Capítulo 1

Introdução

As Redes Neurais Artificiais (RNAs), objeto de nosso estudo, têm a analogia neurobiológica como fonte de inspiração.

> Uma Rede Neural Artificial (RNA) é uma estrutura computacional projetada para mimetizar a maneira pela qual o cérebro desempenha uma particular tarefa de seu interesse.

O cérebro opera de uma forma altamente complexa, não-linear e paralela. O sistema nervoso humano pode ser visto como um sistema de três estágios, conforme descrito no diagrama mostrado na Figura 1. No centro do sistema está o cérebro, representado pela rede neural, o qual recebe continuamente informações, as percebe (compreende) e toma decisões apropriadas.

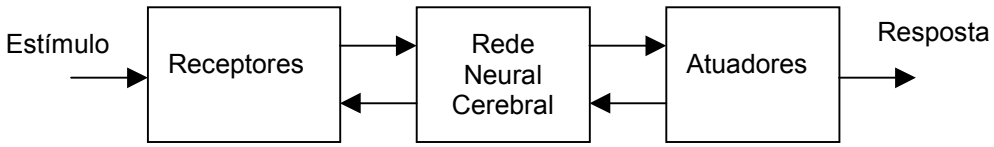

Figura 1: Representação do sistema nervoso em diagrama de blocos.

Na Figura 1, as setas apontando da esquerda para a direita indicam transmissão progressiva de sinais de informação externos, através do sistema. As setas apontando da direita para a esquerda significam a presença de realimentação no sistema. Os receptores convertem o estímulo vindo do corpo humano ou do ambiente externo em impulsos elétricos que conduzem informação para a rede neural, ou seja, o cérebro. Os atuadores convertem os impulsos elétricos gerados pela rede neural em respostas discerníveis como saídas do sistema.

Desde o pioneiro trabalho de Ramón y Cajál, em 1911, os neurônio são considerados as estruturas que constituem o cérebro. O cérebro tem a capacidade de organizar seus componentes estruturais de forma a desempenhar certas operações, tais como reconhecimento de padrões, controle de movimento, etc... muitas vezes mais rápido do que o mais rápido computador digital existente. Os neurônios são cinco a seis ordens de grandeza mais lentos do que as portas lógicas de silício; os eventos em um *chip* de silício acontecem na ordem de nanosegundos (10^{-9} s), enquanto eventos neurais acontecem na ordem de milisegundos (10^{-3} s). Entretanto, o cérebro compensa a taxa de operação relativamente lenta de um neurônio através de um inacreditavelmente grande número de neurônios, com densas interconexões entre eles. Estima-se em aproximadamente 10 bilhões de neurônios no córtex humano, e 60 trilhões de sinapses ou conexões. O cérebro é uma estrutura extremamente eficiente (a eficiência energética do cérebro é aproximadamente 10^{-16} J/operação/s, enquanto que o valor correspondente para o melhor computador existente é de aproximadamente 10^{-6} J/operação/s!).

Como qualquer célula biológica, o neurônio é delimitado por uma fina membrana celular que possui determinadas propriedades essenciais para o funcionamento elétrico da célula nervosa. A partir do corpo celular (ou soma), que é o centro dos processos metabólicos da célula nervosa, projetam-se extensões filamentares, que são os dendritos, e o axônio, conforme pode ser visto na Figura 2. Os dendritos freqüentemente cobrem um volume muitas vezes maior do que o próprio corpo celular e formam uma árvore dendrital. A outra projeção filamentar do corpo celular, o axônio, também referido como fibra nervosa, serve para conectar a célula nervosa a outras do sistema nervoso. Os axiônios são linhas de transmissão e os dendritos, zonas receptivas. O neurônio possui geralmente um único axônio, embora este possa apresentar algumas ramificações.

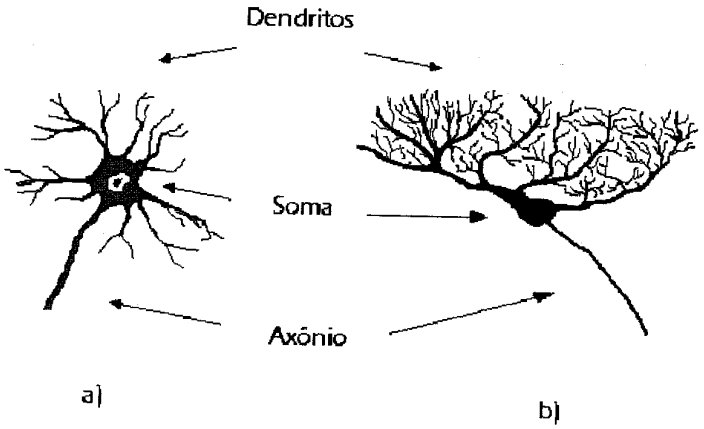

Figura 2: Neurônios do sistema nervoso central dos vertebrados: (a) Neurônio motor da célula espinhal; (b) Célula de Purkinje encontrada no cerebelo. Esta célula é notável pela extensa ramificação da sua árvore dendrital, da qual apenas uma pequena parte é mostrada na figura, podendo ultrapassar várias dezenas de vezes as dimensões do corpo celular.

O neurônio biológico é basicamente o dispositivo computacional elementar do sistema nervoso, que possui entradas - muitas entradas - e uma saída. As entradas ocorrem através das conexões sinápticas, que conectam a árvore dendrital aos axônios de outras células nervosas. Os sinais que chegam dos axônios de outras células nervosas são pulsos elétricos conhecidos como impulsos nervosos ou potenciais de ação e constituem a informação que o neurônio processará de alguma forma para produzir como saída um impulso nervoso no seu axônio. As sinapses são regiões eletroquimicamente ativas, compreendidas entre duas membranas celulares: a membrana pré-sináptica (por onde chega um estímulo proveniente de uma outra célula) e a membrana pós-sináptica (que é a do dendrito). Nesta região intersináptica, o estímulo nervoso que chega à sinapse é transferido à membrana dendrital através de susbstâncias conhecidas como neurotransmissores. O resultado desta transferência é uma alteração no potencial elétrico da membrana pós-sináptica. Dependendo do tipo de neurotransmissor, a conexão sináptica será excitatória ou inibitória. Uma conexão excitatória provoca uma alteração no potencial da membrana que contribui para a formação de um impulso nervoso no axônio de saída, enquanto que uma

conexão inibitória age no sentido oposto. Uma sinapse pode impor excitação ou inibição (uma ou outra) sobre o neurônio receptivo.

A maioria dos neurônios codifica suas saídas como uma série de breves pulsos de tensão. Esses pulsos, conhecidos como potenciais de ativação, originam-se no próprio corpo celular do neurônio (ou próximo a ele) e então se propagam através dos neurônios individuais à velocidade e amplitude constantes.

No cérebro há organizações anatômicas em pequena escala e grande escala, e diferentes funções acontecem em níveis inferiores e superiores. A Figura 3 mostra tais níveis entrelaçados de organização. As sinapses representam o nível mais fundamental, dependendo de moléculas e íons para sua atuação. Nos próximos níveis temos microcircuitos neurais, árvores de dendritos e, então, neurônios. Um microcircuito neural se refere a um agrupamento de sinapses organizadas em padrões de conectividade para produzir uma operação funcional de interesse. Um microcircuito neural pode ser comparado a um chip de silício feito do agrupamento de transístores. Os microcircuitos neurais são agrupados para formar subunidades dendríticas dentro das árvores dendríticas de neurônios individuais. O neurônio contém várias subunidades dendríticas. No próximo nível de complexidade temos os circuitos locais feitos de neurônios com propriedades similares ou diferentes; estes agrupamentos de neurônios desempenham operações características de uma região localizada no cérebro. Isto é seguido pelos circuitos interregionais feitos de caminhos, colunas e mapas topográficos, que envolvem regiões localizadas em diferentes partes do cérebro. Estes mapas topográficos são organizados para responder a informações sensoriais que chegam. Estes mapas são freqüentemente arranjados em placas que são empilhadas em camadas adjacentes, de forma que estímulos vindos de pontos correspondentes no espaço estão acima ou abaixo de outros (exemplo: mapas visuais e auditivos estão empilhados em camadas adjacentes). No nível final de complexidade, os mapas topográficos e outros circuitos interregionais mediam específicos tipos de comportamento no sistema nervoso central.

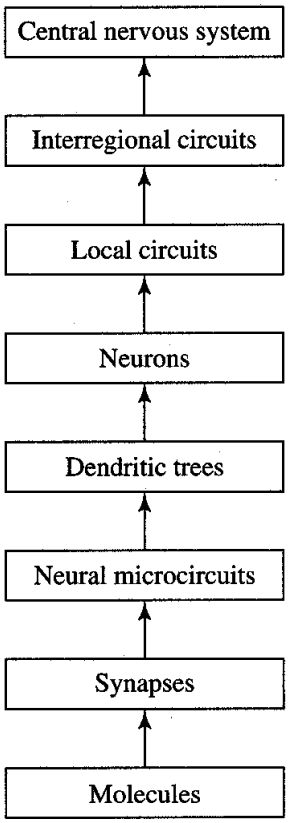

Figura 3: Organização estrutural de níveis no cérebro.

Uma rede cerebral é, portanto, um dispositivo geral de processamento. A função que a rede cerebral computa é determinada pelo padrão das conexões, ou seja, o análogo de um programa de computador baseado em algoritmos totalmente desconhecidos, que são naturalmente implementados no cérebro. Além disto, é importante reconhecer que os neurônios artificiais que usamos para construir nossas RNAs são muito primitivos, se comparados aos neurônios encontrados no cérebro, e as RNAs que somos capazes de projetar atualmente são primitivas se comparadas aos circuitos locais e circuitos interregionais 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.

Um exemplo de tarefa de processamento de informação realizada pelo cérebro é o sistema visual. É função do sistema visual prover uma representação do ambiente que nos cerca e, mais importante ainda, suprir a informação de que necessitamos para interagir com o ambiente. O cérebro rotineiramente realiza tarefas de reconhecimento perceptivo (por exemplo, o reconhecimento de um rosto familiar em meio a uma cena não-familiar) em aproximadamente 100 a 200 ms, enquanto que tarefas de complexidade muito menor podem levar dias em um computador convencional. A questão é: como o cérebro humano executa tais tarefas?

No nascimento, um cérebro tem uma grande estrutura e a habilidade de construir suas próprias regras através da experimentação continuada. A experiência é construída ao longo do tempo, com o desenvolvimento mais dramático do cérebro humano ocorrendo durante os dois primeiros anos a partir dos nascimento, mas o desenvolvimento continua muito além daquele estágio. Neurônios em desenvolvimento são sinônimos de um cérebro "plástico". Esta capacidade "plástica" permite ao sistema nervoso em desenvolvimento se adaptar ao ambiente que o cerca. Em um cérebro adulto, esta "plasticidade" pode ser responsável por dois mecanismos: a criação de novas conexões sinápticas entre neurônios, e a modificação de sinapses existentes.

Esta mesma "plasticidade" essencial ao funcionamento dos neurônios do cérebro humano como unidades de processamento de informação é utilizada pelas RNAs e seus neurônios artificiais. Assim, podemos afirmar que as RNAs assemelham-se ao cérebro humano em dois aspectos:

1. O conhecimento é adquirido pela RNA através de um processo de aprendizado.
2. As transmitâncias das conexões inter-neurônios, conhecidas como pesos sinápticos, às quais está submetido o fluxo de informações através da rede, são utilizadas para armazenar o conhecimento.

Uma RNA nada mais é, portanto, do que uma máquina projetada para modelar a maneira pela qual o cérebro desempenha funções de interesse.

A exemplo das "redes neurais naturais", as redes artificiais consistem da interconexão de um grande número de unidades de processamento chamadas neurônios. As conexões entre as unidades computacionais (ou neurônios) são chamadas sinapses ou pesos sinápticos.

A Figura 4 apresenta a arquitetura de uma rede neural artificial composta de uma camada de entrada e duas camadas de unidades de processamento, ou neurônios.

A camada de entrada, que conecta a rede ao ambiente externo, é composta por elementos chamados nós de entrada ou nós fonte da rede.

A segunda camada de neurônios ou camada escondida de neurônios é conectada à camada de nós de entrada e à camada de neurônios de saída por um conjunto de interconexões chamadas sinapses ou pesos sinápticos.

Tal rede, conforme mostrada na Figura 4, é normalmente referida na literatura como uma RNA de duas camadas (a camada de nós de entrada não conta como camada de unidades processadoras ou neurônios), ou simplesmente referida como uma RNA que apresenta apenas uma camada escondida (ficando naturalmente implícitas as camadas de nós de entrada e de neurônios de saída).

Conforme já sabemos, as RNAs têm a capacidade de obter conhecimento a partir de seu ambiente através de um processo de aprendizado. O conhecimento obtido pelas RNAs é armazenado nos parâmetros livres da rede, que são os pesos sinápticos e os parâmetros que definem a função de transferência das unidades computacionais ou neurônios.

O procedimento utilizado para o processo de aprendizado é chamado Algoritmo de Aprendizagem e tem por função modificar de forma adaptativa os parâmetros livres da rede para atingir um objetivo desejado. Em outras palavras, da mesma forma que em um filtro linear adaptativo convencional, as redes neurais artificiais têm a capacidade de, através da informação de uma resposta desejada, tentar aproximar um sinal alvo durante o processo de aprendizagem.

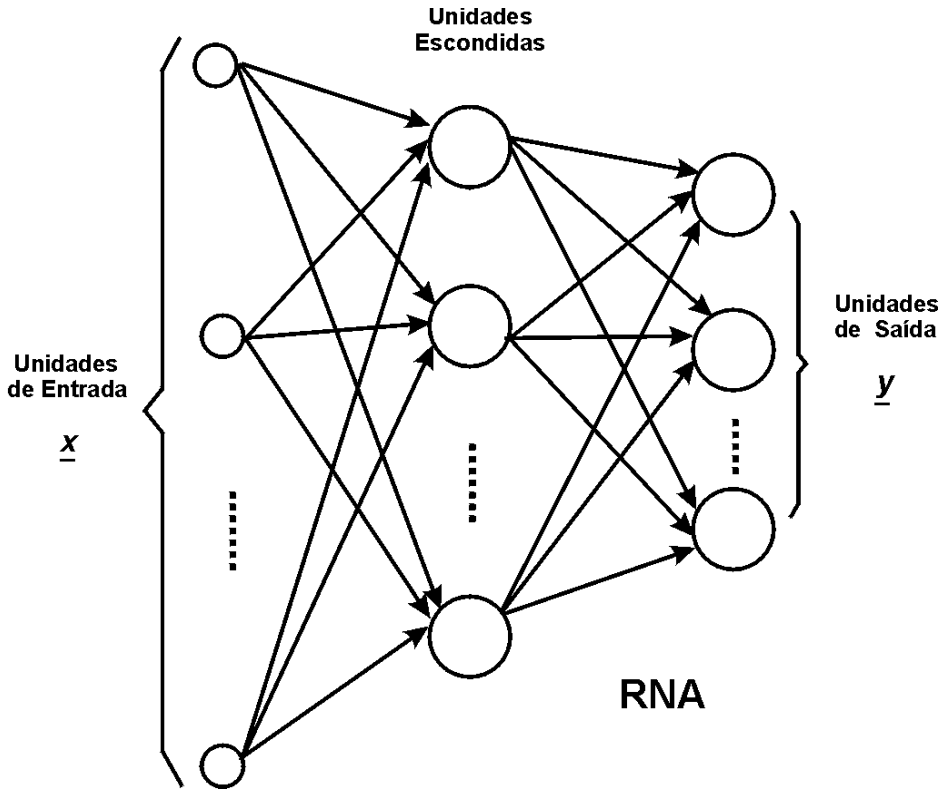

Figura 4: Exemplo de arquitetura de uma Rede Neural Artificial.

Esta aproximação é obtida através do ajuste, de forma sistemática, de um conjunto de parâmetros livres, característico de cada rede neural. Na verdade, o conjunto de parâmetros livres provê um mecanismo para armazenar o conteúdo de informação subjacente presente nos dados que são apresentados à rede na fase de treinamento.

1.1 Características Relevantes das RNAs

O poder computacional de uma RNA é devido basicamente a dois fatores: sua estrutura paralela pesadamente distribuída e sua habilidade de aprender e, conseqüentemente, generalizar.

Algumas características relevantes das Redes Neurais Artificiais são descritas por Simon Haykin em *Adaptive Filter Theory* e *Neural Networks* e aqui citadas:

- 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.

- 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.

- 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. Devido à capacidade de aprendizado, uma RNA tem a possibilidade de encontrar qualquer mapeamento entrada/saída, desde que os dados sejam adequadamente representativos do processo que esteja sendo tratado, e desde que sejam adequadamente escolhidos a arquitetura da rede e seu algoritmo de treinamento.

- Adaptatividade.

  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 idéias 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. Conseqüentemente, informação contextual é tratada com naturalidade pelas RNAs.

- Possibilidade da implementação em VLSI.

  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 tecnologia VLSI (*very-large-scale-integrated*).

1.2 Modelo de um Neurônio

O diagrama de blocos mostrado na Figura 5 apresenta o modelo básico de um neurônio utilizado no projeto de Redes Neurais Artificiais. O modelo consiste de:

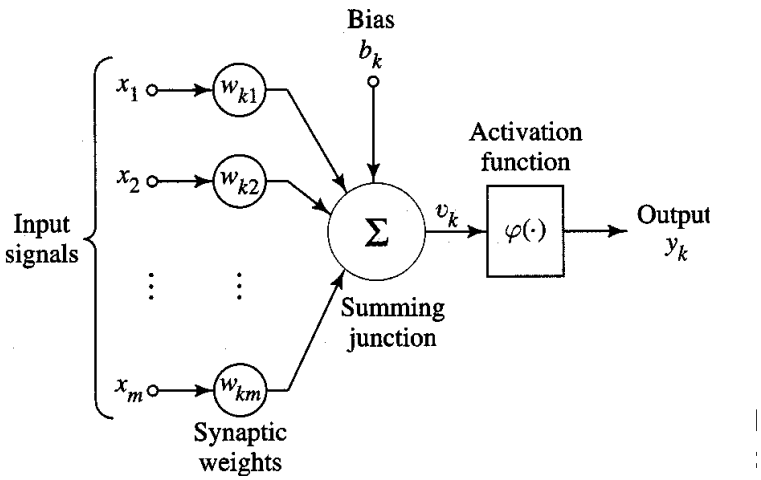

Figura 5: Modelo não-linear de um neurônio.

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 $[0,1]$ ou, alternativamente $[-1,1]$.

O modelo neural da Figura 5 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

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{1.1}$$

e

$$y_k = \varphi(u_k + b_k) \tag{1.2}$$

onde:

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

$w_{k1}, w_{k2}, ..., w_{km}$ 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*;

$\varphi(\cdot)$ 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

$$v_k = u_k + b_k \tag{1.3}$$

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 é conforme mostrada na Figura 6. Observe que, como resultado da transformação, o gráfico de $v_k \times u_k$ não passa mais pela origem.

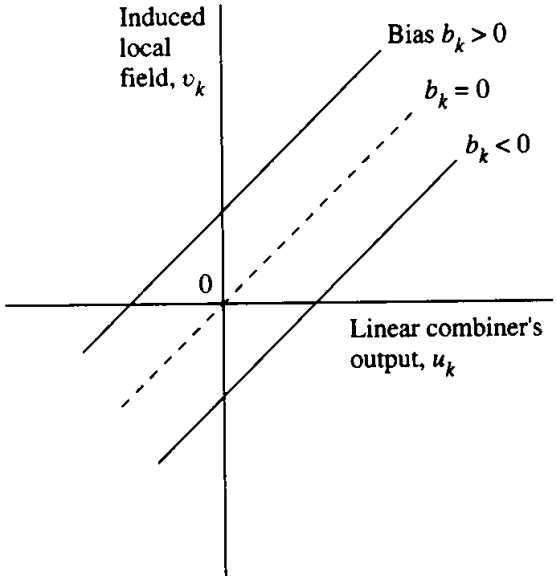

Figura 6: Transformação produzida pela polarização ou *bias* ($v_k = b_k$ para $u_k = 0$).

A polarização é um parâmetro externo do neurônio artificial *k*, conforme expressa a Equação (1.2). Uma outra forma de expressar a presença da polarização é através da combinação das Equações (1.1) e Equação (1.3),

$$v_k = \sum_{j=0}^{m} w_{kj} x_j \tag{1.4}$$

e

$$y_k = \varphi(v_k) \tag{1.5}$$

Na realidade, adicionamos uma nova sinapse na Equação (1.4), cuja entrada é $x_0 = +1$ e cujo peso é $w_{k0} = b_k$. O modelo do neurônio reformulado de acordo com as Equações (1.4) e (1.5) é mostrado na Figura 7. Embora os modelos pareçam diferentes, são matematicamente equivalentes.

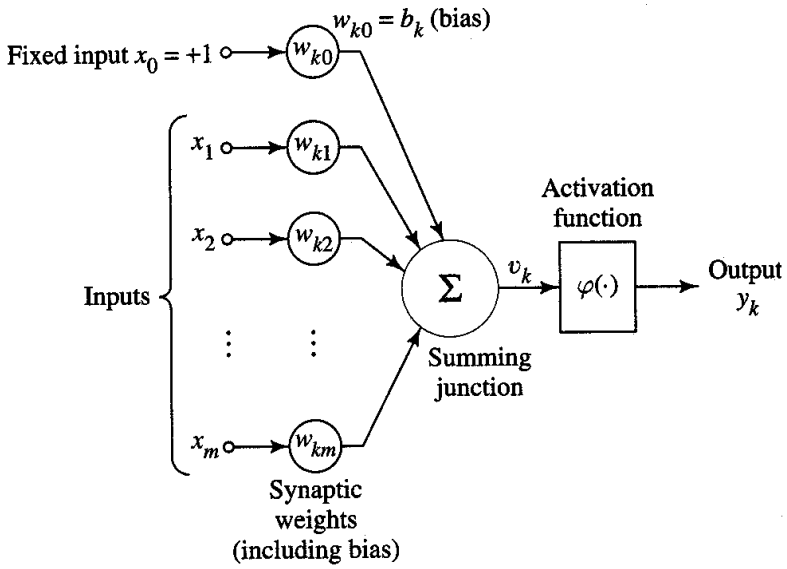

Figura 7: Outro modelo não-linear de um neurônio.

1.3 Tipos de Função de Ativação

Conforme vimos na Seção 1.2, a função de ativação $\varphi(v)$ define a saída do neurônio em termos do potencial de ativação *v*. A Figura 8 apresenta três tipos de função de ativação, a Função *Threshold*, a Função *Piecewise-linear* e a Função Sigmoide.

(a) <u>Função *Threshold* (Função Limiar):</u>

Este tipo de função de ativação, mostrado na Figura 8 (a) é descrito por

$$\varphi(v) = \begin{cases} 1 & se \quad v \geq 0 \\ 0 & se \quad v < 0 \end{cases} \tag{1.6}$$

Correspondentemente, a saída do neurônio *k*, empregando tal função *Threshold* é expressa por

$$y_k = \begin{cases} 1 & se \quad v_k \geq 0 \\ 0 & se \quad v_k < 0 \end{cases} \tag{1.7}$$

onde v_k é o potencial de ativação do neurônio, dado por

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \tag{1.8}$$

Um neurônio assim definido é conhecido como o modelo de McCulloch-Pitts. Neste modelo, a saída do neurônio assume o valor 1 se o potencial de ativação do neurônio é não-negativo e zero em caso contrário.

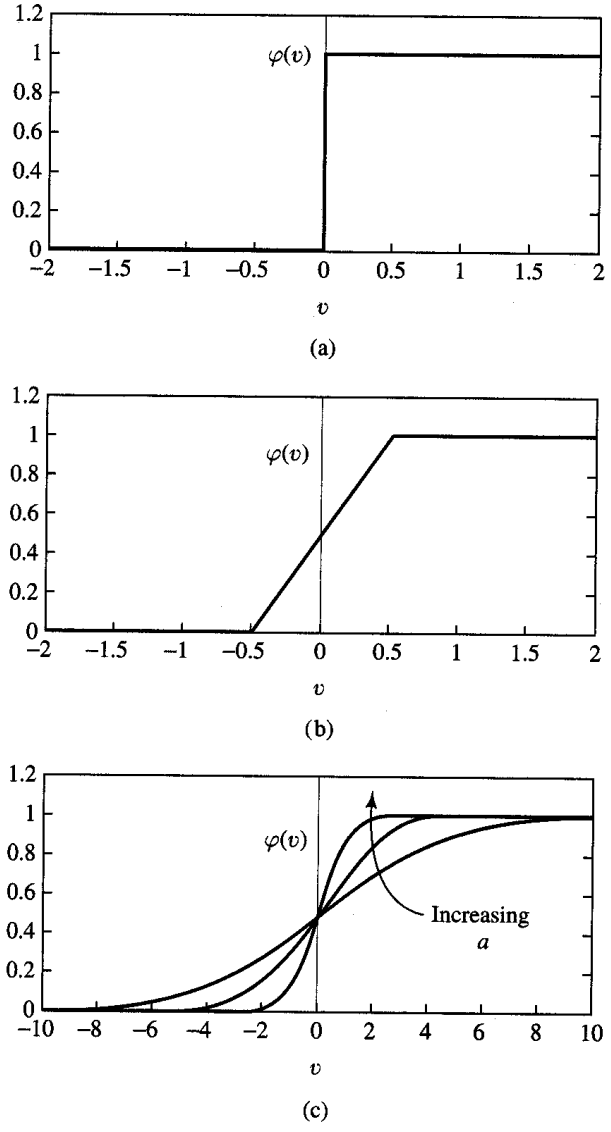

Figura 8: (a) Função *Threshold*, (b) Função *Piecewise-linear*, (c) Função Sigmoide.

(b) Função *Piecewise-linear* (Linear por Partes):

Este tipo de função de ativação, mostrado na Figura 8 (b) é descrito por

$$\varphi(v) = \begin{cases} 1 & se \quad v \geq +\frac{1}{2} \\ v & se \quad +\frac{1}{2} > v > -\frac{1}{2} \\ 0 & se \quad v \leq -\frac{1}{2} \end{cases} \tag{1.9}$$

onde o fator de amplificação dentro a região linear de operação é assumido unitário. Esta função de ativação pode ser vista como uma aproximação de uma amplificação não-linear. Duas situações podem ser vistas como formas especiais da função *Piecewise-linear*:

- Um combinador linear (se a região linear de operação não saturar);
- A função *Piecewise-linear* se reduz a uma função *Threshold* se o fator de amplificação da região linear for feito infinitamente grande.

(c) Função *Sigmoide*

Este tipo de função de ativação cujo gráfico se assemelha a uma curva em "S", é a forma de função de ativação mais utilizada na construção de RNAs. A função, mostrada na Figura 8 (c), é definida como uma função estritamente crescente que exibe um interessante balanço entre o comportamento linear e o comportamento não-linear. Um exemplo de função sigmoidal é a função logística, definida por

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \tag{1.10}$$

onde a é o parâmetro declividade da função sigmoidal. Através da variação do parâmetro a são obtidas funções sigmoidais de diferentes declividades, conforme pode ser observado na Figura 8 (c). Na verdade, a declividade na origem é igual a $a/4$. No limite, quando o parâmetro declividade se aproxima do infinito, a função sigmoidal se torna simplesmente em uma função *Threshold*. Enquanto a função *Threshold* assume o valor 0 ou 1, uma função sigmoidal assume um intervalo contínuo de valores de 0 a 1. É importante notar que a função sigmoidal é diferenciável, enquanto que a função *Threshold* não o é.

As funções de ativação *Threshold*, *Piecewise-linear* e Sigmoide são definidas no intervalo de 0 a 1. Algumas vezes é desejável funções de ativação definidas no intervalo de -1 a +1, caso em que a função de ativação assume uma forma anti-simétrica com respeito à origem; ou seja, a função de ativação é uma função ímpar do potencial de ativação. Especificamente, a função *Threshold* é agora definida por

$$\varphi(v) = \begin{cases} 1 & se \quad v > 0 \\ 0 & se \quad v = 0 \\ -1 & se \quad v < 0 \end{cases} \tag{1.11}$$

que é comumente referida como função Signum.

Para permitir que a função de ativação do tipo sigmoidal assuma valores negativos, utiliza-se a forma correspondente da função sigmoidal, a função tangente hiperbólica, definida por

$$\varphi(v) = \tanh(v) \tag{1.12}$$

1.4 RNAs vistas como Grafos de Fluxo de Sinal

Os diagramas de blocos das Figuras 5 e 7 apresentam uma descrição funcional dos vários elementos que constituem o modelo de um neurônio artificial. A idéia de introduzir os grafos de fluxo de sinal objetiva simplificar a aparência dos modelos, sem perder qualquer detalhe funcional.

Um grafo de fluxo de sinal é uma rede de ramos orientados (com sentido) que são interconectados a certos pontos chamados nós. Um nó típico *j* tem um sinal de nó associado x_j. Um típico ramo direcionado se origina no nó *j* e termina no nó *k*; e tem uma função de transferência associada (ou transmitância) que especifica a maneira pela qual o sinal y_k no nó k depende do sinal x_j no nó *j*. O fluxo de sinais nas várias partes do grafo é regido por três regras básicas:

1. Um sinal flui ao longo de um ramo somente na direção definida pela seta. Dois tipos de ramos podem ser definidos: ramos sinápticos, governados por uma relação linear

entrada/saída, em que o sinal no nó x_j é multiplicado pelo peso sináptico w_{kj} para produzir o sinal no nó y_k e ramos de ativação, governados, em geral, por uma relação não-linear entrada-saída.

2. Um sinal em um nó é igual à soma algébrica de todos os sinais que entram no nó, através dos ramos que chegam ao nó.
3. O sinal no nó é transmitido para cada ramo de saída originado no nó, com a transmissão sendo inteiramente independente das funções de transferência dos ramos que saem do nó.

A Figura 9 mostra o grafo de fluxo de sinal construído a partir das regras descritas para o modelo de neurônio correspondente ao diagrama de blocos mostrado na Figura 7.

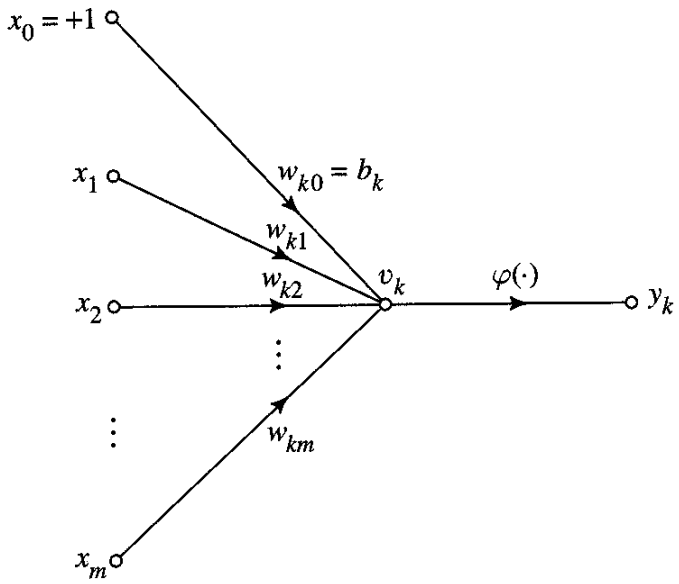

Figura 9: Representação do modelo de neurônio mostrado na Figura 7 sob a forma de grafo de fluxo de sinal.

A partir do grafo de fluxo de sinal mostrado na Figura 9, pode-se apresentar a seguinte definição matemática de uma rede neural:

Uma rede neural é um grafo de fluxo de sinal orientado, consistindo de nós com interconexões sinápticas e ramos de ativação, caracterizado por quatro propriedades:

1. Cada neurônio é representado por um conjunto de ramos sinápticos lineares, uma polarização externa aplicada, e um possível ramo não-linear de ativação. A polarização é representada por um ramo sináptico conectado a uma entrada fixa em +1.
2. Os ramos sinápticos de um neurônio ponderam seus respectivos sinais de entrada.
3. A soma ponderada dos sinais de entrada define o potencial de ativação do neurônio em questão.
4. O ramo de ativação limita o potencial de ativação do neurônio para produzir uma saída.

1.5 Arquiteturas de Redes

O projeto de uma rede neural, ou seja, a maneira pela qual os neurônios da rede são estruturados, está intimamente relacionada ao algoritmo de aprendizagem usado para treinar a rede (conforme poderemos comprovar nos capítulos seguintes). Em geral, podemos identificar três diferentes classes fundamentais de arquiteturas de redes:

Redes *Single-Layer Feedforward*:

As redes *single-layer feedforward* podem ser referidas como "redes progressivas de uma única camada". Esta arquitetura de RNAs é a forma mais simples de redes *layered*, em que os neurônios são organizados em forma de camadas. Na rede progressiva de uma única camada, temos uma arquitetura com uma camada de entrada de nós fontes conectada a uma camada de saída constituída de neurônios (nós computacionais), conforme mostrado na Figura 10.

Esta rede é estritamente progressiva, no sentido de que não há conexões no sentido camada de saída $\rightarrow$ camada de nós fontes (não há elos de realimentação entre as camadas). Conforme já comentamos na introdução deste capítulo, a rede mostrada na Figura 10 é referida na literatura como uma RNA de uma única camada (*single-layered network*), pois a

camada de nós de entrada não é contada, já que não é formada por unidades processadoras, ou neurônios.

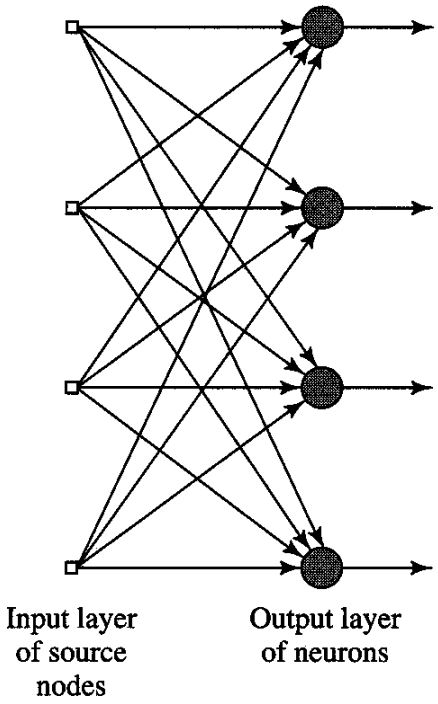

Figura 10: Rede progressiva formada por uma única camada de neurônios (representada com quatro nós na camada de entrada e quatro neurônios na camada de saída).

Redes *Multilayer Feedforward*:

A segunda classe de redes progressivas (redes neurais progressivas multicamadas) tem por característica possuir uma ou mais camadas escondidas, cujos nós computacionais ou neurônios são correspondentemente chamados de neurônios escondidos ou unidades escondidas. A função dos neurônios escondidos é intervir entre a camada externa de entrada e a saída da rede de alguma forma útil. Adicionando uma ou mais camadas escondidas, a rede pode extrair estatísticas de ordem superior. Pode-se dizer que a rede adquire uma perspectiva global apesar de sua conectividade local, devida ao conjunto extra de conexões sinápticas e à dimensão extra de interações neurais.

Os nós fonte na camada de entrada da rede provêem os vetores de entrada, que constituem os sinais de entrada aplicados aos neurônios da segunda camada (primeira camada escondida). Os sinais de saída da segunda camada são usados como entradas para a terceira camada, e, assim, sucessivamente, para o resto da rede. O conjunto de sinais de saída dos neurônios da camada de saída da rede constituem a resposta global da rede ao padrão de ativação provido pelos nós fonte na camada de entrada.

A Figura 11 ilustra uma rede neural progressiva multicamadas, para o caso de uma única camada escondida, em que cada nó de cada camada da rede é conectado a cada outro nó da camada adjacente. Neste caso, a rede é dita completamente conectada. Se, no entanto, algumas das conexões sinápticas estiverem faltando, a rede é dita parcialmente conectada.

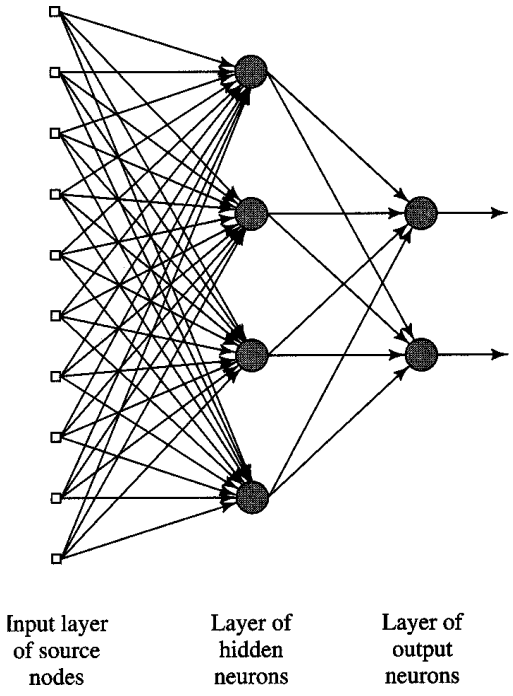

Figura 11: Rede progressiva multicamadas completamente conectada, formada por uma única camada escondida de neurônios e uma única camada de saída (representada com 10 nós fonte na camada de entrada, 4 neurônios escondidos e 2 neurônios na camada de saída).

Redes Recorrentes:

Uma rede neural recorrente difere de uma rede neural progressiva (*feedforward*), pelo fato de possuir pelo menos um *loop* de realimentação (*feedback loop*).

Por exemplo, uma rede recorrente pode consistir de uma única camada de neurônios, em que cada neurônio alimenta seu sinal de saída de volta para as entradas de todos os outros neurônios, conforme ilustra a Figura 12.

Já a Figura 13 ilustra uma rede recorrente em que há uma camada de neurônios escondidos e em que as conexões de realimetação são originadas tantos dos neurônios escondidos, quanto dos neurônios de saída.

A presença de *loops* de realimentação em estruturas recorrentes tem um grande impacto na capacidade de aprendizagem da rede e em seu desempenho.

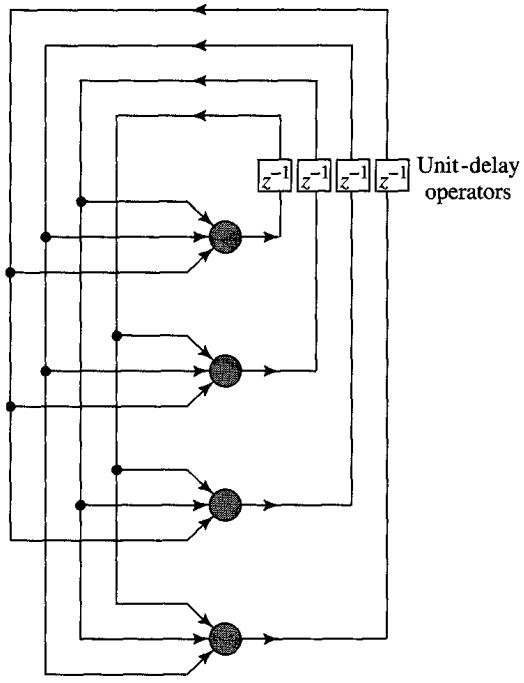

Figura 12: Rede recorrente em que não há *loops* auto-realimentados, nem neurônos escondidos.

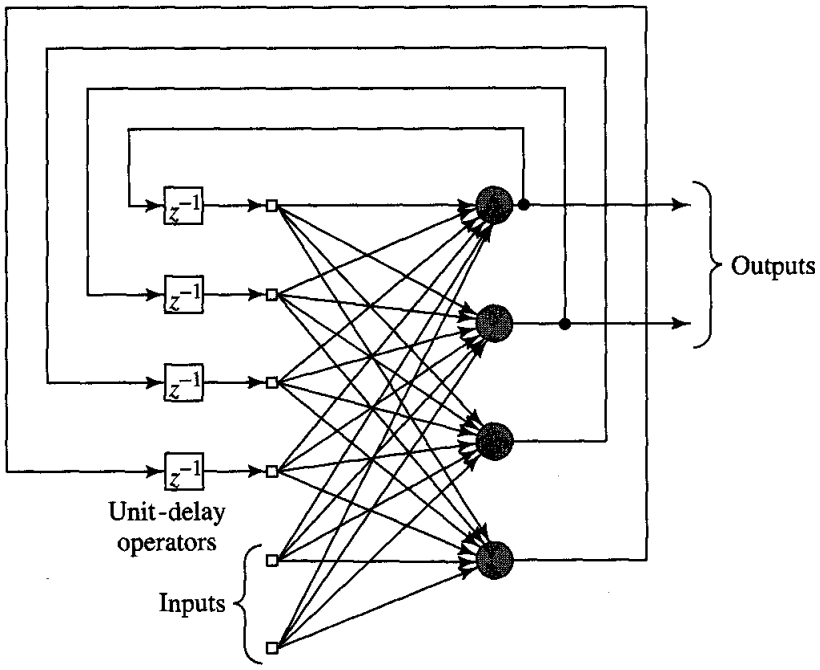

Figura 13: Rede recorrente com neurônios escondidos e *loops* auto-realimentados.

1.6 Referências Bibliográficas do Capítulo 1:

[1] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice Hall, Upper Saddle River, New Jersey, 1996.

[2] S. Haykin, *Neural Networks*, 2nd ed., Prentice Hall, Upper Saddle River, New Jersey, 1999.

[3] [3] Z.L.Kovács, *Redes Neurais Artificiais*, Editora Acadêmica São Paulo, São Paulo, 1996.