo de sinais pode ser resumida pela seguinte proposição.

PRINCÍPIO BÁSICO DO PROCESSAMENTO NEURAL-ADAPTATIVO DE SINAIS

"Uma rede neural corresponde a um conjunto de vários filtros adaptativos (possivelmente não-lineares e treinados de forma supervisionada ou não), interligados conforme determinada arquitetura e que se influenciam mutuamente em resposta a um estímulo externo. (Portanto, o comportamento da rede é resultado deste complexo conjunto de interações). Por sua vez, filtros adaptativos correspondem a neurônios (redes neurais extremamente rudimentares), que apresentam formas simples de propriedades coletivas" .

Finalmente, deve-se lembrar que os resultados a serem alcançados por este tratamento unificado representam contribuições não somente à vertente de pesquisa aplicada de redes neurais e da filtragem adaptativa, como também à vertente básica (compreensão de funcionamento do cérebro humano).

Exemplifica-se a seguir a utilidade prática da cooperação redes neurais - filtragem adaptativa através de uma breve análise de recentes resultados da literatura [22-25].

6.4 - PROCESSAMENTO NEURAL-ADAPTATIVO DE SINAIS: UMA COOPERAÇÃO ÚTIL.

A inter-relação espontânea de conceitos, passo inicial para o estabelecimento do processamento neural-adaptativo de sinais, pode ser considerada uma "descoberta" extremamente jovem. Kohonen [5], em 1983, já comparava o Perceptron a um filtro adaptativo não-linear. Posteriormente, em 1991, Haykin (através de uma nota de rodapé em [10]) e Sylvie Marcos et alli [22-23] evidenciaram outras analogias, apresentadas no capítulo 4 desta tese.

Os primeiros artigos a proporem explicitamente uma cooperação entre as redes neurais e os filtros adaptativos, dos quais se tem conhecimento, correspondem a [11,22,23] (1991) e a [24] (1992). Algumas destas contribuições são comentadas logo a seguir.

A referência [22] apresenta uma análise unificada dos algoritmos de treinamento adaptativos supervisionados (baseados nos critérios de otimização dos mínimos quadrados e dos mínimos quadrados médios [10,17,18]), que podem ser utilizados para o treinamento de estruturas da filtragem adaptativa (filtros FIR e IIR), bem como de redes neurais (Perceptron multi-camadas estático ou dinâmico) e de neurônios. Isto permite demonstrar a equivalência matemática entre determinadas técnicas de aprendizado típicas da filtragem adaptativa e o treinamento por retropropagação de diversas estruturas neurais, que podem ser incorporados à inter-relação intrínseca entre os dois campos (seção 6.1, fig. 6.2) como novos elos de ligação. Uma destas equivalências será aprofundada no presente capítulo (seção 6.5).

Já em [23], apresenta-se uma extensão dos resultados do artigo anterior. Os autores definem "famílias" de algoritmos de aprendizado, de forma a permitir a utilização de diversas regras de treinamento neurais, desconhecidas no campo do processamento adaptativo, para o aprendizado de filtros adaptativos não-lineares.

Neuvo et alli [24,25] difinem a estrutura de uma rede neural Perceptron multi-camadas, denominada de "filtro neural", a partir dos conceitos de filtro de pilha e de filtro mediano [56], apresentados nas analogias A3 e A4 da seção 4.2.

Desta forma, o processamento paralelo distribuído da estrutura pode ser analisado em termos da álgebra booleana e da decomposição de limiar (princípios básicos do filtro de pilha). Simulações computacionais demonstram que o filtro neural, aplicado à equalização, apresenta erro quadrático de regime permanente igual (no caso de ruído gaussiano branco e aditivo) ou inferior (no caso de ruído não-gaussiano, por exemplo, ruído impulsivo) àquele de filtros adaptativos transversais. Além disso, a definição matemática do filtro neural unifica várias estruturas do processamento adaptativo (por exemplo, os filtros não-lineares booleanos - filtros de pilha generalizados - e os transversais) ao Perceptron multi-camadas, visto que todos estes sistemas correspondem a casos especiais do filtro neural.

Os objetivos finais do processamento neural-adaptativo de sinais são bem caracterizados pelos resultados apresentados em [25]. Através da definição de uma rede neural com base no formalismo da filtragem adaptativa (visto que os nós constituintes do filtro neural correspondem a filtros de pilha - vide analogia A3, seção 4.2 -), é possível não somente equalizar sinais de forma mais eficiente que filtros adaptativos isolados (por exemplo, que o filtro transversal - para o caso de ruído não-gaussiano -, e que o filtro mediano - para o caso de ruído gaussiano branco e aditivo-), como também compreender a o processamento paralelo distribuído da rede neural empregada.

Além destas contribuições, deve-se destacar também, dentre outras, [37,38,54]. Embora o propósito explicíto de seus autores não fosse exatamente estabelecer um tratamento unificado, os resultados alcançados representam novos elos de ligação entre as redes neurais e a filtragem adaptativa.

A seguir, propõe-se uma aplicação do processamento neural-adaptativo de sinais e discutem-se alguns resultados já obtidos.

6.5 - APLICAÇÃO : ANÁLISE MATEMÁTICA DO PROCESSAMENTO PARALELO DISTRIBUÍDO DO PERCEPTRON MULTI-CAMADAS E APRIMORAMENTO DO ALGORITMO DO GRADIENTE ESTOCÁSTICO PARA A FILTRAGEM ADAPTATIVA EM CASCATA.

Nesta seção, será apresentada uma aplicação do processamento neural-adaptativo de sinais, baseada numa analogia que envolve a cascata de filtros adaptativos transversais e uma rede neural linear Perceptron multi-camadas parcialmente interconectada, proposta em [22]. As características desta rede possibilitam a análise do processamento paralelo distribuído de forma exata, independentemente da não-linearidade da função de ativação.

Inicialmente, discute-se a filtragem adaptativa em cascata aplicada à predição linear. Analisa-se então, sob a ótica do processamento neural-adaptativo de sinais, uma rede neural Perceptron multi-camadas linear parcialmente interconectada e a cascata de filtros adaptativos tranversais, com base na analogia estabelecida em [22]. Isto permite a proposição de uma expressão matemática para o processamento paralelo distribuído do Perceptron multi-camadas linear, bem como de uma versão modificada para o algoritmo do gradiente estocástico na forma cascata. A seguir, define-se um critério para avaliar a dependência de um algoritmo relativamente às condições iniciais impostas aos vetores de coeficientes do sistema, critério este utilizado para a comparação do desempenho de algoritmos na parte experimental. Finalmente, verificam-se as hipóteses formuladas através de simulação computacional, que evidenciam como a nova versão do algoritmo do gradiente estocástico na forma cascata é mais eficiente que a original, em termos da velocidade de convergência e da independência relativamente às condições iniciais, para o caso de um sinal de entrada auto-regressivo de ordem 4.

FILTRAGEM ADAPTATIVA: FORMA DIRETA E EM CASCATA.

A fig. 6.2(a) apresenta um filtro de erro de predição linear de ordem N, na forma direta, descrito pelo seguinte conjunto de equações:

$$f_D(k) = \sum_{j=0}^{N} a_j(k).x(k-j) \qquad (6.1a)$$

$$A_N(z) = \sum_{j=0}^{N} a_j.z^{-j} \qquad (6.1b)$$

$$a_0(k) \triangleq 1 \;\forall\; k \qquad (6.1c)$$

Onde $f_D(k)$ representa o erro de predição linear progressivo de ordem N e $A_N(z)$ corresponde à função de transferência do filtro em regime permanente. A eq. (6.1a) pode ser reescrita da seguinte forma:

$$f_D(k) = \mathbf{A}_N(k)^T.\ \mathbf{X}(k) \qquad (6.2a)$$

$$\mathbf{A}_N(k)^T = [1\ -a_1(k)\ -a_2(k)\ \ldots\ -a_N(k)] \qquad (6.2b)$$

$$\mathbf{X}(k)^T = [x(k)\ x(k-1)\ x(k-2)\ \ldots\ x(k-N)] \qquad (6.2c)$$

O treinamento do filtro adaptativo da fig. 6.2(a) objetiva minimizar a seguinte função de custo:

$$J_D = (1/2)\ E[f_D(k)^2] \qquad (6.3)$$

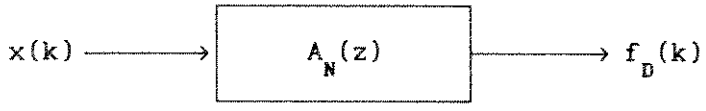
Demonstra-se que, se o sinal de entrada x(k) for estacionário no sentido amplo e de média nula, os coeficientes ótimos teóricos $\mathbf{A}_{OP}(N)$ que minimizam o critério da eq. (6.3) são estabelecidos pelas equações de Yule-Walker [10]:

$$\mathbf{A}_{OP}(N) = [a_{1op}\ a_{2op}\ \ldots\ a_{Nop}]^T = \mathbb{R}_x^{-1}.\mathbf{R}_{xx} \qquad (6.4a)$$

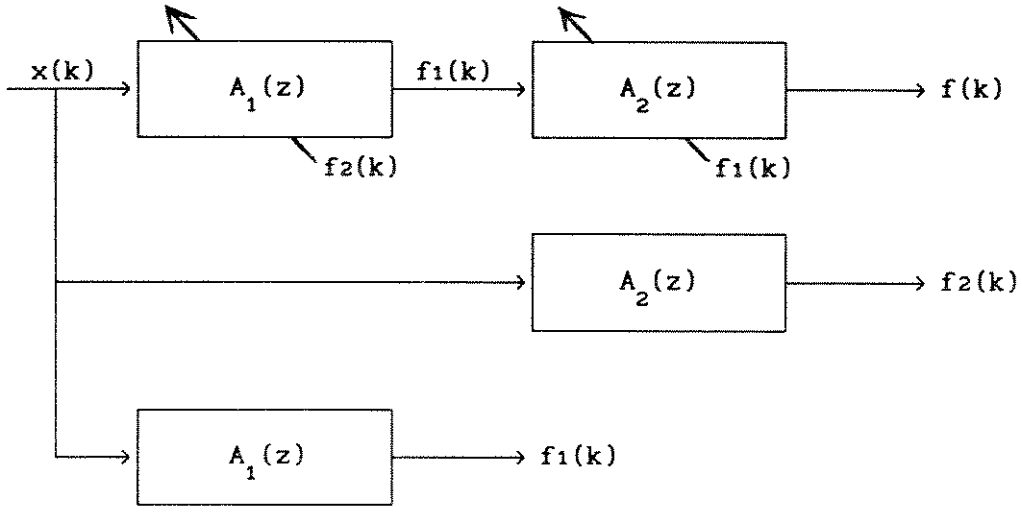
$$\mathbb{R}_x = E[\mathbf{X}(k).\mathbf{X}(k)^T] \qquad (6.4b)$$

$$\mathbf{R}_{xx} = E[\mathbf{X}(k).x(k+1)] \qquad (6.4c)$$

Onde o sinal de referência externo d(k) corresponde ao valor x(k+1).



(a)



(b)

Figura 6.2: Filtragem Adaptativa

(a) Forma Direta;

(b) Forma Cascata.

Alternativamente, pode-se estabelecer uma aproximação para a função de custo da eq. (6.3), a partir da qual deriva-se o algoritmo do gradiente estocástico (eq. (6.7)), também abreviado de LMS:

$$J_D(k) = (1/2)\, f_D(k)^2 \tag{6.5}$$

$$\nabla J_D(k) = \partial J_D(k)/\partial \mathbf{A}_N(k) = f_D(k).\mathbf{X}(k) \tag{6.6}$$

$$\mathbf{A}_N(k+1) = \mathbf{A}_N(k) + (\mu/k^2).f_D(k).\mathbf{X}(k) \tag{6.7}$$

Onde a eq.(6.5) representa uma aproximação estocástica da eq.(6.3), ∇ é o operador gradiente e μ o passo de adaptação (decrescente com o tempo, neste caso).

Se o sinal de entrada x(k) for estacionário no sentido amplo e possuir distribuição probabilística gaussiana, demonstra-se que o algoritmo do gradiente estocástico (eq. (6.7)) converge para os valores ótimos $\mathbf{A}_{OP}(N)$ (eq. (6.4a)) se o passo μ obedecer a seguinte restrição [22]:

$$\mu \leq \mu_{MAX} = 2/(N.\sigma_x^2) \tag{6.8}$$

Onde σ_x^2 representa a potência média do sinal de entrada.

A fig. 6.2(b) representa a realização na forma cascata da mesma estrutura da fig. 6.2(a). A estrutura $A_N(z)$ é subdividida em dois filtros $A_1(z)$ e $A_2(z)$, ambos de ordem N/2. O sistema é descrito pelas seguintes equações:

$$f(k) = \sum_{i=0}^{N/2} \sum_{j=0}^{N/2} a_i^2(k).a_j^1(k).x(k-i-j) \tag{6.9a}$$

$$a_0^1(k) \triangleq 1 \;\forall\; k,\; 1 \;;\; 1 = 1,\; 2 \tag{6.9b}$$

$$\mathbf{A}_1(k)^T = [a_1^1(k)\; a_2^1(k)\; \ldots\; a_{N/2}(k)] \;;\; 1 = 1,\; 2 \tag{6.9c}$$

$$A_1(z) = \sum_{j=0}^{N/2} a_j^1.z^{-j} \;;\; 1 = 1,\; 2 \tag{6.9d}$$

Onde f(k) representa a saída do sistema. De agora em diante, o

índice i denota o i-ésimo filtro da cascata, podendo assumir os valores $i = 1$ ou $i = 2$. Assim, a_i^1 corresponde ao i-ésimo coeficiente do i-ésimo filtro, enquanto que $\mathbf{A}_i(k)$ representa o vetor de coeficientes desta mesma estrutura, caracterizada pela função de transferência $A_i(z)$ em regime permanente. Deve-se notar que os coeficientes $a_0^1(k)$ são constantes, devido à operação de predição linear.

O treinamento do filtro em cascata objetiva minimizar a seguinte função de custo:

$$J = (1/2).E[f(k)^2] \qquad (6.10)$$

O algoritmo do gradiente estocástico na forma cascata pode ser derivado definindo-se uma aproximação para a eq. (6.10) e calculando-se o gradiente associado a cada filtro, através da diferenciação da nova função de custo relativamente ao vetor de coeficientes $\mathbf{A}_i(k)$ do filtro considerado [1,59]:

$$J(k) = (1/2).f(k)^2 \qquad (6.11a)$$

$$\nabla J_i(k) \triangleq \partial J(k)/\partial \mathbf{A}_i(k) = f(k).\mathbf{F}_m(k) \; ; \; i,m = 1, 2 \qquad (6.11b)$$

$$\mathbf{F}_m(k)^T = [f_m(k-1) \; f_m(k-2) \; \ldots \; f_m(k-N/2)] \; ; \; i,m = 1, 2 \qquad (6.11c)$$

$$f_m(k) = \mathbf{A}_m(k)^T.\mathbf{X}_p(k) \qquad (6.11d)$$

$$\mathbf{X}_p(k)^T = [x(k) \; x(k-1) \; \ldots \; x(k-(N/2))] \qquad (6.11e)$$

Onde $J(k)$ representa a aproximação estocástica da eq. (6.10). O índice i corresponde ao filtro considerado, enquanto o índice m denota o outro. Por exemplo, se considerarmos a adaptação do filtro 1, então $i = 1$ e $m = 2$ nas equações (6.11a-e).

Deve-se notar que o sinal $f_m(k)$ (vide fig. 6.2(b)), definido pela eq. (6.11d), é obtido pela filtragem da entrada $x(k)$ pela estrutura caracterizada por $\mathbf{A}_m(k)$, e corresponde ao "sinal de gradiente" utilizado para a adaptação de $\mathbf{A}_i(k)$ (vide eq. (6.12a)).

Finalmente, o algoritmo do gradiente estocástico na forma cascata é escrito utilizando-se a eq. (6.11b), através do método de "steepest descent" [18]:

$$\mathbf{A}_i(k+1) = \mathbf{A}_i(k) - \mu.f(k).\mathbf{F}_m(k) \; ; \; i,m = 1, 2 \qquad (6.12a)$$

$$\Delta\mathbf{A}_i(k) = \mathbf{A}_i(k+1) - \mathbf{A}_i(k) = -\mu.f(k).\mathbf{F}_m(k) \; ; \; i,m = 1, 2 \qquad (6.12b)$$

Verifica-se experimentalmente que o algoritmo do gradiente estocástico na forma cascata (eq. (6.12a)) apresenta baixa velocidade de convergência [58], especialmente se comparado a técnicas rápidas de mínimos quadrados [59]. Isto está associado, de certa forma, ao fato do algoritmo do gradiente estocástico na forma cascata não levar em consideração a correlação estatística existente entre os sinais de gradiente $f_1(k)$ e $f_2(k)$ [59].

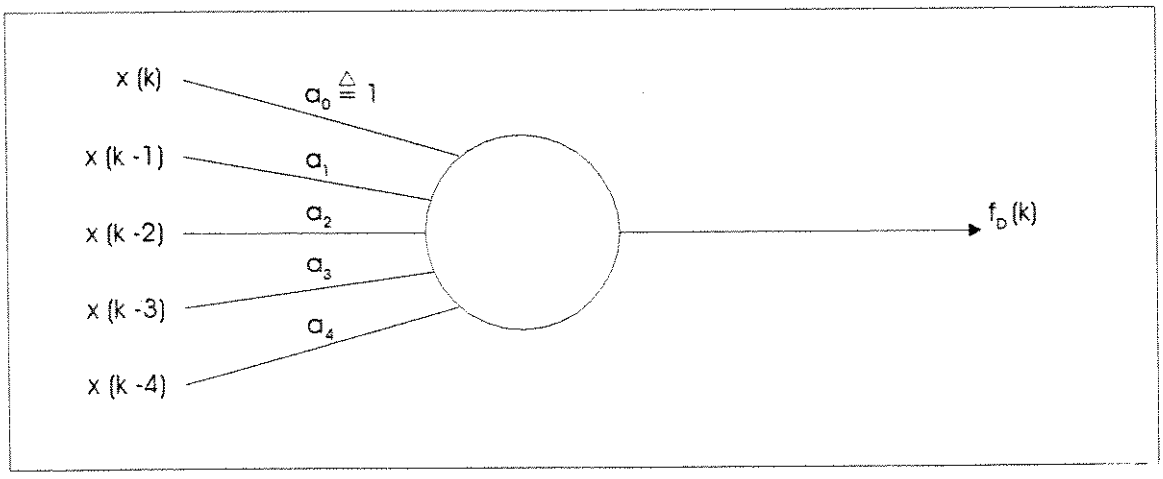
A filtragem adaptativa em cascata apresenta algumas vantagens se comparada à sua respectiva realização direta. Em termos de predição linear, por exemplo, uma estrutura em cascata, composta por filtros de ordem 2, permite cálculo rápido e eficiente dos zeros do modelo auto-regressivo do sinal de entrada, através da resolução de vários sistemas de equações de segunda ordem. (Deve-se notar que, para uma estrutura direta de ordem 8, por exemplo, o cálculo destes zeros envolveria a resolução de um sistema de ordem 8). Além disso, tal estrutura possibilita impor restrições matemáticas aos zeros do sistema de forma mais simples que a realização direta, o que representa um procedimento muitas vezes necessário nas aplicações da filtragem espacial [10].

Apresenta-se a seguir a analogia proposta em [22].

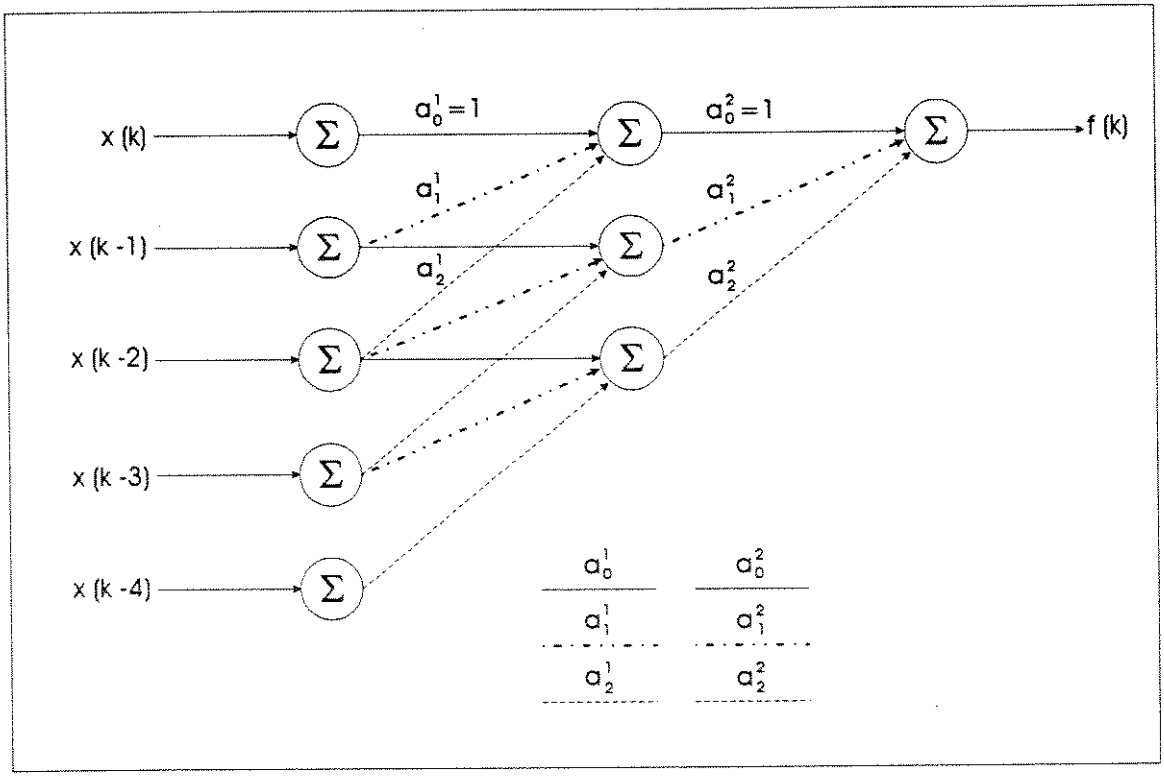
ANALOGIA ENTRE A CASCATA DE FILTROS ADAPTATIVOS TRANSVERSAIS E O PERCEPTRON MULTI-CAMADAS LINEAR.

A fig. 6.3(a) apresenta um neurônio Perceptron linear. De acordo com a analogia (A1) da seção 4.2, as equações de filtragem (eqs. (4.4b)) desta estrutura são análogas àquelas de um filtro adaptativo transversal, e em particular, àquelas do filtro de erro de predição linear (eq. (6.1a)), mostrado na fig. 6.2(a). Além disso, demonstra-se a igualdade das equações do algoritmo do gradiente estocástico, aplicado ao treinamento deste Perceptron linear e do filtro de erro de predição [22].

Portanto, as estruturas mostradas pelas figs. 6.2(a) e 6.3(a) são matematicamente equivalentes, no sentido que suas equações de filtragem e de adaptação (considerando-se aqui o algoritmo do gradiente estocástico) são iguais.



(a)



(b)

Figura 6.3 : Analogia entre a cascata de filtros adaptativos transversais e o Perceptron multi-camadas linear.
(a) Neurônio Perceptron linear (N = 4);
(b) Perceptron multi-camadas linear parcialmente interconectado [22].

Seja agora a rede neural Perceptron multi-camadas com função de ativação linear. A conclusão do parágrafo anterior, aliada ao princípio fundamental do processamento neural adaptativo de sinais (seção 6.3), sugere que esta estrutura corresponda a uma associação de diversos filtros adaptativos. Ou seja, espera-se que um Perceptron multi-camadas linear seja matematicamente equivalente a uma cascata de filtros transversais. De fato, ambas estruturas apresentam características comuns, por exemplo, o processamento distribuído (vide seção 5.2).

Esta hipótese pode ser validada por resultados apresentados em [22]. Seja a rede neural Perceptron multi-camadas apresentada na fig. 6.3(b), que possui as seguintes características:

- Função de ativação linear para todos os neurônios componentes;

- Os estímulos de entrada possuem a propriedade do deslocamento temporal (vide seção 4.1), definida pela seguinte equação:

$$\mathbf{X}(k)^T = [x(k)\ x(k-1)\ \ldots\ x(k-N)] \qquad (6.13)$$

(Onde N = 4 para a fig. 6.3(b));

- A rede é parcialmente interconectada, de forma que cada neurônio da camada 1 não recebe entradas provenientes de todos os nós da camada 1-1 (ao contrário do Perceptron multi-camadas convencional, fig. (3.5)), e de forma que a quantidade de pesos sinápticos de cada camada seja decrescente à medida que se avança para a saída da rede.

Para o caso aqui considerado (fig. 6.3(b)), os vetores de pesos sinápticos dos nós situados na primeira camada são todos iguais.

Demonstra-se [22] que as equações do algoritmo de retropropagação, aplicadas ao treinamento da estrutura mostrada na fig. 6.3(b) (a qual assume, por hipótese, N = 4), são

matematicamente iguais àquelas do algoritmo do gradiente estocástico na forma cascata (eqs. 6.12a-b), utilizadas para a adaptação dos filtros em cascata mostrados pela fig. 6.2(b), onde se considera N = 4. Além disso, demonstra-se que as equações de filtragem das duas estruturas são também análogas [22]. Este resultado pode ser generalizado para uma realização cascata qualquer de um filtro adaptativo direto, subdividido em um número arbitrário de subseções [22].

Resumindo, as figs. 6.2(a)-(b) e 6.3(a)-(b) apresentam estruturas matematicamente equivalentes, em termos da adaptação através do algoritmo do gradiente estocástico (ou de retropropagação) e das equações de filtragem. Desta forma, as saídas do Perceptron linear e do Perceptron multi-camadas das figs.6.3(a)-(b) possuem, respectivamente, as mesmas denominações das saídas das estruturas apresentadas nas figs. 6.2(a)-(b). Além disso, estabelecem-se as seguintes equivalências entre os vetores de coeficientes e de pesos sinápticos:

$$\mathbf{A}_N(k) = \mathbf{W}(k); \quad N = 4 \tag{6.14a}$$

$$\mathbf{A}_i(k) = \mathbf{W}_i(k,1); \quad i = 1, 2, 3 \tag{6.14b}$$

$$\mathbf{A}_2(k) = \mathbf{W}(k,2) \tag{6.14c}$$

Onde $\mathbf{W}(k)$ é o vetor de pesos sinápticos do Perceptron; $\mathbf{W}_i(k,1)$ é o vetor de pesos do i-ésimo neurônio da primeira camada e $\mathbf{W}(k,2)$ o vetor de pesos do neurônio de saída do Perceptron multi-camadas da fig. 6.3(b). $\mathbf{A}_N(k)$ denota o vetor de coeficientes do filtro direto, de ordem N = 4, e $\mathbf{A}_i(k)$ o conjunto de coeficientes do i-ésimo filtro da cascata (de ordem N/2 = 2).

Estas analogias serão utilizadas nas próximas subseções para analisar a cascata de filtros adaptativos e o Perceptron multi-camadas linear sob o enfoque do processamento neural-adaptativo de sinais.

A SUPERFÍCIE DE ERRO QUADRÁTICO MÉDIO DO PERCEPTRON MULTI-CAMADAS, RELAÇÕES ENTRE A DENSIDADE DE INTERCONEXÃO E A CAPACIDADE DE CLASSIFICAÇÃO E A PROPOSIÇÃO DE UMA EXRESSÃO MATEMÁTICA PARA O PROCESSAMENTO PARALELO DISTRIBUÍDO.

A função de custo J (eq. (6.10)) da cascata de filtros adaptativos transversais está associada a uma superfície de erro quadrático médio não-convexa. Entretanto, demonstra-se que seus pontos de mínimo são globais [62], ou seja, todos estão associados ao mínimo valor do erro quadrático médio. Desta forma, para a estrutura da fig. 6.2(b), os vetores de coeficientes ótimos $A_{OP,1}$ e $A_{OP,2}$ correspondem a um mesmo conjunto de valores, exceto por uma permutação arbitrária.

Com base na analogia anteriormente estabelecida, conclui-se que a superfície de erro quadrático médio do Perceptron multi-camadas da fig. 6.3(b) é também não-convexa, e que seus pontos de mínimo também são todos globais.

Deve-se notar que, em regime permanente, tanto o filtro adaptativo direto quanto sua realização em cascata são equivalentes, pois representam um mesmo sinal de entrada por modelos paramétricos iguais. Portanto, o Perceptron linear da fig. 6.3(a) e a rede neural multi-camadas da fig. 6.3(b) são também equivalentes em regime. Consequentemente, o Perceptron multi-camadas linear apresentado é capaz de classificar, no máximo, padrões linearmente separáveis, de forma análoga ao neurônio Perceptron. Tal capacidade é claramente inferior àquela de uma mesma estrutura multi-camadas totalmente interconectada e não-linear, composta por duas camadas, capaz de classificar qualquer conjunto de padrões não-linearmente separáveis [43], de acordo com os resultados teóricos comentados na seção 3.4.

Conclui-se, a partir desta comparação, que a densidade de interconexões, conjugada à não-linearidade, representam os elementos fundamentais que determinam a capacidade de classificação de um Perceptron multi-camadas.

As interações mútuas entre os diversos neurônios das duas camadas da fig. 6.3(b) definem o processamento paralelo distribuído deste Perceptron multi-camadas (abreviado por MLP). De acordo com a analogia em questão, a comparação das figs. 6.2(b) e 6.3(b) (bem como as igualdades estabelecidas pelas eqs. (6.14a-c)) sugere que a primeira camada do MLP (l = 1) representa o filtro $A_1(z)$, enquanto que a segunda (l = 2) pode ser associada a $A_2(z)$. Portanto, o processamento paralelo distribuído deste Perceptron multi-camadas é equivalente à interação entre os filtros $A_1(z)$ e $A_2(z)$ da cascata, a qual será denominada matematicamente de sinal p(k), de agora em diante.

Qual o significado de "interação $A_1(z)$-$A_2(z)$" no contexto do treinamento supervisionado? De acordo com a fig. 6.2(a), a saída $f_1(k)$ do primeiro filtro corresponde à entrada do segundo. Além disso, de acordo com o algoritmo do gradiente estocástico (eq. (6.12a)), a adaptação de um dos filtros depende do estado do outro. Por exemplo, a modificação $\Delta A_1(k)$ do filtro 1 - eq. (6.12b) - é diretamente proporcional ao termo $F_2(k)$, calculado a partir do vetor de coeficientes do filtro 2 (vide eq. (6.11d)). Com base nesta discussão, conclui-se que a influência mútua p(k) deve ser uma função dos sinais de gradiente $f_1(k)$ e $f_2(k)$.

Suponhamos que $f_1(k)$, gradiente calculado a partir do estado do primeiro filtro, possua amplitude elevada. Em conformidade com a eq. (6.12b), isto acarretará uma elevada modificação $\Delta A_2(k)$ dos coeficientes do filtro 2. Ou seja, isto significa que o segundo filtro será intensamente influenciado pelo filtro 1. Por um raciocínio análogo, se $f_2(k)$ for elevado, então $\Delta A_1(k)$ também o será, e o filtro 1 será intensamente influenciado pelo segundo. Portanto, tal influência mútua será reduzida apenas no caso de $f_1(k)$ e de $f_2(k)$ apresentarem amplitudes pequenas.

Neste contexto, conclui-se que a interação p(k) entre os filtros 1 e 2 será elevada sempre que $f_1(k)$ ou $f_2(k)$ possuírem amplitudes elevadas, e reduzida apenas se $f_1(k)$ e $f_2(k)$ forem

pequenos. Esta proposição pode ser aproximada pela seguinte equação:

$$p(k) = f_1(k) + f_2(k) \qquad (6.15)$$

A expressão anterior pode ser considerada uma possível definição matemática para o processamento paralelo distribuído do Perceptron multi-camadas da fig. 6.3(b).

Nesta subseção, através da inter-relação intrínseca existente entre as redes neurais e a filtragem adaptativa (vide fig. 6.1), foi possível analisar o Perceptron multi-camadas da fig. 6.3(b) com base na filtragem adaptativa. Isto permitiu a análise da sua superfície de erro quadrático médio, de sua capacidade de classificação e de seu processamento paralelo distribuído. A seguir, será realizado o procedimento inverso para a cascata de filtros adaptativos.

PROPOSIÇÃO DE UMA VERSÃO MODIFICADA DO ALGORITMO DO GRADIENTE ESTOCÁSTICO NA FORMA CASCATA (ou ALGORITMO LMS NEURAL).

A principal diferença entre o Perceptron multi-camadas da fig. 6.3(b) e o neurônio Perceptron da fig. 6.3(a) reside no processamento paralelo distribuído desta rede. Enquanto o neurônio atua isoladamente no padrão de entrada x(k), diversos nós processam o estímulo externo de forma conjunta, parcialmente interconectados.

Analogamente, a interação mútua existente entre os filtros $A_1(z)$ e $A_2(z)$ (fig. 6.2(b)) representa a diferença fundamental entre a cascata de filtros adaptativos da fig. 6.2(b) e sua respectiva realização direta da fig. 6.2(a).

Entretanto, os algoritmos de treinamento da cascata de filtros adaptativos, bem como aqueles utilizados para o

aprendizado de redes neurais multi-camadas, não consideram de forma explícita o processamento paralelo distribuído intrínseco a estas estruturas. Uma possível aproximação para incluir o efeito de p(k) durante o treinamento pode ser realizada através de uma definição alternativa para a o sinal de saída destes sistemas, expressa pela seguinte equação:

$$f_p(k) = f(k) + p(k) \qquad (6.16)$$

$f_p(k)$: sinal a ser considerado durante a adaptação do sistema analisado (Perceptron multi-camadas ou cascata de filtros adaptativos), onde se leva em conta o efeito do processamento paralelo distribuído p(k).

f(k) : saída do sistema analisado (Perceptron multi-camadas ou cascata), que não considera o processamento paralelo distribuído p(k).

Particularmente, para a cascata da fig. 6.2(b), a eq.(6.16) pode ser reescrita utilizando-se a eqs. (6.9a-d):

$$f(k) = a^1(k) * a^2(k) * x(k) \qquad (6.17)$$

$$f_p(k) = a^1(k) * a^2(k) * x(k) + p(k) \qquad (6.18)$$

Onde $a^1(k)$ e $a^2(k)$ denotam, respectivamente, a resposta ao impulso dos filtros 1 e 2 no instante k. O símbolo * representa a operação de convolução discreta.

Neste contexto, deriva-se agora uma versão modificada do algoritmo do gradiente estocástico na forma cascata. Primeiramente, define-se uma aproximação estocástica alternativa para a função de custo da eq. (6.10), com base na eq. (6.16), utilizando a aproximação da eq. (6.15) para definir o termo p(k). Esta nova função de custo está aqui definida apenas a título de dedução do algoritmo alternativo.

$$J_p(k) = (1/2).f_p(k)^2 \qquad (6.19a)$$

$$f_p(k) = f(k) + f_1(k) + f_2(k) \qquad (6.19b)$$

Em seguida, calcula-se o gradiente da nova função de custo, diferenciando-se a eq. (6.19a) relativamente ao vetor de coeficientes $\mathbf{A}_1(k)$ da cascata, considerando-se que os sinais $f_1(k)$ e $f_2(k)$ são definidos pelas eqs. (6.11d-e).

$$\nabla J_{p1}(k) = \partial J_p(k)/\partial \mathbf{A}_1(k) = f_p(k).\mathbf{F}_m(k) + f_p(k).\mathbf{X}_p(k) \qquad (6.20)$$

Finalmente, utiliza-se a equação anterior para se expressar o algoritmo adaptativo de acordo com o método "steepest-descent" [18]:

$$\mathbf{A}_1(k+1) = \mathbf{A}_1(k) - \mu.f_p(k).\mathbf{F}_m(k) - \mu.f_p(k).\mathbf{X}_p(k) \qquad (6.21)$$

A equação (6.21) representa uma versão modificada do algoritmo do gradiente estocástico na forma cascata (ou do algoritmo de retropropagação aplicado ao treinamento do Perceptron multi-camadas da fig. 6.3(b)), quando a estrutura da fig. 6.2(b) é considerada no contexto de redes neurais. Por este motivo, tal algoritmo será denominado de agora em diante como LMS neural (ou NLMS).

A complexidade computacional C do algoritmo do gradiente estocástico convencional na forma cascata e do LMS neural correspondem, respectivamente, a:

$$C\ (\text{LMS cascata}) = 2.N + 1 \qquad (6.22a)$$

$$C\ (\text{LMS neural}) = 4.N + 1 \qquad (6.22b)$$

Onde se supõe a realização direta de ordem N e a cascata composta por dois filtros de ordem N/2. A nova versão proposta apresenta, portanto, a mesma ordem de grandeza de complexidade computacional que o algoritmo original.

A aproximação neural permite avaliar o comportamento da potência do sinal de saída da cascata (ou erro de predição linear de ordem N). Para isto, expressa-se esta grandeza a partir da eq. (6.19b):

$$E[f_p(k)^2] = E[f(k)^2] + \left\{ E[f_1(k)^2] + E[f_2(k)^2] \right\} + \\ + 2.\left\{E[f_1(k).f_2(k)] + E[f_1(k).f(k)] + E[f_2(k).f(k)]\right\} \qquad (6.23)$$

Suponha-se agora um valor suficientemente elevado para k, de forma que o sistema esteja próximo ao regime permanente (porém não ainda em regime). Nesta situação, é razoável supor que a saída "convencional" da cascata f(k) (vide fig. 6.2(b)) seja aproximadamente igual ao sinal de saída $f_D(k)$ da respectiva realização direta (fig. 6.2(a)). Utilizando-se esta aproximação, a eq. (6.23) pode ser reescrita da seguinte forma:

$$E[f_p(k)^2] \cong E[f_D(k)^2] + \left\{ E[f_1(k)^2] + E[f_2(k)^2] \right\} + \\ + 2.\left\{E[f_1(k).f_2(k)] + E[f_1(k).f_D(k)] + E[f_2(k).f_D(k)]\right\} \qquad (6.24)$$

A eq. (6.24) sugere que a potência do sinal de saída da cascata, quando considerada no contexto neural, é superior à potência da saída da filtragem direta, pelo menos para um instante k próximo ao regime permanente. Além disso, a presença do termo $E[f_1(k).f_2(k)]$ nas eqs. (6.23-6.24) sugere que o algoritmo LMS neural leva em consideração a correlação estatística entre os sinais de gradiente $f_1(k)$ e $f_2(k)$ durante o aprendizado do sistema. Portanto, em conformidade com [59], espera-se que este algoritmo convirja mais rapidamente que o gradiente estocástico convencional para a forma cascata. Além disso, já que o LMS neural assume explicitamente o processamento paralelo distribuído intrínseco da cascata, espera-se que seja mais independente das condições iniciais que o algoritmo original.

Nesta subseção, através da inter-relação intrínseca existente entre as redes neurais e a filtragem adaptativa, foi possível analisar a cascata de filtros tranversais sob um novo enfoque (aproximação do processamento neural-adaptativo de sinais), o que possibilitou a definição de um algoritmo de treinamento alternativo.

Antes da verificação experimental das propostas e das hipóteses formuladas nas duas últimas subseções, é necessário definir um critério para avaliar a dependência de algoritmos adaptativos relativamente às condições iniciais. Supõe-se sempre que o filtro de erro de predição linear (realização direta ou em cascata, que correspondem aos sistemas de interesse nesta seção) seja treinado muitas vezes com passo de adaptação μ constante, sendo que em cada vez inicializam-se os vetores de coeficientes com valores aleatórios.

Tal critério é denominado "CRITÉRIO DO DESVIO PADRÃO", pois caracteriza o quanto um algoritmo é influenciado pelas condições iniciais através do cálculo do desvio padrão da grandeza IT (ou parâmetro IT), definida como sendo a quantidade de iterações para o algoritmo considerado convergir.

O desvio padrão da grandeza IT é calculado de acordo com as seguintes equações [63,64]:

$$s(IT) = \left[(1/n) \cdot \sum_{j=1}^{n} (IT_j - \hat{E}[IT])^2 \right]^{1/2} \qquad (6.25a)$$

$$\hat{E}[IT] = (1/n) \cdot \sum_{j=1}^{n} IT_j \qquad (6.25b)$$

$s(IT)$: desvio padrão do parâmetro IT.

n : quantidade de experimentos realizados (treinamentos do sistema analisado).

IT_j : valor do parâmetro IT para o j-ésimo experimento.

$\hat{E}[IT]$: estimativa do valor médio do parâmetro IT.

Desta forma, quanto maior o desvio padrão de IT, maior a dependência do algoritmo relativamente às condições iniciais. Teoricamente, espera-se que os desvios padrões de IT associados à filtragem direta sejam menores que aqueles da realização em cascata. Isto porque a superfície de erro quadrático médio da filtragem direta é convexa, ao passo que a respectiva superfície associada à cascata possui vários mínimos globais.

A seguir, analisam-se os resultados experimentais.

RESULTADOS DE SIMULAÇÃO E ANÁLISE.

Os objetivos principais das simulações realizadas foram verificar a validade das hipóteses elaboradas após a análise conjunta das estruturas das figs. 6.2(a)-(b) e 6.3(a)-(b) sob a ótica do processamento neural-adaptativo de sinais (especialmente as aproximações matemáticas das eqs. (6.15), (6.16), (6.23) e (6.24)), bem como analisar a dinâmica do algoritmo LMS neural. Para isto, considerou-se uma aplicação simples.

Simularam-se um filtro de erro de predição linear direto de ordem 4 (correspondente à estrutura da fig. 6.2(a) para N = 4) e sua respectiva realização cascata, composta por dois filtros de ordem 2 (correspondente à fig. 6.2(b) para N/2 = 2). Estas estruturas foram aplicadas para a análise de um sinal estacionário x(k) descrito por um modelo paramétrico auto-regressivo de ordem 4. A fig. 6.4 apresenta o esquema utilizado para simulação. O sinal x(k) foi gerado pela filtragem de um ruído branco gaussiano de média nula v(k) através de uma estrutura recursiva definida pela seguinte função de transferência:

$$A(z) = 1 + h_0 . z^{-1} + h_1 . z^{-2} + h_2 . z^{-3} + h_3 . z^{-4} \qquad (6.26)$$

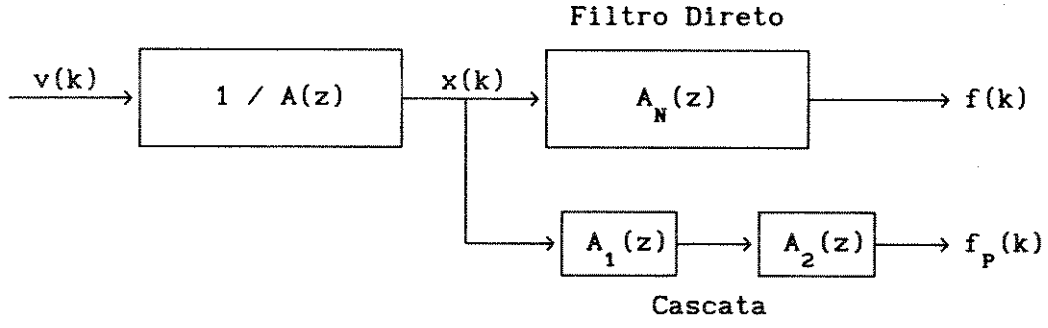


Figura 6.4: Sistema utilizado para as simulações

O sistema da fig. 6.4 foi simulado para vários conjuntos de parâmetros h_0, h_1, h_2 e h_3 do filtro recursivo 1/A(z). Fixado um determinado conjunto de parâmetros, o filtro de erro de predição linear direto e sua realização cascata foram treinados várias vezes, impondo-se diferentes condições iniciais a estes sistemas para cada simulação. Utilizaram-se os algoritmos do gradiente estocástico convencional para o filtro direto (eq. (6.7), aqui abreviado por LMS)) e, para a cascata, os algoritmos LMS neural (eq. (6.21)), abreviado NLMS) e o do gradiente estocástico original na forma cascata (eq. (6.12a), abreviado LMSc). Para todas as simulações, os melhores resultados foram alcançados para o seguinte valor do passo de adaptação:

$$\mu = 0.007\ \mu_{max} \qquad (6.27a)$$

$$\mu_{max} = 0.1605 \qquad (6.27b)$$

O valor da eq. (6.27b) obedece à restrição da eq. (6.8).

O gráfico da fig. 6.5 corresponde à média realizada considerando-se todas as simulações. A tabela 1 compara o desempenho dos algoritmos quanto à velocidade de convergência, enquanto que na tabela 2 compara-se a independência quanto à inicialização em termos do desvio padrão da grandeza IT, estabelecido anteriormente.

Antes da análise dos resultados propriamente dita, é importante relembrar que a comparação entre o filtro direto (fig. 6.2(a)) e sua respectiva realização cascata (fig. 6.2(b)), através das simulações, é equivalente à comparação do transitório do neurônio Perceptron linear (fig. 6.3(a)), treinado pela regra delta, e o do Perceptron multi-camadas parcialmente interconectado (fig. 6.3(b)), treinado pelo algoritmo de retropropagação. Consequentemente, as diferenças de comportamento transitório das estruturas simuladas representam efeitos associados ao processamento paralelo distribuído.

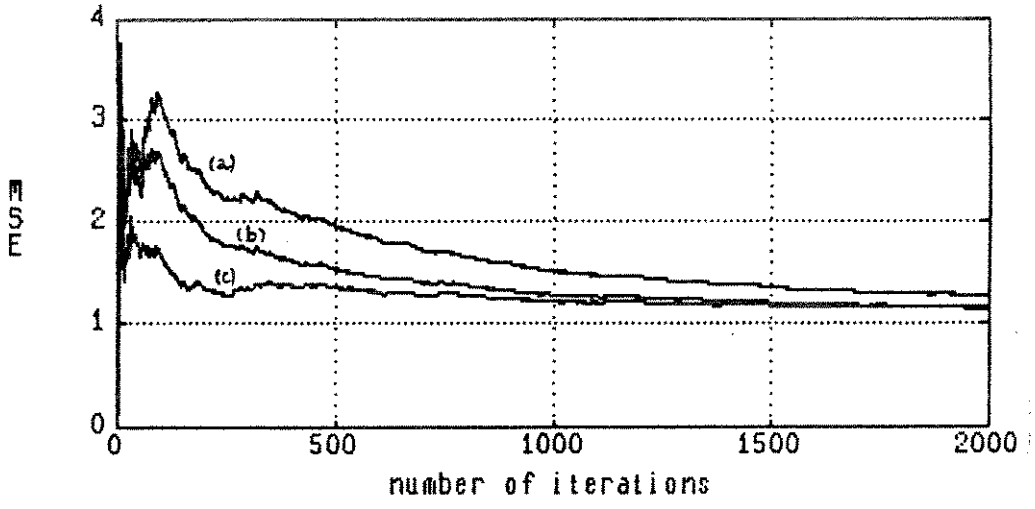


Figura 6.5 : Comparação da dinâmica dos algoritmos LMS, LMSc e NLMS. Evolução temporal do erro quadrático médio (ou MSE).
(a) Algoritmo LMSc;
(b) Algoritmo NLMS;
(c) Algoritmo LMS.

Tabela 1			
Comparação em termos da Velocidade de Convergência			
Algoritmo	LMS	LMSc	NLMS
$\hat{E}[IT]$	1208	1275	1125

Tabela 2			
Comparação em termos da Independência quanto à Inicialização - Critério do Desvio Padrão.			
Algoritmo	LMS	LMSc	NLMS
s(IT)	248.45	518.42	439.75

A fig. 6.5 evidencia as seguintes diferenças entre os algoritmos:

D1) O erro quadrático médio, para o caso da cascata treinada pelo algoritmo LMSc convencional, é superior aos demais, durante todo o transitório.

D2) O algoritmo do gradiente estocástico convencional LMS converge mais rapidamente que sua versão cascata LMSc (vide tabela 1).

D3) O algoritmo LMS neural converge mais rapidamente que sua versão original, o algoritmo LMSc (vide também a tabela 1), além de apresentar comportamento transitório mais próximo ao do gradiente estocástico para a forma direta.

A diferença (D1) não apenas confirma a expectativa teórica da eq. (6.24), como também a generaliza, de certa forma, para todo o transitório da adaptação. A diferença (D2) pode ser justificada pelo fato do algoritmo do gradiente estocástico convencional na forma cascata não levar em consideração o processamento paralelo distribuído do sistema. Finalmente, a diferença (D3) confirma as expectativas anteriores quanto à velocidade de convergência do novo algoritmo.

A tabela 2 evidencia que o desvio padrão do parâmetro IT do sistema analisado diminuiu com a utilização do novo algoritmo, aproximando-se do respectivo valor para a adaptação do filtro direto. Conclui-se, portanto, que este é mais independente das condições iniciais da cascata que a versão original LMSc, o que confirma as expectativas anteriores.

6.6 - CONCLUSÃO

Definiu-se neste capítulo o processamento neural-adaptativo

de sinais, um tratamento unificado das técnicas de redes neurais e da filtragem adaptativa.

O objetivo principal deste tratamento unificado é conjugar as potencialidades intrínsecas das propriedades coletivas emergentes de redes neurais ao sólido formalismo matemático que fundamenta a filtragem adaptativa. Desta forma, será possível uma análise matemática mais aprofundada dos princípios básicos do processamento de informação neural (auto-organização e processamento paralelo distribuído) - o que propiciará uma melhor utilização de redes neurais -, bem como generalizar a aplicação de filtros adaptativos para situações mais complexas (por exemplo, no caso de ruído não-gaussiano e não-aditivo).

Comentaram-se recentes publicações da literatura [22-25] a respeito da cooperação entre os dois campos e apresentou-se uma possível aplicação do processamento neural-adaptativo de sinais, baseada em uma analogia estabelecida entre a cascata de filtros adaptativos transversais e uma rede neural Perceptron multi-camadas linear parcialmente interconectada [22].

Com base na inter-relação intrínseca de conceitos e no formalismo matemático que fundamenta a filtragem adaptativa, foi possível concluir que a superfície de erro quadrático médio de um determinado Perceptron multi-camadas linear e parcialmente interconectado é não-convexa, mas que seus mínimos são globais. (Isto sugere que a superfície de erro quadrático médio de um Perceptron multi-camadas linear, totalmente interconectado, também seja não-convexa). Além disso, estabeleceu-se a capacidade de classificação da estrutura analisada. Reconheceu-se que o processamento paralelo distribuído (PDP) deste Perceptron parcialmente interconectado corresponde à interação entre os dois filtros adaptativos da cascata, o que possibilitou a proposição de uma expressão para o PDP. Através das diferenças entre o comportamento transitório da cascata e aquele de sua respectiva realização direta, observadas nas simulações realizadas,

analisaram-se os efeitos do processamento paralelo distribuído sobre o sistema, de forma independente da não-linearidade da função de ativação.

Com base na inter-relação espontânea de conceitos e no contexto de redes neurais, foi possível evidenciar que o processamento paralelo distribuído corresponde a um conceito intrínseco à cascata de filtros adaptativos transversais, bem como propor um algoritmo de treinamento que leva em consideração os efeitos associados ao PDP da cascata. Através de simulações computacionais da predição linear de um sinal de entrada auto-regressivo de ordem 4, constatou-se que o algoritmo do gradiente estocástico neural converge mais rapidamente e apresenta uma maior independência das condições iniciais do sistema que sua versão original, o que confirmou as expectativas anteriores formuladas na análise teórica.

Finalmente, é ainda interessante observar que o sinal p(k), representativo do processamento paralelo distribuído, pode ser interpretado como uma espécie de não-linearidade do sistema. De fato, pode-se verificar que a eq. (6.18) (relação de entrada-saída da cascata de filtros adaptativos, considerada no contexto neural e para um sinal p(k) genérico) caracteriza um sistema que não obedece ao princípio da superposição para qualquer instante k considerado, justamente devido ao fato de se incluir o sinal p(k) na eq. (6.18). Neste contexto, conclui-se então que o processamento paralelo distribuído pode ser considerado como uma espécie de não-linearidade sistêmica por si só, independentemente da função de ativação da rede neural e do fato da cascata ser constituída de filtros adaptativos lineares ou não-lineares.

A seguir, as principais conclusões desta tese são apresentadas e discutem-se suas possíveis extensões.

CAPÍTULO 7

CONCLUSÃO E PERSPECTIVAS

Redes neurais correspondem a modelos matemáticos simplificados de sistemas biológicos descritos experimentalmente pela neurofisiologia. Graças ao processamento paralelo distribuído, apresentam diversas propriedades coletivas emergentes, que as capacitam a executar tarefas cognitivas. Um exemplo clássico destas tarefas é a classificação de padrões, que exige do sistema a capacidade de manipulação, em tempo real, de dados de entrada simbólicos e estruturalmente complexos, perturbados por ruído de difícil caracterização matemática. Deve-se notar que o projeto de redes neurais possui caráter empírico, e que seu treinamento é, em geral, complexo e lento.

Os filtros adaptativos correspondem a sistemas com parâmetros variantes no tempo, solidamente fundamentados pela teoria da otimização, de processos estocáticos e de comunicações. Através do treinamento contínuo e sob algumas restrições (que correspondem em geral aos casos de ruído gaussiano e aditivo), são capazes de processar eficientemente sinais não-estacionários (séries temporais, em geral), bem como de caracterizá-los matematicamente através de modelos paramétricos. Uma aplicação clássica da da filtragem adaptativa é a equalização de sinais digitais, transmitidos por canais de comunicação ruidosos.

Embora desenvolvidos independentemente, as redes neurais e os filtros adaptativos compartilham a propriedade coletiva de compressão de informação, utilizada de forma distinta nos dois casos devido às características dos sinais envolvidos e das aplicações clássicas. Para as redes neurais, a compressão de informação significa elevada GENERALIZAÇÃO, que capacita o sistema a processar eficientemente ruídos ou degradações impostas aos estímulos externos (de estrutura extremamente complexa) em regime permanente, ou seja, sem a necessidade de uma adaptação dos pesos

sinápticos. Esta capacidade das redes neurais é adquirida com base no aprendizado prévio que utiliza um conjunto limitado de estímulos de treinamento. Para a filtragem adaptativa, a compressão de informação pode ser traduzida por elevada ADAPTATIVIDADE, ou seja, a capacidade de acompanhar as variações da estrutura estatística do sinal de entrada em tempo real (o que implica em máxima velocidade de convergência do algoritmo de treinamento), através da constante modificação dos parâmetros do sistema.

É importante mais uma vez destacar a necessidade de uma formação multidisciplinar mínima para a pesquisa eficiente em redes neurais, quer seja ela aplicada ou básica, e independentemente da aplicação ou do objetivo final. De fato, são justamente os conceitos inspiradores de redes neurais que fundamentam e validam a definição da inter-relação intrínseca de conceitos, sustentáculo do processamento neural-adaptativo de sinais. Além disso, deve-se destacar como foi possível, através de argumentos eminentemente biológicos, justificar algumas constatações práticas observadas no treinamento das arquiteturas neurais (vide seção 3.8).

É necessário reafirmar o caráter de cooperação mútua do processamento neural-adaptativo de sinais, que o diferencia da pesquisa tradicional nos campos de redes neurais e da filtragem adaptativa. É no sentido de fornecer os fundamentos teóricos de base para a cooperação entre os dois campos que a aparece a motivação maior desta tese, assim como suas principais contribuições.

Dentre estas contribuições, destacam-se as analogias apresentadas nos capítulos 4 e 5, que estabelecem a inter-relação entre os conceitos de redes neurais e filtragem adaptativa, através da qual espera-se possibilitar resultados novos e alternativos aos da pesquisa tradicional. Para se chegar a estas analogias, foi necessário um estudo aprofundado nas duas áreas e a

busca de um formalismo comum.

Uma primeira e significativa etapa na busca desta cooperação é justamente a proposição do processamento neural-adaptativo, apresentado no capítulo 6. O estudo de um caso aparentemente simples de uma cascata de dois filtros adaptativos tornou possível obter conclusões a respeito da não-convexidade da superfície de erro quadrático médio de um Perceptron multi-camadas linear, de uma possível medida a ser relacionada ao conceito de processamento paralelo distribuído e de sua utilização no sentido de melhorar a técnica de treinamento da cascata.

Em consequência, o algoritmo NLMS foi proposto para uma cascata de filtros adaptativos, obtendo-se resultados superiores ao convencional.

Algumas perspectivas imediatas decorrem das contribuições acima citadas. Por exemplo, a própria análise matemática do filtro em cascata pode ser aprofundada no sentido de obter uma melhor aproximação para o termo relativo ao processamento paralelo distribuído, que seja generalizável para estruturas quaisquer e mesmo para um Perceptron multi-camadas não-linear totalmente interconectado.

Em termos das analogias propostas, destacam-se possíveis perspectivas, relacionando a teoria da desconvolução cega, área emergente em filtragem adaptativa, e suas associações com técnicas de auto-organização neural. Também é bastante interessante a idéia de se prosseguir o estudo de algoritmos adaptativos e sua relação com o modelo matemático do neurônio biológico. Isto possibilitaria o uso da filtragem adaptativa como ferramenta auxiliar da neurofisiologia e também, como perspectiva a mais a curto prazo, a busca de algoritmos mais eficientes de treinamento de redes neurais.

Em suma, almeja-se que o formalismo bem estabelecido da teoria de filtragem adaptativa possa possibilitar um suporte importante ao tratamento matemático das redes neurais. Do mesmo modo, pode-se explorar a filtragem adaptativa em tarefas cognitivas e como mais uma área engajada na vertente de pesquisa básica para a compreensão do cérebro humano. Neste sentido, longe de pretender a resultados conclusivos, esta tese se propôs a estabelecer um primeiro passo em direção a esta cooperação.

CAPÍTULO 8

REFERÊNCIAS

[1] D. R. Hush and B. G. Horne; "Progress in Supervised Neural Networks", IEEE SP Mag.,pp 8-39,Jan. 1993.

[2] W.S. McCulloch and W. Pitts; "A Logical Calculus of the Ideas Imminent in Nervous Activity", Bulletin of Mathematical Biophysics, 5, 115-133, 1943.

[3] B. Widrow and M.E. Hoff; "Adaptive Switching Circuits", 1960 IRE WESCON Conv. Record, Part 4, 96-104, Aug. 1960.

[4] D.E. Rumelhart, J.L. McClelland and the PDP Research Group; "Parallel Distributed Processing: Explorations in the Microstructure of Cognition", vol.1, MIT Press, Massachusetts, 1989.

[5] T. Kohonen; "Self-Organization and Associative Memory", Springer-Verlag, Berlin, 1989.

[6] S. Grossberg; "Adaptive Pattern Classification and Universal Recoding, II: Feedback, Expectation, Olfaction and Illusions", Biolog. Cybernetics, vol. 23, pp. 187-202, 1976.

[7] J. A. Feldman and D. H. Ballard; "Connectionist Models and Their Properties", Cognitive Science, vol. 6, 205-254, 1982.

[8] G. E. Hinton, R. J. Sejnowski and D. H. Ackley; "Boltzmann Machines: Constraint Satisfaction Networks that Learn", Tech. Rep. CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computers Science, 1984.

[9] S.-I. Amari; "Mathematical Foundations of Neurocomputing", Proc.of the IEEE, vol. 78, no. 9, pp. 1443-1463, Sep. 1990.

[10] S. Haykin ; "Adaptive Filter Theory", Prentice-Hall Inc., New Jersey, 1991.

[11] G. Favier; "Traitement du signal, réseaux de neurones et systèmes experts: Le mariage à trois pour les futurs systèmes intelligents de traitement de l'information?", Éditorial, Traitement du Signal, vol. 8, no. 6, pp. 383-385, 1991.

[12] C. E. Shannon; "A Mathematical Theory of Communication", Bell System Tech. J., vol. 27, pp. 379-423 and pp. 623-656, 1948.

[13] N. Wiener; "Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications", The MIT Press, Cambridge, Mass., 1949.

[14] R. E. Kalman; "A new approach to linear filtering and prediction problems", Trans. ASME, J. Basic Eng., Ser. 82D, pp. 85-45, March 1960.

[15] F. Rosenblatt; "The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain", Psychological Review, 65:386-408, 1958.

[16] M. J. D. Powell; "Radial Basis Functions for Multivariate Interpolation: A Review", Technical Report DAMPT 1985/NA12, Dept. of App. Math. and Theor. Physics, Cambridge University, Cambridge, England, 1985.

[17] M. G. Bellanger; "Adaptive Digital Filters and Signal Analysis", Marcel-Dekker Inc., New York, 1987.

[18] B. Widrow and S. D. Stearns; "Adaptive Signal Processing", Prentice-Hall, New Jersey, 1985.

[19] E. R. Kandel and J. H. Schwartz; "Principles of Neural Sciences", 3rd Edition, Elsevier Science Publishing Co., 1991.

[20] D. O. Hebb; "The Organization of Behavior", John Wiley & Sons, New York, 1949.

[21] J. Piaget; "Six Études de Psychologie", Éditions Gonthier S. A., Genève, 1964.

[22] S. Marcos, O. Macchi, C. Vignat, G. Dreyfus, L. Personnaz and P. Roussel-Ragot; "A Unified Framework for Gradient Algorithms Used for Filter Adaptation and Neural Network Training", Int. Journal of Circuit Theory and Applications, vol. 20, pp. 159-200, 1992.

[23] S. Marcos, P. Roussel-Ragot, L. Personnaz, O. Nerrand, G. Dreyfus et C. Vignat; "Réseaux de neurones pour le filtrage non linéaire adaptatif", Traitement du Signal, vol.8, no.6, pp. 409-421, 1991.

[24] L.Yin, J. Astola and Y.Neuvo; "Neural Filters: A Class of Filters Unifying FIR and Median Filters", Proc. of ICASSP 92, vol.IV, pp. 53-56, San Francisco, 1992.

[25] L. Yin, J. Astola and Y. Neuvo; "A New Class of Nonlinear Filters- Neural Filters", IEEE Trans. on Signal Processing, vol. 41, no. 3, pp.1201-1222, March 1993.

[26] J. J. Hopfield; "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", Proc. Natl. Acad. Sci., vol. 79, pp. 2554-2558, Apr. 1982.

[27] T. Sejnowski and C. R. Rosenberg; "NETTalk: a Parallel Network That Learns to Read Aloud", Johns Hopkins Univ. Technical Report JHU/EECS-86/01, 1986.

[28] A. J. Robinson and F. Fallside; "Static and Dynamic Error Propagation Networks with Application to Speech Coding". In D.Z. Anderson, editor, *Neural Information Processing Systems*, pp. 632-641, New York, NY, 1988.

[29] R.J. Williams and D. Zipser; "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", Neural Computation, 1(2):270-280, 1989.

[30] G. N. Reeke,Jr., O. Sporns and G. M. Edelman; "Synthetic Neural Modeling: The "Darwin" Series of Recognition Automata", Proc. IEEE, vol.78, no.9, pp.1498-1530, Sep. 1990.

[31] B. Widrow and M. A. Lehr; "30 Years of Adaptive Neural Networks: Perceptron, Madaline and Backpropagation", Proc. of the IEEE, vol.78, no.9, pp.1415-1442, Sep. 1990.

[32] R. P. Lippman; "An Introduction to Computing with Neural Nets", IEEE ASSP Mag., vol. ASSP-38, vol.4, pp.4-22, Apr. 1987.

[33] T. Kohonen, "The Self-Organizing Map", Proc. of the IEEE, vol. 78,no. 9, pp.1464-1480, Sep. 1990.

[34] D. Essen; "Functional Organization of Primate Visual Cortex", in Cerebral Cortex, vol. 3, Plenum Press, pp.259-329, 1985.

[35] R. A. Reale and T. J. Imig; "Tonotopic Organization in Auditory Cortex of the Cat", Journal Comp. Neurol., vol. 192, pp. 265-291, 1989.

[36] P. Guillaume; "Manuel de Psychologie", Presses Universitaires de France", Paris, 1960.

[37] Z. Xiang and G. Bi; "Complex Neuron Model with its Applications to M-QAM Data Communications in the Presence of Co-Channel Interferences", Proc. of ICASSP 92, vol.II, pp.305-308, San Francisco, 1992.

[38] G. J. Gibson, S. Siu and Colin F. N. Cowan; "The Application of Nonlinear Structures to the Reconstruction of Binary Signals", IEEE Trans. Signal Processing, vol.39, no. 8, pp. 1877-1884, Aug. 1991.

[39] P. M. Grant and J. P. Sage; "A Comparison of Neural Network and Matched Filter Processing for Detecting Lines in Images", in J.S. Denker (Ed.) AIP Conference Proceedings, 151, Neural Networks for Computing, Snowbird Utah, AIP, 1986.

[40] T. Kohonen, K. Masisara and T. Saramaki; "Phonotopic Maps - Insightful Representation of Phonological Features for Speech Representation", Proceedings IEEE 7th Inter. Conf. on Pattern Recognition, Montreal, Canada, 1984.

[41] K. Akazawa e K. Kato; "Neural Network Model for Control of Muscle Force Based on the Size Principle of Motor Unit", Proc. IEEE, vol. 78, no. 9, pp. 1531-1535, Sep. 1990.

[42] R. Rosenblatt; "Principles of Neurodynamics", New York, Spartan Books, 1959.

[43] J. Makhoul, A. El-Jaroudi and R. Schwartz; "Formation of Disconnected Decision Regions with a Single Hidden Layer", in Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 455-460, 1989.

[44] Y. le Cun, J.S. Denker and S.A. Solla; "Optimal Brain Damage". In D. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pp.598-605, Morgan Kaufmann, 1990.

[45] Y. Linde, A. Buzo and R. M. Gray; "An Algorithm for Vector Quantization", IEEE Trans. Communications, vol. COM-28, pp. 84-95, 1980.

[46] A. Gersho; "On the Structure of Vector Quantizers", IEEE Trans. Inform. Theory, vol. IT-25, no. 4, pp. 373-380, July 1979.

[47] E. J. Hartman, J. D. Keeler and J. M. Kowalski; "Layered Neural Networks with Gaussian Hidden Units as Universal Approximations", Neural Computation, 2 (2), pp. 210-215, 1990.

[48] J. Moody and C. J. Darken; "Fast Learning in Networks of Locally-Tuned Processing Units", Neural Computation, 1:281-293, 1989.

[49] D. H. Hubel, T. N. Wiesel; J. Comp. Neurol., 158, 307, 1974.

[50] D. I. Perrett, E. T. Rolls, W. Caan; Exp. Brain Res., 47, 329, 1982.

[51] K. Fukushima; Biol. Cyb., 36, 193, 1980.

[52] E.Harrigan, J.R. Kroh, W.A.Sandham, T.S. Durrani; "Seismic Horizon Picking using an Artificial Neural Network", Proc. ICASSP 92, vol. III ,pp.105-108, San Francisco, 1992.

[53] R. Nambiar, C. K. K. Tang and P. Mars; "Genetic and Learning Automata Algorithms for Adaptive Digital Filters", Proc. of ICASSP 92, vol. IV, pp.41-44, San Francisco, 1992.

[54] A. Feuer and R. Cristi; "On the Optimal Weight Vector of a Perceptron with Gaussian Data and Arbitrary Nonlinearity", IEEE Trans. on Sig. Proc., vol. 41, no. 6, pp. 2257-2259, June 1993.

[55] N. S. Jayant and P. Noll; "Digital Coding of Waveforms - Principles and Applications to Speech and Video", Prentice-Hall Inc., New Jersey, 1984.

[56] P. D. Wendt, E. J. Coyle and N. C. Jr. Gallagher; "Stack Filter", IEEE Trans. ASSP, vol. ASSP-34, no. 4, pp. 898-911, Aug. 1986.

[57] J. B. Destro Filho; "Filtragem Neuro-Adaptativa", Trabalho de Curso IA353, FEE-UNICAMP, Dez. 1992.

[58] L. B. Jackson and S. L. Wood; "Linear Prediction in Cascade Form", IEEE Trans. on ASSP, vol. 26, no. 6, pp. 518-528, Dec. 1978.

[59] J. M. Travassos Romano; "Localisation de Fréquences Bruitées par Filtrage Adaptatif et Implantation d'Algorithmes des Moindres Carrées Rapides", Thèse de Doctorat, Orsay 1987.

[60] P. Foulquier et G. Deledalle; "La Psychologie Contemporaine", Presses Universitaires de France, Paris, 1951.

[61] F. S. Keller and W. N. Schoenfeld; "Principles of Psychology", Appleton-Century-Crofts Inc., New York, 1950.

[62] L. C. Coradine; "Filtragem Adaptativa em Cascata: Proposta de Estrutura e Algoritmo, Análise e Aplicações", Tese de Doutorado, Faculdade de Engenharia Elétrica, UNICAMP, Setembro 1993.

[63] G. K. Bhattacharyya and R. A. Johnson; "Statistical Concepts and Methods", John Wiley & Sons Inc., 1977.

[64] J. P. Holman; "Experimental Methods", McGraw-Hill International, 1983.