UMA PROPOSTA PARA COMPARAÇÃO DE PERFORMANCE ENTRE REDES NEURAIS ARTIFICIAIS

José Pinto Ramalho [2] zeramalho@infosolda.com.br
Flávio Michel Ramilli Issa [1] flavio.issa@poli.usp.br
Mauro Conti Pereira [1,3] mauro@lac.usp.br
Sergio Duarte Brandi [2] sebrandi@usp.br
Cláudio Garcia [1] clgarcia@lac.usp.br
Fuad Kassab Junior [1] fuad@lac.usp.br
[1]Lab. de Automação e Controle, Depto. de Eng. de Telecomunicações e Controle,
Escola Politécnica da USP, Caixa Postal 61548 CEP 05424-970, São Paulo, SP, Brasil
[2]Depto de Eng. Metalúrgica e de Materiais, EPUSP, São Paulo ,SP, Brasil.
[3]Curso de Eng. de Computação – Universidade Católica Dom Bosco (UCDB)
Av.Tamandaré 6000 – CEP 79117-900 - Campo Grande, MS, Brasil.

Resumo Configurar os parâmetros de redes neurais artificiais (RNA) aproxima-se de uma arte, dependendo muito da experiência do projetista, encontrando-se dificuldades para compará-las. Este artigo descreve uma proposta de metodologia para comparação de desempenho de redes, indo além de um simples percentual de acerto, procurando-se definir escalas de intensidade do erro correspondente. Para duas redes com mesmo percentual de acerto, procuramos definir qual estava "mais errada". É proposta a definição de um índice qualitativo em que utilizamos conceitos semelhantes a matriz de confusão e distância de Hamming, ponderando-se também a dificuldade de ajuste e/ou conseqüência dos erros da rede em teste. É dado um exemplo de aplicação em reconhecimento de caracteres manuscritos, mas salientando a possibilidade de uso em outras áreas, tais como metrologia dimensional, entre outras.

Palavras Chaves: Redes neurais artificiais, reconhecimento de caracteres.

Abstract: To adjust the parameters of artificial neural networks (ANN) relies too much on the designer's experience, and one of the problems is to choose among several tested ANNs. This paper proposes a methodology to compare neural nets performances, comparing not only the correctness percentage but also defining a scale of corresponding error intensity. For two neural nets with the same correctness percentage we try to define which one has the worst errors that should be avoided. A new qualitative index is proposed, using concepts similar to confusion matrix and Hamming distance, but also weighting how difficulty is the adjustment of the neural net, and the error consequence for the net under test. It is given an example for handwritten character recognition, but suggesting the use in other areas, such as dimensional metrology.

Keywords: Artificial neural networks, character recognition.

1 INTRODUÇÃO

As redes neurais artificiais (RNA) oferecem boa abordagem para problemas que requeiram ajuste de funções, inclusive reconhecimento, identificação, associação ou classificação de padrões. Encontram-se aplicações em diversas areas tais como Administração, Química, Medicina, Controle de Qualidade, Linhas de Produção, Segurança e Telecomunicações.

O ajuste dos parâmetros de configuração de RNA permitem uma grande liberdade na concepção e implementação, o que dá liberdade ao projetista para que adeque a RNA para cada aplicação individualmente, mas que exige dele experiência como fator determinante para o sucesso da aplicação.

As múltiplas possibilidades na implementação e nos diversos tipos de pré-processamento aplicáveis aos dados, acabam por dar condições a que se desenvolvam RNAs adaptadas a resolução de problemas em maior ou menor intensidade. Isto justifica a comparação de desempenho entre diferentes RNA, ou entre a mesma RNA com diferentes modos de configuração para a resolução de um mesmo problema.

A própria comparação do desempenho entre RNAs também é complexa. Normalmente se utiliza uma taxa de acerto, mas isto não é suficiente quando temos de escolher entre duas redes com mesma taxa de acerto.

Faz-se necessária uma metodogia que diferencie dentre elas qual é a melhor, isto é, dos erros das duas qual foi mais dificil de ocorrer, ou ainda, qual erro tem consequências mais graves. Um identificador de placas de automóveis para emissão de multas tem a mesma consequência caso erre a leitura da placa, ou seja, daria uma multa para o veículo errado. Neste caso predomina a dificuldade de reconhecimento, por exemplo reconhecer um 3 como 8 é mais fácil que reconhecê-lo como um 7. Por outro lado, uma rede neural em aplicações médicas que gere um diagnóstico errado poderia ter consequências desastrosas. Este tipo de informacao deve pesar na escolha dos parâmetros da rede a ser usada.

A seguir citamos como alguns autores comparam RNAs, e partindo de um deles, propomos uma metodologia que leve em conta o que foi mencionado acima a respeito de quais erros são "piores" que outros. Foi usado um exemplo específico para

reconhecimento de algarismos manuscritos mas a mesma idéia pode ser usada para outras aplicações.

2 REDES NEURAIS ARTIFICIAIS (RNA)

Uma das estruturas de redes neurais mais comumente utilizadas é a *multilayer perceptron*, com o algoritmo de treinamento *bakcpropagation*. Conforme Haykin (2002), o *perceptron* é uma forma construtiva de RNA para padrões linearmente separáveis. Foi proposto em 1958 por Rosenblatt como sendo o primeiro modelo de aprendizagem supervisionada. É constituído de um único neurônio com um polarizador conhecido como *bias* e um peso sináptico ajustável. A limitação de um perceptron construído por um único neurônio é exatamente classificar os padrões em apenas duas hipóteses. A classificação em mais que dois planos cria a necessidade de mais neurônios, de onde vem o nome *multilayer perceptron* (MLP).

Não existe nenhuma metodologia clara para implementação de uma RNA. Conforme Diniz (1997), os principais parâmetros ficarão a cargo do conhecimento prévio do projetista. Entre estes parâmetros a autora cita as seguintes escolhas:

- número de camadas ("*layers*");
- número de neurônios por camada;
- topologia das interconexões;
- funções de transferência nas diferentes camadas;
- representação dos dados;
- dinâmica de treinamento, verificação e operação.

Todos estes fatores estão correlacionados. O aumento do número de camadas, por exemplo, implica em um aumento da complexidade e conseqüentemente do tempo de processamento. Um pequeno número de camadas pode fazer com que a RNA não tenha a flexibilidade necessária para a generalização; por outro lado, um número excessivo pode fazer com que a RNA "decore" os dados ao invés de generalizá-los.

2.1 Aspectos relativos à comparação entre Redes Neurais

Visando comparar a eficiência das diferentes redes, ou de uma mesma rede com diferente arquitetura construtiva, diversos autores têm proposto abordagens comparativas. A seguir é apresentado um resumo das abordagens de comparação das estudadas:

Dietterich (2000) mostra em seu trabalho uma metodologia comparativa onde são apresentados três tipos de dados para avaliação e treinamento da RNA. O autor separa os dados em:

- Dados Reais, que são os de experimentos práticos.
- Dados Realísticos, que são os dados artificiais com desvios controlados.
- Dados Sintéticos que são dados artificialmente desenvolvidos para ensinar a resposta "correta" para o sistema.

De posse destes diferentes conjuntos de dados, o autor propõe quando cada um dos conjuntos deve ser utilizado para as tarefas de aprimoramento de uma rede, ou comparação entre RNAs. Esta proposição é resumida na tabela 1.

Flexer (1996) apresenta em seu trabalho os requisitos mínimos necessários para uma avaliação comparativa entre redes. Primeiro separa os diferentes tópicos a serem analisados: como analisar a configuração da rede em si, com a verificação da influência da escolha de cada parâmetro, e como analisar estatisticamente os dados gerados por uma rede. Para esta segunda avaliação este autor propõe uma fórmula onde o desvio padrão dos resultados corretos é analisado.

Tabela 1 – *Synthetic versus Real World* – Dietterich (2000)

Operação a realizar	Tipo de dado recomendado
1 - Teste e verificação de funcionamento	Sintéticos
2 – Estimativa de desempenho em novos dados	Reais
3 – Estimativa de desempenho em tarefas similares	Realísticos
4 - Comparação com outros algoritmos:	
4.1 – Qual algoritmo é melhor em uma dada tarefa?	Reais
4.2 – Qual algoritmo é melhor para um conjunto de tarefas?	Reais ou Realísticos
4.3 – Verificar se 2 algoritmos distintos aprendem a mesma coisa.	Qualquer um
5 - Análise do comportamento do algoritmo:	
5.1 - Porque o algoritmo A tem melhor desempenho que B?	Qualquer um
5.2 - Em problemas o algoritmo A terá desempenho melhor que B?	Sintéticos

Prechelt (1994) apresenta em seu trabalho diferentes abordagens de testes de redes aplicadas em diferentes aplicações finais. O artigo comenta as particularidades da RNA aplicada às áreas médica, química e biológica, e ressalta os aspectos da escolha da função de ativação, a inicialização dos pesos sinápticos e o critério adotado para o encerramento do treinamento.

Dietterich (1997) em outro trabalho apresenta uma descrição entre cinco métodos para avaliação comparativa entre redes. Entre estes métodos aparece o teste de McNemar, onde é proposta a construção de uma tabela de duas linhas e duas colunas com os resultados corretos e incorretos identificados pelas duas redes na diagonal principal, e os identificados por uma das RNA e não pela outra na diagonal secundária. Esta abordagem permite explorar com clareza a questão do falso positivo, ou seja, algo falso que a rede identifique como verdadeiro.

3 METODOLOGIA PROPOSTA PARA A COMPARAÇÃO ENTRE AS REDES

Conforme já foi mencionado, a proposta deste trabalho é criar um indice que leve em conta não apenas os acertos, mas tambem os erros e uma ponderacao via sua gravidade (consequência dos erros) ou a dificuldade de ocorrerem.

Para facilitar a compreensao o método será descrito com o auxílio de um exemplo prático, o reconhecimento dos algarismos 0 a 9 manuscritos.

Assim, cada RNA testada originou uma matriz de confusão do memso tipo utilizado em comunicação de dados, de tamanho 10x10 com os dados a serem reconhecidos nas colunas, e o valor que a RNA reconheceu nas linhas. No cruzamento de cada linha e coluna é indicado o numero de vezes que aquele algarismo da coluna foi interpretado como sendo o algarismo da linha. Desta forma, na diagonal principal foi gerado o número de vezes que o algarismo foi lido corretamente, e nas linhas de cada coluna, a quantidade de dados reconhecidos

incorretamente conforme a freqüência com que foram identificados. Exemplificando: se oito algarismos 5 foram identificados uma vez como 6 e sete vezes como 5, a sétima linha da sexta coluna apresenta a quantidade um, enquanto a célula da sexta linha e sexta coluna apresenta a quantidade sete.

Com isto, gerou-se a informação relativa a cada algarismo e quantas vezes foi identificado corretamente. Nos casos em que não foi identificado, qual algarismo foi identificado em seu lugar e qual a freqüência desta ocorrência. Esta metodologia, proposta pelos autores, pode ser vista na tabela 2.

A partir disto, visando quantificar quais erros foram mais críticos, foi criado um índice de comparação. Por ex. identificar como 1 o algarismo 6 é mais crítico que identificar como 1 o algarismo 7. Para isto tomou-se como digito padrão o *display* de 7 segmentos, conforme mostrado na figura 3.

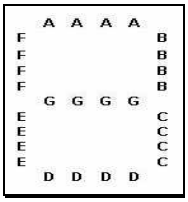

Figura 3 – *Display* de 7 segmentos

Tabela 2 – Matriz de identificação de dados

		Algarismo correto									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Algarismo reconhecido	**0**	**6**	3	5	0	2	6	0	7	3	8
	1	3	**10**	4	9	4	0	0	6	4	5
	2	7	2	**17**	0	9	6	3	5	6	0
	3	3	6	1	**24**	2	0	3	0	0	9
	4	3	2	0	3	**31**	9	8	4	7	0
	5	6	2	4	7	2	**6**	2	0	4	5
	6	3	8	9	7	8	3	**11**	5	0	6
	7	7	6	6	4	7	5	6	**18**	7	9
	8	2	4	0	6	0	6	2	0	**25**	6
	9	7	5	5	0	1	9	0	2	4	**32**

De acordo com este desenho, e atribuindo 0 e 1 para cada segmento utilizado, os números foram interpretados da seguinte maneira:

Tabela 3 – segmentos ativos para cada algarismo

		Segmentos utilizados						
		A	**B**	**C**	**D**	**E**	**F**	**G**
Algarismo	**1**	0	1	1	0	0	0	0
	2	1	1	0	1	1	0	1
	3	1	1	1	1	0	0	1
	4	0	1	1	0	0	1	1
	5	1	0	1	1	0	1	1
	6	1	0	1	1	1	1	1
	7	1	1	1	0	0	0	0
	8	1	1	1	1	1	1	1
	9	1	1	1	1	0	1	1
	0	1	1	1	1	1	1	0

Foi calculada então a distância de Hamming, a quantidade de *bits* que diferem entre as duas seqüências de 1 e 0. Realizando uma operação XOR, obtém-se resposta 0 para os segmentos coincidentes. A soma desta operação propiciou então o grau de diferença entre os números estudados.

Exemplificando, a comparação entre os algarismos 1 e 2 seria:

1 = 0110000
2 = 1101101
XOR = 1011101 soma =5

Desta forma, foram então contadas as diferenças de segmentos não coincidentes para todos os números, e estas foram classificadas desde a diferença de apenas um segmento (caso do algarismo 1 ser lido como 7 por ex.) até a diferença de seis segmentos (caso do algarismo 6 ser interpretado como 1). Com estes valores, foi montada uma tabela que classifica os erros de interpretação da RNA de acordo com a "gravidade" do erro. Isto é apresentado na tabela 4.

De acordo com a quantidade de segmentos não coincidentes, os pares de algarismo original e como este foi reconhecido, foram divididos em 3 grupos, listados abaixo com o % de ocorrência sobre a quantidade total de erros:

- GRAU A - Erros leves: pares dos grupos 1 e 2 (40% dos casos)
- GRAU B - Erros intermediários: pares dos grupos 3 e 4 (46,67% dos casos)
- GRAU C - Erros grosseiros: pares classificados nos grupos 5 e 6 (13,33% dos casos)

Tabela 4 – Classificação de sobreposição dos segmentos dos algarismos, com fonte igual ao display de 7 segmentos

Pares de números	Quantidade e classes de erros											
	A				B				C			
	1		2		3		4		5		6	
	5 6		2 3		3 4		7 8		1 2		1 6	
	8 9		9 0		4 5		4 6		6 7			
	6 8		3 5		1 3		5 7		2 4			
	8 0		1 4		7 9		2 5		1 5			
	5 9		5 8		3 6		2 7		1 8			
	1 7		6 9		4 7		4 0					
	3 9		3 7		7 0		1 9					
			6 0		2 6		1 0					
			3 8		4 8							
			4 9		5 0							
			2 8		2 9							
					3 0							
					2 0							

Foram ponderados em seguida estas porcentagens de erros por grupo com pesos para cada grupo. Os valores são atribuídos visando aplicar um peso maior a um erro classe "C", mais difícil de ocorrer que um erro classe "A", neste exemplo. Lembramos que em outra aplicação as classes podem representar também a gravidade da conseqüência de um erro. Assim, o modelo proposto é flexível a outras distribuições de peso, de acordo com a classificação do grau dos erros (gravidade da conseqüência ou sua dificuldade em ocorrer).

À soma destes erros ponderados foi denominada IEP (Índice de Erro Ponderado). Este conjunto de pesos pode ser alterado de acordo com a importância da aplicação RNA. Em um caso, por exemplo, em que o erro grosseiro possa ser catastrófico, o valor atribuído ao grau C aumentaria em detrimento dos demais.

Para o caso da comparação de RNA para reconhecimento de caracteres, foram definidos os pesos com valores 1, 4 e 7, para que a variação entre a classe A e B e a variação entre B e C fosse a mesma.

Uma vez calculado o IEP, este foi subtraído da taxa de acerto (TA), usando a própria taxa de acerto e seu complementar como pesos.

Peso aplicado na taxa de acerto (Pa) = TA / 100
Peso aplicado na taxa de erro (Pe) = (1 - Pa)

Com isto, foi criado um índice que leve em conta o acerto e o erro, ponderados pelos seus respectivos pesos, que resulta num valor entre –1 e 1. Dividindo-se por 2 e soma-se 0.5 para resultar num índice mais intuitivo, o índice é transformado então em percentagem para normalizar o valor entre 0 e 100, que denominamos de Índice de Avaliação da Rede, IAR:

$$IAR = \left[\left[\frac{(TA \times Pa) - (IEP \times Pe)}{2}\right] + 0{,}5\right] \times 100$$

Desta forma é possível comparar a performance de duas RNA, mesmo quando estas tenham índice de acertos iguais, uma vez que podem ser diferenciadas pelo IEP e apresentarão conseqüentemente IAR diferente.

Um exemplo numérico hipotético que detalha a metodologia apresentada é aqui apresentado na tabela 5, onde duas redes hipotéticas reconheceram corretamente 70% dos caracteres. Os 30% classificados errados foram distribuídos como apresentado na tabela 5.

*IEP rede 1 = [(0,33*1)+(0,33*4)+(0,33*7)]/12 = 33%*
*IAR rede 1 = [[((0,7*70)-(0,3*33,3))/2]+0,5]*100 = 69,49*

*IEP rede 2 = [(0,16*1)+(0,16*4)+(0,66*7)]/12 = 45%*
*IAR rede 2 = [[((0,7*70)-(0,3*45))/2]+0,5]*100 = 67,62*

Tabela 5 – Exemplo de cálculo do IAR

		Rede 1	Rede 2
Total de amostras		1000	1000
Classificados corretamente		70%	70%
Quantidade de erros	Grau A	100	50
	Grau B	100	50
	Grau C	100	200
% de erros sobre total de erros	Grau A	33,33%	16,66%
	Grau B	33,33%	16,66%
	Grau C	33,33%	66,67%
IEP		33,33	45
IAR		69,49	67,62

4 ESTUDO DE CASO

Neste trabalho foi utilizado o software Matlab 6.1 para a implementação das RNAs, utilizando um *toolbox* específico, o que faz com que as funções de projeto e treinamento da RNA sejam simplificadas.

Foi desenvolvido um conjunto de dados de treinamento através da seguinte metodologia: obtenção na *internet* de sete fontes manuscritas desenvolvidas para o ambiente *Windows* e com elas foi feita a geração de sete conjuntos com 10 números cada. Este conjunto foi denominado "conjunto de dados sintéticos".

Foi obtido na internet, da página de Le Cun (2002), um conjunto de 10000 dados contendo números manuscritos. Estes dados, que foram denominados dados reais, estão no formato 28x28 *pixels*. Com estes dados foi feito um pré-processamento na imagem reduzindo-a para a resolução 14x14 *pixels*, substituindo cada quatro pontos (2x2) da imagem original por sua média aritmética, e depois binarizados. Este tratamento foi aplicado tanto ao conjunto de Le Cun (2002) quanto ao conjunto de dados sintéticos produzido.

Desta forma, foram quatro os conjuntos de dados utilizados que são identificados com as seguintes nomenclaturas de dados:

- Reais – 1000 dados dos 10000 obtidos na internet
- Reais Reduzidos – reais com tratamento de redução
- Sintéticos – 70 dados obtidos de fontes para Windows
- Sintéticos Reduzidos – sintéticos com tratamento de redução

Foi utilizada uma RNA tipo MLP com *backpropagation*. A inicialização foi feita com os parâmetros taxa de aprendizagem e taxa de "*momentum*" sugeridos no artigo de Silva e Oliveira (2001). Os pesos foram inicializados com o valor 0,01 e para a polarização "*bias*" foi escolhido o valor 0,1.

Para cada um dos quatro conjuntos de dados para treinamento foi então desenvolvida uma RNA, com diversos valores em seus parâmetros de configuração, de modo a identificar aqueles que apresentassem melhor desempenho no reconhecimento dos algarismos. Resultou num total de 48 configurações para cada um dos quatro grupos de dados de treinamento. As variações foram: taxa de aprendizagem, momentum e número de neurônios na camada escondida. Foram escolhidas as taxas de erro: 0,1 0,05 0,02 e 0,02, respectivamente para as quantidades de neurônios 10, 20, 30 e 89 (ou 44). A quantidade de neurônios na camada escondida teve como estimativa inicial a média geométrica entre o número de dados de entrada e de saída. Por esta razão os valores são: 44 para os dados reduzidos (14x14 *pixels*, 196 entradas e 10 saídas) e 89 (para 28x28 *pixels*, 784 entradas e 10 saídas) para os dados originais.

Os testes de validação foram realizados utilizando-se 8 conjuntos de 1000 caracteres cada. Para cada conjunto de teste foi calculado o IAR. Em seguida, foram calculadas a média e o desvio padrão do IAR para todas as RNA de mesma configuração, e mostrados em 4 gráficos para uma visualização mais facilitada.

Isto resultou em um conjunto de dados com desvio padrão e média de cada uma das configurações, sendo que a melhor configuração entre as RNA geradas foi considerada a que apresentou o maior limite inferior para o valor de Média menos o Desvio Padrão, ou seja, a RNA que garante a melhor performance mínima.

5 DISCUSSÃO DOS RESULTADOS

Para cada conjunto de dados usado para treinamento, escolheu-se a RNA com melhor performance mínima garantida, resumindo-as na tabela 7. Nos gráficos das figuras 4 a 7, que mostram para cada conjunto o valor médio do IAR menos o desvio padrão, pode-se observar visualmente que escolhe-se o traço mais curto e com seu ponto mínimo mais alto. Assim, vale destacar que escolhe-se uma rede com menor taxa de acerto do que outra, mas que terá erros menos graves, ou mais aceitáveis.

Para os ajustes dos parâmetros de configuração das RNA, seguiu-se a metodologia de utilizar três diferentes taxas de aprendizado, quatro diferentes constantes de momento e quatro diferentes números de neurônios na camada escondida, totalizando 48 diferentes configurações para cada conjunto de dados de treinamento.

Tabela 7 – Resultados das melhores RNA

Conjunto de Treinamento	RNA escolhida	Taxa% de acerto média	IAR - DP	Neurônio na camada escondida	Cte de Momentum	Taxa de aprendizagem
1000 reais originais	28	76,75	74,64	30	0,5	0,3
1000 reais reduzidos	48	83,48	80,46	44	0,9	0,9
70 sintéticos originais	41	15,91	41,50	89	0,5	0,6
70 sintéticos reduzidos	25	15,90	40,60	30	0,3	0,3

Esta decisão foi em razão das diferentes recomendações encontradas nas literaturas consultadas sobre este tópico. Com isto, avaliou-se um maior universo de configurações para cada RNA, do qual foi selecionada, com o IAR, a configuração de melhor resultado. Como já mencionado, eventualmente pode-se escolher uma rede com menor taxa de acerto do que outra, mas que terá erros menos graves, ou mais aceitáveis.

O conjunto de dados sintéticos tinha por objetivo ser aplicado na metodologia de avaliação proposta por Dietterich (2000). Porém, as taxas de acerto obtidas com esses dados mostraram que os mesmos não eram suficientemente representativos do conjunto de dados reais. Visando aumentar esta representatividade, estes dados foram replicados 10 vezes e empregados no treinamento de uma RNA. O que se observou disto é que a maior quantidade de dados, com a mesma variedade, não melhora o desempenho da RNA.

O pré-processamento visou inicialmente diminuir o número de neurônios de entrada de 784 (28x28) para 196 (14x14). Com isto, visava-se reduzir proporcionalmente o tempo de treinamento. O objetivo era comparar com o tempo de treinamento da RNA com os dados originais, e verificar comparativamente os resultados para determinar se o pré-processamento pode ser aplicado sem prejuízo da performance da RNA. Notou-se que há uma redução considerável no tempo de treinamento sem afetar a performance da RNA.

Era também objetivo desta redução, verificar como a diminuição das possibilidades de variação dos *pixels* influenciaria os resultados da RNA. O que se observou foi a melhora no desempenho da RNA, provavelmente devido à diminuição da possibilidade de variação dos *pixels* que constituem a imagem.

Foi também realizado o teste conforme Flexer (1996), dividindo o conjunto de dados em 8 conjuntos de 1000 dados cada. Assim, todas as RNA foram testadas oito vezes e os resultados comparados em termos de taxa de acerto e IAR. Calculou-se a média e o Desvio Padrão destes valores, com o que foi possível determinar a amplitude da performance de cada RNA.

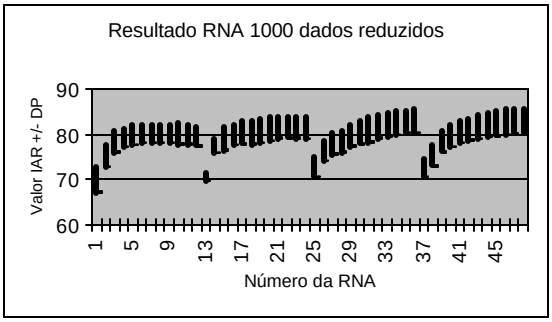

Figura 4 – Resultados da RNA treinada com 1000 dados reduzidos

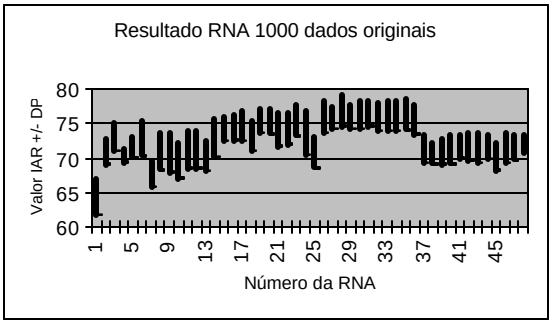

Figura 5 – Resultados da RNA treinada com 1000 dados originais

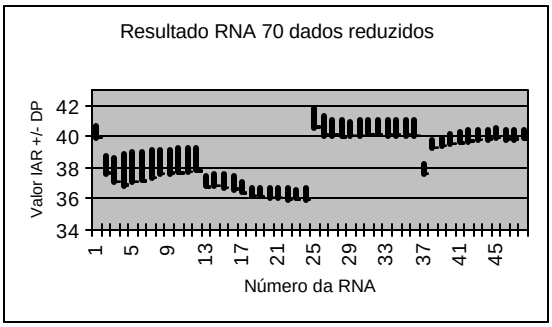

Figura 6 – Resultados da RNA treinada com 70 dados reduzidos

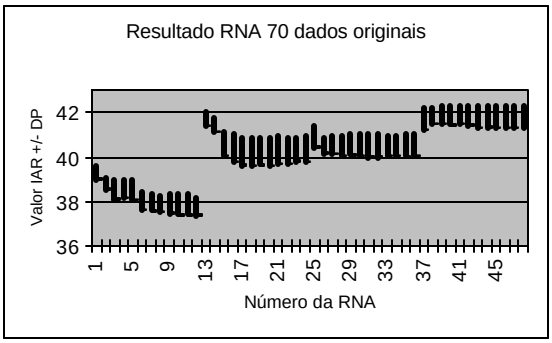

Figura 7 – Resultados da RNA treinada com 70 dados originais

Para exemplo utilizado de reconhecimento de algarismos, foi escolhido como critério para seleção a RNA com a maior performance mínima. Mas em outros casos pode basear-se em um dos critérios apresentados a seguir:

- Maior performance máxima,
- Maior performance média,
- Maior performance mínima ou
- Menor desvio padrão.

O IAR foi desenvolvido tomando por base a metodologia de teste de McNemar citada no trabalho de Dietterich (2000), porém difere do enfoque deste por julgar a importância do erro de identificação cometido pela RNA.

O teste de McNemar classifica entre duas RNA os acertos e erros de ambas e os erros individuais, dando importância ao "falso positivo", quando uma RNA classifica algo falso como sendo verdadeiro.

A proposta do IAR é avaliar a intensidade do erro para cada RNA, considerando, por exemplo, o reconhecimento do algarismo 1 como 7 algo menos crítico que reconhecer o algarismo 1 como 8. Esta separação entre diferentes tipos de erros foi dividida em classes e devidamente ponderada.

É importante destacar que o desempenho do IAR depende dos valores de pesos adotados nas diferentes classes de erros. Exemplificando, na avaliação de uma RNA de diagnóstico de paciente, os erros ao classificar um paciente doente como são teriam peso maior, enquanto o inverso disto teria peso menor. Ao atribuir importância aos diferentes tipos de erros, o IAR possibilita uma melhor avaliação comparativa de diferentes RNA, não baseando esta decisão apenas na taxa de acerto, uma vez que neste caso, os erros poderiam vir a ser catastróficos.

6 CONCLUSÕES

O índice IAR mostrou ter aspectos não destacados nos outros testes citados nos trabalhos de análise de performance de RNA estudados, devido a:

- Não comparar RNA com base apenas em sua taxa de acerto.
- Realizar analise qualitativa e ponderada dos erros cometidos
- A possibilidade de atribuição de diferentes pesos às diferentes classes de erro, permite a utilização do índice IAR na comparação de RNA onde a tolerância aos erros varia de inaceitáveis a toleráveis.
- Procuramos selecionar a RNA com melhor desempenho mínimo garantido, escolhendo-se aquela com maior valor de [Média $_{(IAR)}$ – DP $_{(IAR)}$].

Na aplicação de reconhecimento de caracteres manuscritos usada como exemplo pode-se observar que:

- Dentre as configurações testadas, o melhor desempenho no reconhecimento de caracteres foi obtido com 44 neurônios na camada escondida, 0,9 de taxa de aprendizagem, 0,9 de constante de momentum, pré-processamento de redução de imagem e treinada com dados originais.

- Nos dados artificialmente gerados o pré-processamento apresenta uma diminuição do IAR, uma vez que estes dados possuem originalmente menor variação de tonalidade e forma que os dados reais.

- O pré-processamento de redução de 28x28 para 14x14 *pixels* melhorou o desempenho da RNA treinada com dados reais, devido a diminuição das possibilidades de variação dos *pixels* que constituem a imagem.

- A RNA treinada com os dados reais apresentou melhor taxa de acerto e IAR do que a RNA treinada com dados sintéticos. Deste fato interpretou-se que:
  - A RNA com dados reais apresenta maior capacidade de generalização para reconhecimento de caracteres
  - Os dados sintéticos não possuem suficiente representatividade para reconhecimento de dados reais.

.

Vale ainda destacar que o índice e metodologia propostos podem ser adaptados a qualquer outra aplicação aonde seja possível ou desejável avaliar os erros. Pode-se escolher a quantidade de níveis de classificação de erros e os pesos usados para ponderá-los de acordo com o quanto se deseja evita-los. Isto já foi aplicado também na área de metrologia dimensional, em comparação de qualidade de solda.

REFERÊNCIA BIBLIOGRÁFICA

Dietterich, Thomas G. – Experimental Methodology for Benchmarking – Department of Computer Science, Oregon State University – USA 2000 http://wwwipd.ira.uka.de/~prechelt/nipsbench/expmeth.ps.gz

Dietterich, Thomas – Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. - Neural Computation, Department of Computer Science, Oregon State University – USA 1997 ftp: //ftp.cs.orst.edu/pub/tgd/papers/nc-stats.ps.gz

Diniz, Suelaine dos Santos - Uso de Técnicas Neurais para o Reconhecimento de Comandos à Voz - Dissertação de Mestrado em Engenharia Elétrica - Rio de Janeiro, Instituto Militar de Engenharia, Julho/1997.

Flexer, Arthur – Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice – The Austrian Research Institute For Artificial Intelligence – Vienna, Austria 1996 ftp://ftp.ai.univie.ac.at/papers/oefai-tr-95-16.ps.gz

Haykin, Simon – Redes Neurais, princípios e prática. Editora Bookman, Porto Alegre RS 2002.

Le Cun, Yann - Página pessoal http://yann.lecun.com/exdb/mnist/index.html

Prechelt, Lutz – Proben 1 – A set of Neural Network benchmark problems and benchmark rules – Fakultat für Informatik von Karlsuhe – Alemanha 1994 http://www.poli.usp.br/d/pmr5406/Download/Metodologia/proben.ps

Silva, Eugênio e Oliveira, Anderson Canêdo - Dicas para a configuração de Redes Neurais - Rio de Janeiro - RJ, Outubro 2001. http://www.nce.ufrj.br/labic/downloads/dicas_cfg_rna.pdf