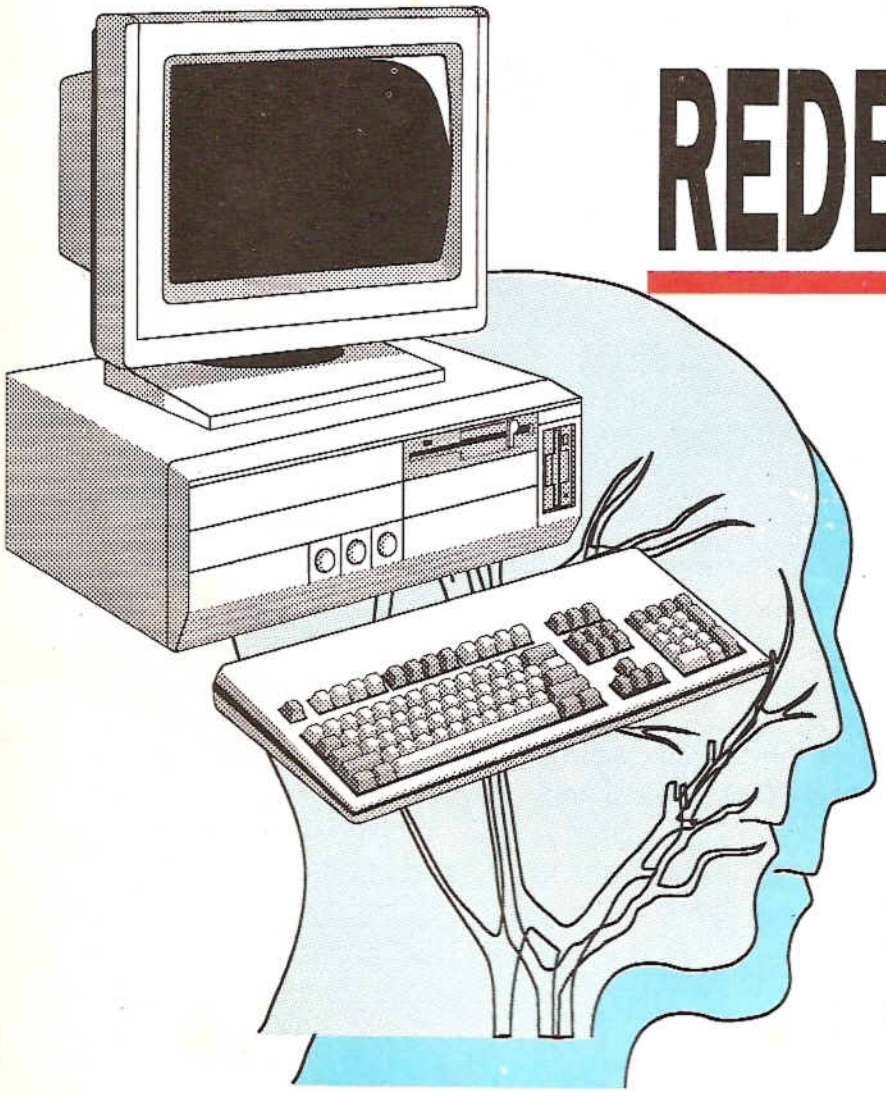
SEÇÃO Bitmap

ANO XIII - Nº133 - CR$ 780,00

Micro Sistemas

A PRIMEIRA REVISTA BRASILEIRA DE MICROCOMPUTADORES



REDES NEURAIS

COMO INSTALAR NOVOS DISPOSITIVOS NO SEU PC

E MAIS

MULTIPLICAÇÃO POR TRÊS

CLASSE WINDOW EM CLIPPER 5.x

CAPA

Redes Neurais Artificiais

Uma abordagem revolucionária em Inteligência Artificial

Antonio Augusto Gorni

É apresentada neste trabalho a técnica das redes neurais artificiais, que consiste na emulação de sistemas nervosos biológicos em programas ou circuitos digitais. Tais redes são capazes de estabelecer relações complexas entre dados, de forma relativamente precisa, sem que seja necessário informar a elas qual a função que os relaciona. A seguir, é feita uma breve descrição dos aspectos gerais da neurocomputação, seguida de uma retrospectiva histórica, comparação com as demais técnicas utilizadas em modelamento de processos e recomendações quanto a sua implementação e aplicação.

INTRODUÇÃO

Desde o advento do primeiro computador digital efetivamente útil – o ENIAC, em 1946 – até o final da década de 1980, praticamente todas as aplicações para processamento de dados e informações adotaram uma única abordagem básica: a **computação programada**. Ela consiste no desenvolvimento prévio de um algoritmo (ou seja, um conjunto detalhado e rígido de regras previamente estabelecidas) para a resolução do problema em questão, o qual era codificado em alguma linguagem de computação. Isto implicou no fato de que tal abordagem somente poderia ser utilizada em casos onde o processamento a ser efetuado pudesse ser descrito em termos de um conjunto de regras conhecido. Contudo, nem sempre isso acontece; muitas vezes a dedução desse conjunto de regras pode ser difícil. Além disso, uma vez que os computadores atuais trabalham de forma totalmente lógica, o programa final tem de estar praticamente perfeito para poder funcionar a contento. Logo, o desenvolvimento de programas para computador é, na verdade, uma sucessão de ciclos "projeto-teste-melhoria interativa", que pode vir a demandar muito tempo, esforço e dinheiro.

No final da década de 1980 surgiu uma abordagem revolucionária para o processamento de dados e informações: a neurocomputação ou, como é mais conhecida, as redes neurais artificiais, também conhecidas como "perceptrons". Ela não requer o desenvolvimento de algoritmos ou conjunto de regras para analisar os dados, o que freqüentemente reduz de forma significativa o trabalho de desenvolvimento de programas que uma dada aplicação venha a requerer. Na maior parte dos casos, a rede neural passa por um processo de treinamento a partir de casos reais conhecidos, adquirindo a partir daí a sistemática necessária para executar adequadamente o processamento desejado dos dados fornecidos. Ou seja: ela tem capacidade de extrair as regras básicas desejadas a partir de dados reais, dispensando qualquer modelo prévio já conhecido. Esta é a abordagem utilizada nos sistemas nervosos biológicos, particularmente em seres humanos [1].

Com efeito, pode-se perceber melhor a diferença entre a computação programada e as redes neurais comparando-se computadores e seres humanos. Por exemplo, um computador efetua operações matemáticas com rapidez e precisão muito superiores aos seres humanos. Em contrapartida, estes conseguem reconhecer faces e imagens complexas de maneira muito mais precisa, eficiente e rápida que o melhor computador disponível atualmente.

Uma das razões dessa diferença de desempenho em tarefas diversas pode estar na forma como se organizam computadores e sistemas nervosos. Geralmente um computador consiste de um processador trabalhando sozinho, executando instruções fornecidas por um programador, uma a uma. Já sistemas nervosos consistem de bilhões de células nervosas - ou seja, neurônios - com alto grau de interconexão entre elas, que efetuam cálculos simples sem que haja a necessidade de que sejam programadas [1,2].

Logo, as redes neurais artificiais são um modo de se simular e tentar entender o que se passa nos sistemas nervosos biológicos, na esperança de se conseguir tomar

proveito dos poderosos recursos desses sistemas orgânicos. Esta técnica revolucionária de Inteligência Artificial, apesar de ainda se encontrar em pleno desenvolvimento, já se encontra num estágio suficientemente adiantado para ser bastante útil em aplicações que vão desde a análise para a aprovação de crédito pessoal até o controle de processos industriais. O desenvolvimento de microcomputadores cada vez mais poderosos tornou essa técnica plenamente acessível: o requisito mínimo para que ela possa ser aplicada é que se disponha de um microcomputador IBM-AT 286.

FUNDAMENTOS DAS REDES NEURAIS ARTIFICIAIS

O elemento básico que constitui uma rede neural artificial chama-se, naturalmente, **neurônio**, conhecido ainda por nó ou elemento processador. Ele pode ser visto esquematicamente na figura 1.

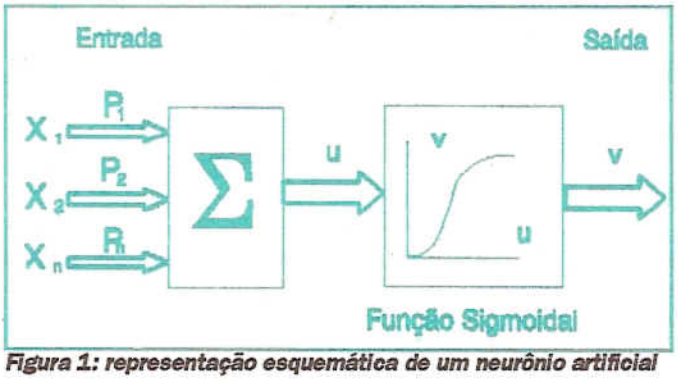


Figura 1: representação esquemática de um neurônio artificial

Usualmente os neurônios estão dispostos em camadas. Geralmente estão ligados a outros neurônios, em camadas anteriores ou posteriores. Tais ligações são chamadas de sinapses. Ocasionalmente pode haver ligações entre neurônios de uma mesma camada ou mesmo entre eles próprios. Os sinais fornecidos a um dado neurônio são, na verdade, o estado ou o valor de ativação dos neurônios precedentes, os quais são multiplicados por um peso correspondente. O estado, ou valor de ativação, do neurônio em questão, é calculado a partir da aplicação de uma função de limiar ao valor de entrada fornecido ao neurônio, ou seja, a somatória dos valores de ativação dos neurônios precedentes, multiplicados pelos respectivos pesos.

A função de limiar (ou de ativação) geralmente é algum tipo de função não-linear, a qual garante a plena funcionalidade das redes neurais com múltiplas camadas. Conforme mostrado na figura 1, ela deve ser uma função com formato sigmoidal no caso de redes neurais que processam dados analógicos, como, por exemplo, tangente hiperbólica, seno, funções logísticas, etc. Já para redes neurais discretas – ou seja, cujos neurônios assumem apenas dois valores: zero ou um – ela deve ser uma função degrau [3].

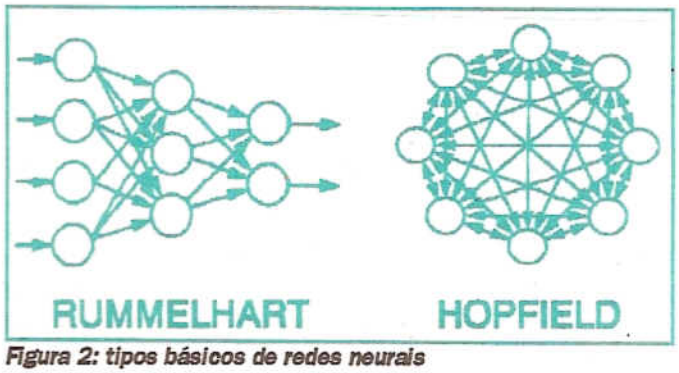


Figura 2: tipos básicos de redes neurais

Nos primordios de seu desenvolvimento, as redes neurais eram emuladas através de circuitos analógicos. A energia correspondente às conexões entre os neurônios (ou seja, a sinapse) era controlada por potenciômetros. Atualmente, é mais comum a utilização de "software" para a geração de redes neurais, ou seja, o funcionamento dos neurônios é simulado por programas. Uma tendência bastante recente é a implementação de redes neurais em circuitos digitais, visando à maximização de sua velocidade de processamento. Isto é conseguido através de arquiteturas digitais que permitam o processamento de dados de forma paralela. Normalmente elas são fornecidas na forma de placas que podem ser acopladas a micro-computadores ou estações de trabalho.

De forma bastante geral, pode-se subdividir as redes neurais em duas categorias básicas, em função de como são dispostos os neurônios e de como é efetuado seu treinamento. Ou seja:

- Em função do fluxo de dados: elas podem propagá-los unidirecionalmente, apenas para a frente ("feed-forward networks") ou nos dois sentidos ("feedback network").
- Em função da ausência ou não de supervisão durante o treinamento da rede neural.

A figura 2 mostra os dois tipos básicos de redes neurais em função do fluxo de dados e tipo de treinamento [4].

A primeira delas, do tipo **Rummelhart**, é aquela onde o fluxo dos dados se dá apenas em um sentido – ou seja, são unidimensionais. Elas são muito utilizadas devido a sua simplicidade e estabilidade, sendo aplicadas para classificação, análise e interpolação de dados, o que as torna particularmente adequadas para o modelamento de processos em geral.

Uma caraterística fundamental deste tipo de rede é a disposição dos neurônios em camadas. Deve haver no mínimo duas camadas: a de entrada (de dados) e a de saída (de resultados). Uma vez que o desempenho de redes neurais deste tipo que contenham apenas duas camadas é muito limitado, normalmente se inclui pelo menos uma camada intermediária entre as duas, também chamada de oculta. Cada neurônio está ligado a todos os neurônios das camadas vizinhas, mas não há ligações entre os neurônios de uma mesma camada. O comportamento deste tipo de rede é estático; ela se comporta de modo a tornar sua saída um reflexo da respectiva entrada. Ela deve ser treinada de modo a produzir os resultados desejados a partir da apresentação de dados reais.

A outra rede neural da figura 2, do tipo **Hopfield**, se caracteriza por apresentar fluxo de dados multidirecional. Seu comportamento é dinâmico, mais complexo que o das redes de Rummelhart, mas freqüentemente apresentam

complicações inesperadas. Note-se que neste caso não há camadas discretas de neurônios: há total integração entre os dados de entrada e os resultados obtidos, pois todos os neurônios são ligados entre si. Tais redes são aplicadas em estudos sobre a otimização de conexões como, por exemplo, para se determinar o percurso ótimo de um caminhão de entregas ou um vendedor. Seu treinamento é feito de modo a minimizar a energia da rede, gerando um comportamento independente [4].

Todo o "conhecimento" de uma rede neural está armazenado em suas sinapses, ou seja, nos pesos relativos às conexões entre os neurônios. Ele é adquirido por um processo de treinamento, que consiste na apresentação de seqüência de dados conhecidos à rede neural, gerando um processo de ajuste dos pesos das sinapses de forma a capturar o "conhecimento". O treinamento pode ser supervisionado ou não. No primeiro caso, é incluída no processo de treinamento uma etapa de verificação dos resultados calculados pela rede neural em treinamento. O erro observado a partir da comparação de resultados reais com os calculados pode ser utilizado para ajustar os pesos das sinapses de forma a aumentar a precisão da resposta da rede.

Nas redes neurais onde o treinamento não é supervisionado ocorre um processo interno de organização dos dados que leva a um grau ótimo de ajuste. Essas redes auto-organizadas podem, por exemplo, dividir dados fornecidos em categorias, em função do grau de similaridade entre eles, de forma totalmente automática.

No presente trabalho serão consideradas em detalhe apenas as redes neurais do tipo Rummelhart, ou seja, unidimensionais, com fluxo unidirecional de dados e treinamento supervisionado. No momento, esta é a rede neural mais utilizada nos diversos campos da tecnologia, inclusive em controle de processos.

Como já foi visto anteriormente, as aplicações típicas para redes neurais são aquelas onde a implementação de algoritmos ou regras de controle é impossível ou desaconselhável. Há casos onde se desconhece um algoritmo adequado; em outros, o desenvolvimento de um programa para implementá-lo pode ser custoso e demorado. As aplicações que se enquadram nesse contexto e são particularmente adequadas para a aplicação de redes neurais são as seguintes:

- Análise e processamento de sinais;
- Controle de processos;
- Robótica;
- Classificação de dados;
- Reconhecimento de padrões em linhas de montagem;
- Filtros contra ruído eletrônico ("smoothing");
- Análise de imagens;
- Análise de voz;
- Diagnóstico médico;
- Previsões do mercado financeiro;
- Avaliação de solicitações de crédito ou empréstimos;
- "Marketing" orientado;

e muitas outras mais.

BREVE HISTÓRICO DO DESENVOLVIMENTO DAS REDES NEURAIS

Uma vez que o conceito e as características atuais das redes neurais já foram adequadamente definidos, é interessante verificar como ocorreu a concepção e aperfeiçoamento desta técnica.

Entre o final do século XIX e o início do século XX, pesquisadores como SIGMUND FREUD e WILLIAM JAMES estabeleceram as bases do funcionamento da neurologia dos seres vivos e, em particular, do ser humano.

Em 1943, WARREN McCULLOUGH e WALTER PITTS estabeleceram as bases da neurocomputação, concebendo procedimentos matemáticos análogos ao funcionamento dos neurônios biológicos. Este desenvolvimento foi puramente conceitual, uma vez que esses autores não sugeriram aplicações práticas a partir de seu trabalho, mesmo porque os sistemas propostos por eles não tinham capacidade de aprendizado. Note-se que o desenvolvimento dos computadores naquela época era extremamente rudimentar. Entretanto, através de cálculos manuais, foram propostos modelos muito simplificados de redes neurais artificiais, os quais podiam efetuar operações aritméticas ou lógicas.

DONALD HEBB, em 1949, sugeriu um modo de se proporcionar capacidade de aprendizado às redes neurais artificiais. Através de experiências com animais, ele propos que as mudanças nas forças das sinapses (ou seja, nas conexões entre os neurônios) são proporcionais às ativações dos mesmos. Este princípio, traduzido matematicamente, viabilizou o desenvolvimento de redes neurais artificiais eficazes.

Esses princípios foram aplicados por MARVIN MINSKY na construção do Snark, o primeiro neurocomputador de que se tem notícia, em 1951. Tecnicamente ele foi um sucesso, uma vez que ele ajustava automaticamente os pesos entre as diversas sinapses, ou seja, demonstrava, ao menos teoricamente, que tinha capacidade de aprendizado. Contudo, este dispositivo nunca executou qualquer função útil no campo do processamento de informações, constituindo-se portanto numa curiosidade acadêmica.

Em 1957, FRANK ROSENBLATT concebeu um dispositivo denominado de perceptron, ou seja, uma rede neural automatizada, com duas camadas de neurônios, capaz de aprender de acordo com as regras propostas por HEBB. Tal aparelho, o Mark I, tinha capacidade de treinamento supervisionado e foi utilizado com sucesso no reconhecimento de caracteres. Pela primeira vez, as redes neurais artificiais foram utilizadas com sucesso numa aplicação prática. O êxito conseguido por esta abordagem fez com que muitos considerem ROSENBLATT como o verdadeiro pai da neurocomputação.

BERNARD WIDROW, em 1962, desenvolveu um tipo diferente de processador para redes neurais, denominado Adaline, o qual dispunha de uma poderosa estratégia de aprendizado. Ele ainda fundou a primeira empresa comercial para a produção de circuitos neurais digitais, a *Memistor*

Corporation, que operou até meados da década de 1960.

Contudo, após esses espetaculares desenvolvimentos, a neurocomputação entrou numa grande crise. O perceptron de ROSENBLATT, que dispunha de apenas duas camadas de neurônios, era incapaz de realizar funções mais complexas que o reconhecimento de caracteres. Ele propos como solução aumentar o número de camadas mas, apesar de toda sua visão e perspicácia neste campo, não logrou desenvolver um método de aprendizado eficaz para essas redes neurais mais avançadas. A verdade é que o desenvolvimento das redes neurais artificiais até aquele momento tinha ocorrido de forma predominantemente empírica. Havia ainda uma enorme carência de embasamento teórico e matemático formal para que sua evolução pudesse continuar de forma segura.

Este impasse incentivou o início de uma campanha orquestrada por grupos de pesquisa rivais para se suprimir as verbas concedidas ao desenvolvimento das redes neurais. Em 1967 ela logrou sucesso: as linhas de financiamento originalmente atribuídas ao estudo da neurocomputação foram realocadas para outras áreas de pesquisa dentro do campo da Inteligência Artificial.

Apesar do descrédito gerado sobre a área da neurocomputação, entre 1967 e 1982 os estudos neste campo continuaram, ainda que englobadas em outras linhas de pesquisa, como processamento adaptativo de sinais, reconhecimento de padrões, modelamento biológico, etc. Este trabalho, ainda que silencioso, construiu as bases necessárias para que o desenvolvimento das redes neurais pudesse continuar de forma consistente.

Em 1974, PAUL WERBOS conseguiu o maior progresso em termos de redes neurais desde o perceptron de ROSENBLATT: ele lançou as bases do algoritmo de retropropagação ("back-propagation"), que permitiu que redes neurais com múltiplas camadas apresentassem capacidade de aprendizado. Em 1982, DAVID PARKER desenvolveu um método similar, de forma aparentemente independente. Contudo, a potencialidade desse método tardou a ser reconhecida.

O desenvolvimento de uma base teórica sólida sobre as redes neurais artificiais e o vertiginoso desenvolvimento dos microprocessadores digitais fizeram com que IRA SKURNICK, um gerente de pesquisas ligado ao Departamento de Defesa dos Estados Unidos, lançasse oficialmente um programa de desenvolvimento para a neurocomputação, em 1983. A atitude corajosa de SKURNICK influenciou fortemente outros Centros de Pesquisa, que já queriam retomar os estudos neste campo, mas hesitavam devido ao descrédito que havia sobre ele há mais de quinze anos atrás.

Os primeiros resultados da retomada do desenvolvimento sobre redes neurais foram publicados em 1986 e 1987, onde ficou consagrada a técnica de treinamento por retropropagação. Nesta fase destacaram-se o surgimento do treinamento não-supervisionado, proposto por TEUVO KOHONEN em 1984 e as novas topologias de redes neurais, como a proposta por BERNARD KOSKO, em 1987 *(B.A.M.: "Bi-Directional Associative Memory")*.

Todo esse desenvolvimento resultou no advento de programas comerciais para microcomputadores e estações de trabalho para o desenvolvimento e implementação de redes neurais com diversas topologias. Isso levou à popularização desta técnica e, de fato, ela está sendo crescentemente aplicada nos mais diversos campos do conhecimento humano. Contudo, a pesquisa sobre redes neurais artificiais continua em ritmo acelerado, uma vez que ainda há uma grande carência de formalismo matemático para que se possa explicar consistentemente seu modo de operação. Além disso, o nível de complexidade das novas topologias vem aumentando dia a dia, o que promete seu uso em aplicações cada vez mais intrincadas e desempenho cada vez melhor.

De toda esta evolução há alguns pontos que devem ser destacados. É interessante notar que tópicos aparentemente modernissimos, como redes neurais e reconhecimento de caracteres, na verdade vem sendo pesquisados já há várias décadas. Pode-se também perguntar por que a computação programada se desenvolveu mais rapidamente que a neurocomputação, uma vez que ambas foram concebidas na mesma época. Isto pode ser explicado pelo fato de que as redes neurais, com sua multiplicidade de sinapses e complicados algoritmos de aprendizado, requerem computadores de alto desempenho para que possam ser desenvolvidas, equipamentos esses que só se tornaram facilmente disponíveis a partir do final da década de 1980, com o advento da tecnologia VLSI para a fabricação de circuitos para computador. Já a computação programada apresentou desempenho muito bom desde o surgimento de máquinas rudimentares como o ENIAC, em 1946, e por isso, desde então, se tornou a abordagem preferida para o processamento comercial de dados e informações [1,3].

CARACTERÍSTICAS DO DESEMPENHO DAS REDES NEURAIS

Em sua essência, redes neurais tendem a trabalhar com os dados de uma forma inteiramente diferente que sistemas baseados em algoritmos ou conjuntos de regras. Enquanto que as redes neurais processam dados com base em transformações, a computação programada faz uso de algoritmos e regras. A experiência tem mostrado que essas duas abordagens de processamento da informação são complementares do ponto de vista operacional, mas incompatíveis em termos conceituais.

De fato, as redes neurais não prescindem totalmente da computação programada. Afinal, em qualquer aplicação, a aquisição dos dados, sua formatação e a saída dos resultados ainda utiliza esse recurso. Redes neurais são geralmente tratadas como "procedures software-callable", ou seja, sub-rotinas que podem ser amalgamadas com programas onde sua capacidade se faz necessária [1].

De modo geral, o uso de redes neurais apresenta as seguintes vantagens [3]:

– Há menor necessidade de se determinar a priori quais são os fatores determinantes sobre o modelo que está

sendo desenvolvido;
- É permitida a interferência de múltiplos fatores de entrada (ou seja, múltiplas variáveis), permitindo um inter-relacionamento muito mais complexo entre elas;
- Alta tolerância a falhas, uma vez que é permitida a entrada de grande número de parâmetros;
- Modelamento direto do problema, sem a necessidade de se seguir um modelo preestabelecido, como no caso da regressão estatística;
- Paralelismo inerente: cada sinapse na rede neural pode ser seu próprio processador.

De fato, certas características das redes neurais, como tolerância a falhas, robustez e capacidade de implementar uma classe particular de transformações, são garantidas por teoremas matemáticos. Ou seja, eles asseguram que as redes neurais podem ser empregadas de forma útil e confiável.

Por outro lado, esses teoremas nada afirmam sobre como (em termos conceitualmente mais altos) a rede neural apreende o "conhecimento". Suspeita-se que a descoberta desse mecanismo requeira uma verdadeira revolução intelectual na área do processamento de informações. Essa falta de embasamento teórico das redes neurais ainda é uma séria desvantagem desta técnica, uma vez que gera alguma desconfiança por parte dos especialistas quanto a sua confiabilidade. De fato, como já foi visto, no passado tal fato levou a uma virtual paralisação no desenvolvimento desta área devido a problemas matemáticos aparentemente insolúveis. Note-se, contudo, que atualmente a utilização de redes neurais está consolidada em diversas áreas, a pesquisa básica nesse campo é febril e os avanços teóricos são animadores [1].

Uma outra grande desvantagem das redes neurais é o tempo requerido em sua fase de aprendizado, particularmente nas do tipo Rummelhart, que utilizam o método da retropropagação. A rigor, tais tempos podem tender ao infinito. Quanto mais sutis as relações entre as variáveis, e maior a precisão requerida nos resultados, maior será o tempo de treinamento. Em alguns casos críticos, poderão ser necessários dias de treinamento, mesmo utilizando-se microcomputadores tão avançados quanto um IBM-AT/386 com "clock" de 33 MHz. Mas mesmo este problema vem sendo resolvido, e em duas frentes:
- Através de algoritmos de treinamento tão sofisticados a ponto de serem patenteados;
- Através da utilização de arquitetura digital paralela nos circuitos digitais.

COMPARAÇÃO COM OUTRAS TÉCNICAS DE MODELAMENTO

Redes neurais, sistemas especialistas e lógica difusa ("fuzzy logic") são técnicas de Inteligência Artificial bastante diferentes entre si, bem como da programação tradicional.

Sistemas Especialistas são diferentes da programação tradicional, uma vez que a base de conhecimento é separada do processador de conhecimentos (ou seja, o motor de inferência). Isto permite que se adicione conhecimento suplementar sem que seja necessário reprogramar o sistema.

Esta técnica requer que se disponha de um especialista na área de conhecimento em questão, para que ele defina as regras que codifiquem a informação pertinente. Freqüentemente se associa um grau de confiança a cada regra estabelecida - por exemplo, 95%.

Já as redes neurais podem ser implementadas sem que haja a necessidade de se impor regras explícitas ou conhecimento formal prévio. Muito pelo contrário, a rede neural aprende as regras necessárias para lidar com o conhecimento, através do ajuste dos pesos relativos às sinapses, de modo a minimizar o erro entre os dados reais e os calculados pela rede. Pode-se encarar a rede neural como um modelo muito generalizado cuja parametrização se dá através do ajuste dos pesos das sinapses. Logo, o desenvolvimento de uma rede neural prescinde do trabalho intensivo de um especialista na matéria em questão, embora ele seja necessário para definir os critérios de seleção e preparação dos dados a serem fornecidos ao treinamento da rede neural.

Além disso, há uma diferença marcante entre as redes neurais e os sistemas especialistas: nem sempre é possível saber como uma rede neural chegou a um dado resultado, enquanto que um sistema especialista sempre tem condições de dar essa informação. Isso se torna crítico para problemas muito complexos.

Em certos casos torna-se interessante desenvolver sistemas híbridos de inteligência artificial que empregam redes neurais juntamente com sistemas especialistas ou de lógica difusa [4-7]. Estes casos, que serão descritos no decorrer deste trabalho, se caracterizam por aproveitar, de forma otimizada, as melhores características de cada abordagem.

Comparando-se as redes neurais com técnicas estatísticas – como, por exemplo, regressão - fica evidente uma vantagem da neurocomputação, ou seja: não há, a princípio, a necessidade de se determinar a priori quais variáveis são importantes. As redes neurais tendem a determinar automaticamente quais são as importantes. Parâmetros irrelevantes são anulados através das reduções das energias de suas conexões com os demais neurônios a que estão ligados. Ou seja, os valores correspondentes dos pesos das sinapses são muito reduzidos. Além disso, não é necessário impor uma função como modelo, condição imprescindível para o uso da regressão estatística.

Outra vantagem das redes neurais em relação à regressão estatística é sua robustez, ou seja, maior grau de imunidade a ruídos nos dados ou falhas parciais no equipamento. Aliás é até interessante treinar as redes neurais com dados contendo ruído, para que sua precisão aumente quando em uso sob condições reais.

As características acima permitem que as redes neurais incluam o maior número possível de variáveis de entrada no modelo. Do mesmo modo que parâmetros irrelevantes são progressivamente eliminados, são criadas interações múltiplas mais complexas e sutis entre as variáveis

importantes, levando à maximização da precisão no modelo. Isto é particularmente útil em modelos com mais de três variáveis de entrada envolvidas [1].

Por outro lado, o uso de técnicas estatísticas pode ser muito útil na análise preliminar dos dados, refinando a informação a ser fornecida à rede neural e, desse modo, promover a minimização do tempo e esforço requeridos em seu desenvolvimento, aumentando a precisão do modelo final. Há casos de modelos mistos onde se utilizam, de forma conjunta, redes neurais e regressão estatística, onde a primeira técnica é utilizada para se calcular os coeficientes que ajustam uma equação empírica, de acordo com o conjunto de dados fornecido [8].

IMPLEMENTAÇÃO DAS REDES NEURAIS.

APLICAÇÕES

A relativa facilidade do uso das redes neurais tem feito com que especialistas em áreas específicas do conhecimento prático as tenham utilizado com grande sucesso. De fato, é mais fácil aprender neurocomputação do que outras disciplinas. Isso é o inverso do que ocorre em outros campos da Inteligência Artificial, onde apenas "engenheiros do conhecimento" conseguem aplicar tais tecnologias de forma eficaz. A situação poderá melhorar ainda mais no futuro, a medida que se for acumulando conhecimento prático sobre as características da aplicação de redes neurais a problemas reais. Essas informações poderão ser incorporadas aos aplicativos de neurocomputação, e o usuário poderá até esquecer os detalhes técnicos relativos às redes neurais, concentrando-se apenas nas minúcias do seu problema específico. No momento, esse modo transparente de trabalho ainda não está plenamente disponível. Ainda é altamente recomendável que o usuário tenha um bom nível de conhecimento sobre o método.

Infelizmente, o embasamento teórico sobre redes neurais ainda apresenta muitas lacunas, o que torna seu uso muito empírico até o momento. Ele requer uma combinação de planejamento e pesquisa cuidadosos, "adivinhações" feitas com bom senso e várias tentativas até se atingir um resultado eficaz.

De modo geral, as redes neurais são um método de modelamento altamente recomendável para se lidar com sistemas abertos ou mais complexos, pouco entendidos e que não podem ser adequadamente descritos por um conjunto de regras ou equações. Outras tarefas indicadas para esta técnica são aquelas que requeiram tolerância a falhas, onde haja dados contaminados com ruído, que envolvam reconhecimento ou detecção de padrões, diagnóstico, abstração ou generalização. Situações onde os dados de entrada estejam incompletos ou ambíguos são um campo de atuação atualmente recomendado para a neurocomputação.

A partir do que já foi visto neste trabalho, as recomendações que acabaram de ser feitas não são uma grande novidade.

Surge então uma questão: o desempenho das redes neurais pode ser superior a outras técnicas de controle, como modelos matemáticos, estatística ou inteligência artificial?

Alguns autores consideram que, até o momento, as redes neurais são uma técnica geralmente comparável, mas não superior, aos melhores modelos matemáticos e estatísticos, para a resolução de problemas em geral. Sempre que possível, convém verificar, para uma dada aplicação, qual foi o desempenho obtido pelas redes neurais e pelos métodos convencionais [9].

Tal recomendação é válida, entre outros motivos, pelo fato de que as redes neurais são, na realidade, caixas pretas. Embora essa desvantagem não seja exclusiva desta técnica - afinal, muitas equações empíricas obtidas via regressão múltipla também o são - muitas vezes fica difícil verificar como uma rede neural chegou a um dado resultado. Contudo, o exame dos valores dos pesos e do grau da sensibilidade da resposta final à varia-ções na magnitude das variáveis de entrada podem fornecer alguma informação a esse respeito.

Inúmeros trabalhos tem apontado melhor previsão e facilidade do uso das redes neurais em relação à regressão linear e não-linear [1,8,10-13]. De fato, neste campo a aplicação de redes neurais parece ser desnecessária apenas quando a precisão fornecida por uma equação de regressão estatística já existente satisfazer as necessidades do usuário.

Isto posto, serão discutidas a seguir algumas regras básicas para o dimensionamento, treinamento e teste de redes neurais unidimensionais (ou seja, "feed-forward", tipo Rummelhart), treinadas por retro-propagação.

REDIMENSIONAMENTO DA REDE NEURAL

Ao se conceber uma rede neural, logicamente há dois parâmetros a serem definidos: o número de camadas de neurônios que ela deverá ter, bem como o número de neurônios em cada camada.

Uma rede neural unidimensional deve conter no mínimo duas camadas de neurônios. Isto é óbvio: uma camada se destina à entrada dos dados e a outra à saída dos resultados. Contudo, como já pode ser visto no tópico que tratou da evolução das redes neurais, perceptrons com apenas duas camadas apresentam utilidade muito limitada. De fato, eles só são úteis se os casos a serem modelados estão separados de forma linear.

A figura 3 mostra como o aumento do número de camadas de neurônios melhora o desempenho das redes neurais. Sua capacidade de aprendizado aumenta, o que se traduz na melhoria da precisão com que ela delimita as regiões de decisão [14]. Note-se nessa figura que redes neurais sem a camada oculta não conseguem modelar a operação lógica "ou exclusivo". Tal fato foi exaustivamente utilizado pelos detratores das redes neurais em meados da década de 1960, quando só se dispunha de redes neurais desse tipo. Note-se ainda que redes neurais com uma ou duas camadas ocultas, adequadamente treinadas, resolveram adequadamente esse problema.

Embora a necessidade de que haja pelo menos uma camada oculta na rede neural seja praticamente um ponto pacífico, há considerável controvérsia quanto ao número necessário de camadas ocultas necessárias para que o perceptron adquira o conhecimento de forma precisa.

Alguns autores, como HECHT-NIELSEN, basearam-se num famoso teorema de KOLMOGOROV, para afirmar que uma rede neural com apenas uma camada oculta pode calcular uma função arbitrária qualquer a partir de dados fornecidos. De fato, tem sido amplamente comprovado que redes neurais classificatórias efetivamente requerem apenas uma camada oculta.

Já outros pesquisadores, como CYBENKO e LIPPMAN, alegam que uma rede neural deve conter pelo menos duas camadas ocultas, o que parece ser particularmente válido quando a saída da rede consiste de valores que variam de forma contínua.

Uma vez definido o número de camadas da rede neural, o próximo passo a ser cumprido é a definição do número de neurônios por camada.

Na camada de entrada deve haver um número de neurônios igual ao número de variáveis a serem fornecidos à rede. Eventualmente uma variável de entrada pode ser subdividida em vários neurônios, segundo um esquema binário, o que eventualmente melhora o desempenho do perceptron. Normalmente se inclui à camada de entrada um neurônio com valor constante unitário, denominado "bias", que atua analogamente a um "terra" elétrico. Seu papel é aumentar o número de graus de liberdade disponíveis no modelo, permitindo que a rede neural tenha maior capacidade de se ajustar ao conhecimento à ela fornecido.

A camada de saída deve conter um número de neurônios igual ao de variáveis que se deseja calcular. No caso de modelos classificatórios, pode-se utilizar um neurônio para cada item de classificação ou utilizar uma representação mais compacta, empregando-se técnicas binárias para diminuir o número de neurônios.

Note-se contudo, que o uso de representação binária para se reduzir o número de neurônios nas camadas de entrada e saída aumenta a carga de trabalho da camada oculta. Tal fato obriga a um aumento do número de neurônios dessa camada ou mesmo a adição de uma camada oculta suplementar para que a rede neural mantenha o mesmo nível de desempenho.

O maior desafio no dimensionamento de uma rede neural é a escolha correta do número de neurônios das camadas ocultas. Na realidade, a camada oculta é uma camada generalizadora. Ela tende a combinar os neurônios das camadas de entrada e saída em metagrupos. Por exemplo, em aplicações que requeiram o reconhecimento de caracteres, a camada interna deve "aprender" a agrupar os "pixels" provenientes da camada de entrada em forma de linhas.

Foram sugeridos vários critérios para a escolha do número ótimo de neurônios das camadas ocultas:

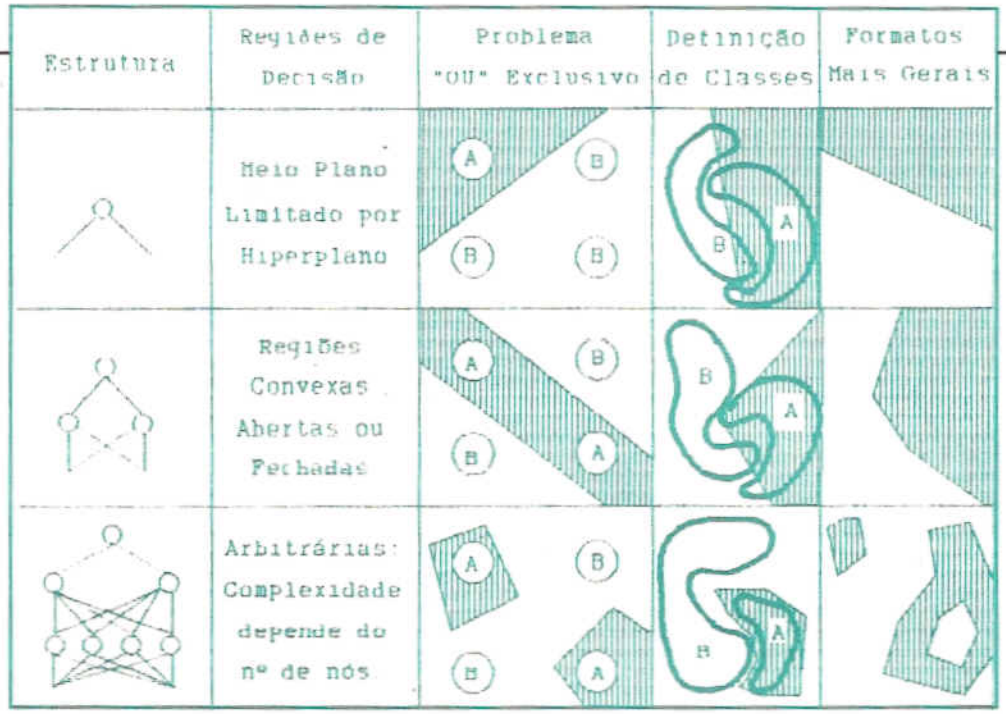


Figura 3: tipos de regiões de decisão que podem ser delimitadas por redes neurais unidirecionais (ou perceptrons), com nenhuma, uma e duas camadas ocultas. As regiões hachuradas indicam as regiões de decisão da classe A geradas pela respectiva rede neural. Regiões com contornos suaves indicam as distribuições reais para as classes A e B; essa informação foi fornecida às respectivas redes neurais durante sua fase de treinamento. Utilizou-se a função degrau como a função-ativação dos neurônios.

- **HECHT-NIELSEN/KOLMOGOROV:** uma rede neural com três camadas (entrada, oculta e saída) pode modelar qualquer função matemática contínua desde que a camada oculta contenha $2.i+1$ neurônios, onde i é o número de variáveis de entrada.
- **KUDRICKY:** verificou empiricamente que, numa rede neural com duas camadas ocultas, consegue-se um desempenho ótimo quando se obedece a uma taxa de 3:1 entre o número de neurônios da primeira e segunda camada ocultas. A regra parece ser válida mesmo no caso de um grande número de variáveis de entrada.
- **LIPPMANN:** no caso de redes neurais com duas camadas ocultas, a segunda camada deve conter o dobro do número de neurônios da camada de saída. Caso a rede neural contiver apenas uma camada oculta, ela deverá ter $s.(i+1)$ neurônios, onde i é o número de neurônios da camada de entrada e s é o número de neurônios da camada de saída;
- Outros autores preferem definir um número máximo de neurônios que a camada oculta deve conter:

$$O_{máx} = \frac{c}{10.(i+s)}$$

onde $O_{máx}$ é o número máximo de neurônios da camada oculta, c é o número de registros de dados utilizados na fase de treinamento da rede neural, i é o número de neurônios da camada de entrada e s é o número de neurônios na camada de saída.

De forma geral, para redes pequenas, onde o número de neurônios na camada de saída é maior do que o da entrada, a média geométrica entre o número de neurônios nas camadas de entrada e saída – ou seja, $\sqrt{(i.s)}$ – é uma boa estimativa para o número de neurônios da camada oculta. Por outro lado, quanto mais complexo for o relacionamento entre as variáveis de entrada e de saída, maior deverá ser o número de neurônios na camada oculta.

Embora o assunto ainda seja muito controverso, em linhas gerais, pode-se afirmar que há um número ótimo de camadas ocultas e de neurônios nelas contidas que, para cada caso, leva à maximização da capacidade preditiva da rede neural utilizada. O problema é determiná-los com base nas poucas diretrizes hoje existentes. Isso normalmente exige um número muito grande de tentativas, se o problema for relativamente complexo [1,3,9,14].

CONDICIONAMENTO DOS DADOS

Uma vez que a rede neural se baseia completamente nos dados a ela fornecidos para extrair o modelo geral que rege seu inter-relacionamento, é absolutamente vital que a informação a ela fornecida durante sua fase de treinamento seja rigorosamente fidedigna. Caso contrário, ela acabará por estabelecer um modelo espúrio, pois ela não dispõe de nenhuma regra preestabelecida para poder efetuar alguma crítica aos dados a ela fornecidos.

Ao menos até o momento, redes neurais somente podem processar dados numéricos. Isso não impede que outros tipos de variáveis possam ser modelados pelas redes neurais: basta apenas que eles sejam codificados previamente em forma numérica. Por exemplo: variáveis lógicas – Falso/Verdadeiro – podem ser traduzidas como 0 e 1, respectivamente. Variáveis de classes – por exemplo, solteiro, casado, divorciado... – como 0, 1, 2... Outra possibilidade para este tipo de variável é designar um neurônio binário para cada tipo de classes e atribuir a ele o valor unitário (ligado) a ela quando for o caso. Analogamente ao que ocorre no caso da regressão estatística, essa última variante evita que se atribua uma ordem de magnitude à cada classe, suprimindo efeitos indesejáveis. Eventualmente,

pode ser feito um programa auxiliar (em BASIC, por exemplo) para facilitar o processo de entrada e codificação de variáveis.

A princípio, a magnitude dos dados não deveria afetar o desempenho das redes neurais. Não é o que se verifica na prática. Contudo, tal fato poderia até ser esperado, uma vez que neurônios biológicos possuem uma faixa dinâmica bastante limitada, ou seja, processam níveis de sinais numa faixa de magnitude muito restrita. A solução para este problema é aplicar técnicas de compressão aos dados originais. Alguns autores recomendam a aplicação de logaritmo aos dados, se sua faixa de magnitude se estender além de algumas oitavas. Em casos especiais, pode-se utilizar outros recursos, como a aplicação de raiz quadrada ou a transformação de GABOR.

Há ainda outra limitação em relação à magnitude dos dados. Eles devem ser adequadamente escalados para poderem ser processados nas redes neurais, quer no estado bruto ou transformado. De fato, perceptrons que utilizam a função sigmoidal como função transferência trabalham numa faixa de valores entre 0 e +1, enquanto que os que operam com a função tangente hiperbólica estendem essa faixa para -1 a +1. Se os dados a serem modelados ultrapassarem a respectiva faixa, é necessário aplicar um fator de correção a eles para que sua magnitude se adeque às condições de sua rede neural.

Deve-se separar de 50 a 90% do número total de registros de dados disponíveis para se obter o sub-conjunto que será utilizado no treinamento da rede neural. O ideal é que a escolha dos dados para esse sub-conjunto se faça de forma totalmente aleatória, de modo a que ele realmente seja representativo em relação ao universo global que está sendo considerado. Um aspecto importante é apresentar à rede, durante seu treinamento, proporções iguais das diversas situações envolvidas, para que ela possa aprender como diferenciá-las de forma eficaz. Caso contrário, ela tenderá a "esquecer" as situações que aparecerem com menor freqüência.

O restante do conjunto de dados - 10 a 50% do total - forma o sub-conjunto de teste. Ele é apresentado à rede neural somente após seu treinamento. Uma vez que ele foi escolhido de forma aleatória, esses dados tem grande possibilidade de serem situações diferentes das apresentadas à rede durante seu treinamento. Logo, se efetivamente a rede neural conseguiu "deduzir" corretamente o inter-relacionamento entre os dados de entrada e saída, a diferença entre os valores calculados por ela e os respectivos dados reais de saída deverá ser mínima quanto ela processar o sub-conjunto de teste [1,3,9,13,14].

Dadas as necessidades de subdivisão do conjunto global de dados e da máxima generalização do treinamento da rede neural, o ideal é se dispor de no mínimo algumas centenas de dados relativos à aplicação que se quer modelar. Há registro do uso de redes neurais treinadas de forma aparentemente bem sucedida como apenas alguns dados. Contudo, nestes casos fica praticamente impossível se a rede neural realmente conseguiu generalizar a relação entre as variáveis ou se ela memorizou os dados.

TREINAMENTO DA REDE NEURAL

O treinamento das redes neurais unidirecionais envolve a otimização dos pesos das sinapses, ou seja, as energias correspondentes às ligações entre os neurônios, de modo a fazer com que ela mapeie corretamente a função proposta entre as variáveis de entrada e de saída.

O aprendizado deste tipo de rede se faz pelo método de retropropagação ("back-propagation"). O processo é iterativo. A cada iteração durante a fase de treinamento, é feita uma comparação entre os valores reais de saída do sub-conjunto de dados de treinamento com os correspondentes valores calculados pela rede neural a partir dos dados de entrada. Em função da magnitude das diferenças (erros) assim constatadas, o algoritmo de aprendizagem recalcula os pesos das sinapses de modo a minimizar os desvios.

Há inúmeras estratégias de aprendizado visando solucionar os diversos problemas apresentados nos primórdios do desenvolvimento das redes neurais. Foram (e eventualmente ainda são!) eles:

- Excessivo número de interações até ocorrer convergência, o que se traduz em tempo de treinamento demasiado longo (eventualmente dias);
- Propensão a travar o treinamento devido ao encontro de mínimos locais na superfície da função que descreve o erro;
- Segmentos planos nos extremos da função de ativação utilizada;
- Capacidade limitada de se lidar com parâmetros de entrada que não sejam invariantes do ponto de vista translacional, rotacional e de magnitude.

Hoje já estão disponíveis algoritmos que garantem rápido treinamento das redes neurais; alguns deles até foram patenteados.

Alguns autores sugerem que a taxa de treinamento da rede neural, definida por um **coeficiente de aprendizado** α, deve ser alta no início do treinamento e decline gradativamente a medida que ele evolui. Isto deve proporcionar rapidez na convergência do treinamento, estabilidade e resistência ao aparecimento de mínimos locais. STUBBS ainda sugeriu que o número de conjuntos de dados disponíveis para treinamento seja maior que cinco vezes o número total de neurônios da rede [1,3,9].

NOTA DA REDAÇÃO: Na próxima edição publicaremos a complementação desta matéria.

ANTONIO AUGUSTO GORNI é Engenheiro da Divisão de Pesquisas Tecnológicas da Companhia Siderúrgica Paulista e Professor-Assistente do Departamento de Metalurgia da Faculdade de Engenharia Industrial - FEI