5
Redes Neurais

O trabalho em redes neurais artificiais, usualmente denominadas “redes neurais” ou “RNA”, tem sido motivado desde o começo pelo reconhecimento de que o cérebro humano processa informações de uma forma inteiramente diferente do computador digital convencional. O cérebro é um computador (sistema de processamento de informação) altamente complexo, não-linear e paralelo. Ele tem a capacidade de organizar seus constituintes estruturais, conhecidos por neurônios, de forma a realizar certos processamentos, por exemplo, reconhecimento de padrões, percepção e controle moto, muito mais rapidamente que o mais rápido computador digital até hoje existente. Considere, por exemplo, a visão humana, que é uma tarefa de processamento de informação. A função do sistema visual é fornecer uma representação do ambiente a nossa volta, e mais importante que isso, fornece a informação de que necessitamos para interagir com o ambiente. O cérebro realiza rotineiramente tarefas de reconhecimento perceptivo, por exemplo, reconhecimento de um rosto familiar inserido em uma cena não familiar, em aproximadamente 200ms, ao passo que tarefas de complexidade muito menor podem levar dias para serem executadas em um computador convencional (HAYKIN, 2001).

É importante reconhecer que os neurônios artificiais que são utilizados para construir as redes neurais artificiais são muito primitivos, se comparados aos neurônios encontrados no cérebro, e as redes neurais artificiais que são capazes de se projetar atualmente são primitivas se comparadas aos circuitos locais e circuitos inter-regionais do cérebro. No entanto, com a profusão de novas teorias, tanto no estudo das redes neurais artificiais, quanto no estudo da fisiologia cerebral, espera-se que nos próximos anos este ramo da ciência seja um estudo muito mais sofisticado do que é atualmente.

Na sua forma mais geral uma RNA é uma máquina que é projetada para modelar a maneira como o cérebro realiza uma tarefa particular ou função de interesse. Para alcançarem bom desempenho, as redes neurais empregam uma interligação maciça de células computacionais simples denominadas “neurônios” ou “unidades de processamento”. Em outras palavras a rede é um processador

maciçamente paralelamente distribuído constituído de unidades de processamento simples, que tem a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos: o conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem e as forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido (HAYKIN, 2001). O "conhecimento" e o "aprendizado" são armazenados nos valores dos pesos, biases e funções contidas em cada uma das conexões e em cada neurônio.

Portanto, todos os neurônios executam sua parcela computacional, onde cada um deles recebe sequencialmente por estímulos externos e possui um limite de ativação inerente (bias), gerando uma saída em função deste limite.

Os elementos básicos de um neurônio podem ser vistos na figura 5.1, apresentando a transformação dos estímulos em uma informação, através do recebimento dos sinais de entrada, da respectiva ponderação pelas sinapses (pesos), da adição ocorrida dentro do elemento principal (bias) e da restrição na amplitude de saída de acordo com a função de ativação do mesmo.

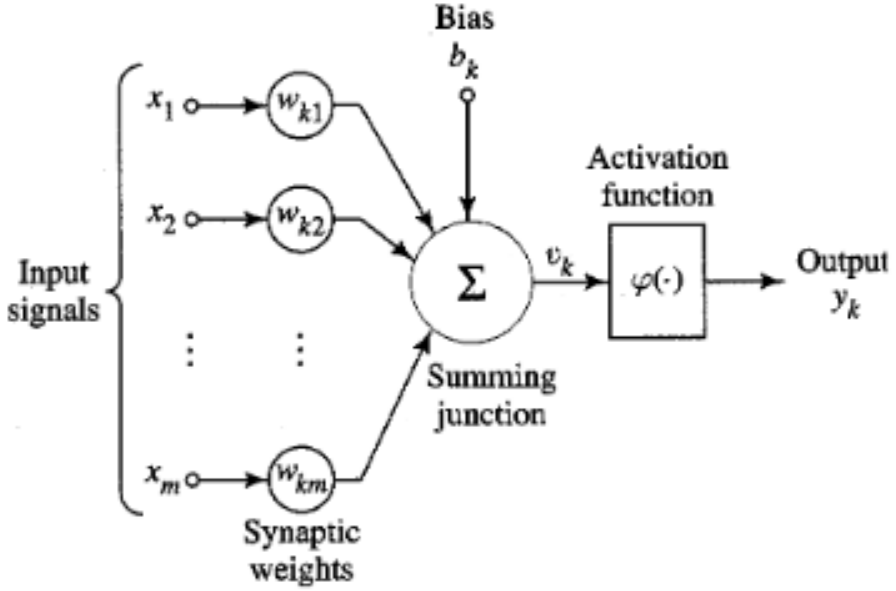

Figura 5.1.Modelo não linear de um neurônio j da camada k+1. Fonte: HAYKIN, 2001

O modelo básico de um neurônio utilizado no projeto de redes neurais artificiais consiste de:

1. Um conjunto de sinapses, cada uma delas caracterizada por um peso característico. Especificamente, um sinal x_j na entrada da sinapse j conectada ao neurônio k é multiplicado pelo peso sináptico w_{kj}. Diferentemente de uma sinapse no cérebro, o peso sináptico de um neurônio artificial pode assumir valores positivos e negativos.

2. Um combinador linear para somar os sinais de entrada, ponderados pela respectiva sinapse do neurônio.

3. Uma função de ativação para limitar a amplitude da saída do neurônio. A função de ativação limita a faixa de amplitude permitida do sinal de saída a algum valor finito. Tipicamente, a excursão da amplitude normalizada da saída de um neurônio é restrita ao intervalo unitário fechado [1 , 0] ou, alternativamente [-1 , 1].

O modelo neural da Figura 5.2 inclui uma polarização externa (*bias*), denotada por b_k . A polarização b_k tem o efeito de aumentar ou diminuir o argumento da função de ativação, caso seja positivo ou negativo, respectivamente.

Em termos matemáticos, um neurônio k pode ser descrito pelas equações

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{5.1}$$

$$y_k = \phi(u_k + b_k) \tag{5.2}$$

onde:

$x_1, x_2, ..., x_m$ são os sinais de entrada;

w_{k1}, w_{k2}, ..., w_{k2} são os pesos sinápticos do neurônio k;

u_k é a saída do combinador linear devida aos sinais de entrada;

b_k é a polarização ou bias;

ϕ (.) é a função de ativação e

y_k é o sinal de saída do neurônio

O uso da polarização ou *bias* tem o efeito de aplicar uma transformação à saída u_k do combinador linear conforme

$$v_k = u_k + b_k \tag{5.3}$$

Dependendo do valor da polarização b_k ser positivo ou negativo, a relação entre o potencial de ativação v_k do neurônio k e a saída do combinador linear u_k conforme figura 5.2 . Como o resultado da transformação, o gráfico de v_k x u_k não passa mais pela origem.

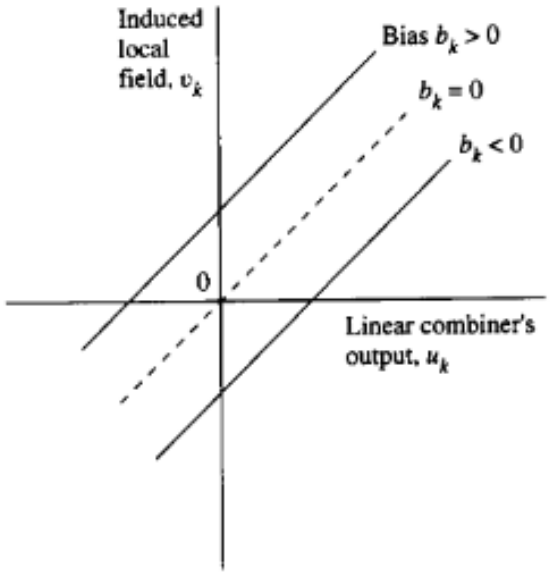

Figura 5.2. Transformação produzida pela polarização ou bias (vk = bk para uk = 0)

Fonte: HAYKIN, 2001

A RNA pode ter diversas camadas e inúmeros neurônios em cada uma destas camadas. O tamanho e a complexidade da rede refletem o nível de complexidade do processamento da informação. As entradas são sinais recebidos em neurônios específicos (da camada de entrada) e se propagam pela rede até gerar uma ou mais saídas. As entradas podem ser variáveis contínuas ou categóricas. (GANDELMAN, 2012)

Aplicada a um processo industrial, a rede deve ser capaz de fornecer respostas baseadas em um conhecimento adquirido, fornecido em uma etapa de aprendizado prévia. Em linhas gerais, esta rede é constituída por uma camada de entrada, recebendo as informações do processo, uma ou mais camadas escondidas, processando internamente as informações através de suas funções de ativação, produzindo as sinapses, e uma camada de saída, disponibilizando o resultado da análise. (GANDELMAN, 2012)

Através de uma rede simples com estrutura *feedforward* (em que os sinais se propagam sempre das camadas mais anteriores para as posteriores), cada neurônio de uma camada recebe as ativações correspondentes às saídas da camada anterior, processando os estímulos enviados pela camada de entrada ao longo das camadas escondidas até a camada de saída, produzindo a resposta (HOSKINS; HIMMELBLAU, 1988). Estas redes são normalmente chamadas de *perceptrons* de múltipla camada (Multilayer Perceptron, MLP) e têm sido aplicadas com sucesso em processos industriais online.

Algumas características relevantes das redes neurais artificiais são descritas por SIMON HAYKIN em Redes Neurais – Princípios e práticas e aqui citada:

- A Possibilidade de considerar o comportamento não-linear dos fenômenos físicos responsáveis pela geração dos dados de entrada. Um neurônio artificial pode ser linear ou não-linear. Uma RNA constituída de interconexões de neurônios não-lineares é uma rede não-linear. É importante observar que a não-linearidade de uma RNA é distribuída por toda a rede. Não-linearidade é uma propriedade altamente importante, particularmente se o mecanismo físico subjacente responsável pela geração do sinal de entrada é inerentemente não-linear, como é o caso, por exemplo, dos sinais de voz.

- Há a necessidade de pouco conhecimento estatístico sobre o ambiente no qual a rede está inserida. Outra característica extremamente importante das RNAs é que, diferentemente da análise estatística tradicional, as redes neurais não requerem prévio conhecimento sobre a distribuição dos dados para analisá-los. Desde que haja uma relação subjacente entre os dados, mesmo que desconhecida, sua representação analítica e/ou estatística, as RNAs podem apresentar um melhor desempenho do que os métodos estatísticos tradicionais. Esta característica as torna de grande utilidade pois, em muitos casos de interesse científico e/ou tecnológico é comum se estar tratando com processos sobre os quais muito pouco ou nada se conhece de seu comportamento estatístico.

- A capacidade de aprendizagem, a qual é atingida através de uma sessão de treinamento com exemplos entrada/saída que sejam representativos do ambiente. O aprendizado supervisionado, ou aprendizado obtido por meio de um tutor, envolve a modificação dos pesos sinápticos da RNA através da aplicação de um conjunto de amostras de treino, para as quais se conhece previamente a saída desejada da rede: cada exemplo consiste de um único sinal de entrada e uma correspondente resposta desejada. Um exemplo tomado aleatoriamente do conjunto de treino é apresentado à rede e os pesos sinápticos da rede (parâmetros livres) são modificados de forma a minimizar a diferença entre a resposta desejada e a resposta atual da rede, produzida pelo sinal de entrada, de acordo com algum critério estatístico apropriado. O treinamento da rede é repetido para muitos exemplos do conjunto de treino até que a rede atinja um estado onde não haja mais mudanças significativas nos pesos sinápticos. Os mesmos exemplos do conjunto de treino podem ser reaplicados durante o processo de treinamento da rede, desde que em outra ordem de apresentação

- Habilidade de aproximar qualquer mapeamento entrada-saída de natureza contínua. A aprendizagem supervisionada envolve a modificação dos pesos sinápticos de uma rede pela aplicação de um conjunto de amostras de treinamento rotulados ou exemplos da tarefa. Cada exemplo consiste de

um dado de entrada único e uma resposta desejada correspondente. Apresenta-se para a rede um exemplo escolhido ao acaso do conjunto, e os pesos sinápticos (parâmetros livres) da rede são modificados para minimizar a diferença entre a resposta desejada e a resposta real da rede, produzida pelo sinal de entrada, de acordo com um critério estatístico apropriado. O treinamento da rede é repetido para muitos exemplos dos conjunto até que a rede alcance um estado estável onde não haja mais modificações significativas nos pesos sinápticos. Os exemplos de treinamento previamente aplicados podem ser reaplicados durante a sessão de treinamento, mas em uma ordem diferente. Assim, a rede aprende dos exemplos ao construir um mapeamento de entrada-saída para o problema considerado.

- Adaptabilidade. As RNAs são ferramentas extremamente flexíveis em um ambiente dinâmico. Elas têm a capacidade de aprender rapidamente padrões complexos e tendências presentes nos dados e de se adaptar rapidamente às mudanças, características estas que são extremamente desejáveis em uma ampla gama de aplicações. As RNAs têm a capacidade de adaptar seus pesos sinápticos a mudanças no ambiente em que está inserida. Uma RNA treinada para operar em um ambiente específico pode ser facilmente retreinada para tratar com pequenas mudanças nas condições operacionais do ambiente. Quando operando em um ambiente não-estacionário (onde a estatística do processo muda com o tempo) uma RNA pode ser projetada para mudar seus pesos sinápticos em tempo real

- Generalização. Capacidade que permite às RNAs um desempenho satisfatório (produzir saídas adequadas) em resposta a dados desconhecidos (não pertencentes ao conjunto de treino, mas que estejam em sua vizinhança

- Tolerância a falhas. Característica que permite à rede continuar a apresentar resultados aceitáveis no caso de falha de alguns neurônios (unidades computacionais básicas das redes neurais artificiais). O projeto de uma RNA é motivado pela analogia com o cérebro, que é a prova viva

de que a tolerância a falhas no processamento paralelo é não apenas fisicamente possível, quanto rápida e poderosa. (Neurobiologistas utilizam RNAs como ferramentas de pesquisa para a interpretação de fenômenos neurobiológicos e engenheiros estudam neurobiologia em busca de novas ideias para resolver problemas complexos).

- Informação contextual. O conhecimento é representado pela própria estrutura da RNA e pelo seu estado de ativação. Cada neurônio da rede é potencialmente afetado pela atividade global de todos os outros neurônios na rede. Consequentemente, informação contextual é tratada com naturalidade pelas RNAs

- Possibilidade da implementação em VLSI (*Very Large Scale Integrated*). Esta característica permite considerar elevado grau de paralelismo no projeto da rede. A natureza fortemente paralela das RNAs as tornam potencialmente rápidas para computar determinadas tarefas. Esta mesma característica possibilita que sejam implementadas usando esta tecnologia VLSI.

Vários autores descreveram a teoria das redes neuronais (DE SOUZA JR, 1993; BISHOP, 1995; HAYKIN, 2001; FORTUNA et al, 2007;), incluindo diferentes abordagens e variantes, algoritmos de treinamento e áreas de aplicação. No entanto, as técnicas mais utilizadas em processos industriais são as redes feedforward do tipo MLP, com ampla capacidade de generalização, conforme a complexidade da topologia da rede (KADLEC; GABRYS; STRANDT, 2009).