Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 295–300, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Replicação *bootstrap* e Análise de sensibilidade em Redes Neurais Artificiais

Renan M.V.R.de Almeida[1]; Antonio Fernando C.Infantosi[2]; Ronaldo C.Gismondi[3]

[1,2]Programa de Engenharia Biomédica - COPPE/UFRJ Caixa Postal 68510 - Rio de Janeiro - RJ - CEP 21945-970
[3]Faculdade de Ciências Médicas - Universidade do Estado do Rio de Janeiro - UERJ
E-mails: renan@peb.ufrj.br, afci@peb.ufrj.br, gismondi@uerj.br

Abstract

The relative importance of the input variables in an Artificial Neural Network (ANN) model was investigated by means of bootstrap replications and a sensitivity analysis. The ANN was developed from 43 variables of the social, economic, environmental and health sector of 59 Brazilian municipalities, using infant mortality as the dependent variable. Eight variables, chosen with the help of a Factor Analysis on the data, were used as inputs. The relative importance of the inputs was investigated with the help of bootstrap replications of the model. The determination coefficient R^2 and the Mean Square Error (MSE) were obtained from bootstrap samples for the entire model, and then compared to those obtained after withdrawing each input variable at a time. For the full model, it was found R^2=.80 and $MSE = (7)(10^3)$. The sensitivity analysis identified as the most important variables *literacy*, *agricultural sector jobs* and *number of commercial establishments.* The method seems a simple alternative for evaluating input importance in ANN models.

1. Introdução

Redes Neurais Artificiais (RNA) representam uma alternativa recente aos métodos clássicos de modelagem, como regressões logística, linear múltipla (RLM) e análise de discriminantes [1,2]. Em saúde, diferentes estudos têm avaliado seu potencial de utilização, mostrando que a RNA pode apresentar desempenho preditivo melhor que os métodos estatísticos convencionais.

No entanto, modelos estatísticos visam usualmente realizar inferências causais. Neste aspecto, modelos de regressão, por exemplo logística, seriam mais adequados que a RNA, pois permitiriam identificar facilmente quais variáveis apresentariam maior valor de determinação em relação à saída, baseando-se na magnitude dos coeficientes β e em seus *odds ratios* associados [1]. Por outro lado, para RNA apenas recentemente técnicas com este objetivo têm sido propostas [3]. A despeito da crescente utilização de aplicações baseadas em inteligência artificial e RNA nas áreas da ciência computacional e do diagnóstico médico, tais recursos encontram-se ainda relativamente inexplorados no campo da epidemiologia [4].

O coeficiente de mortalidade infantil – número de óbitos de menores de um ano para cada mil nascidos vivos - é reconhecido como um dos mais expressivos e sensíveis indicadores das condições de vida e desenvolvimento de uma comunidade. A modelagem da mortalidade infantil tem sido realizada por diversos autores, empregando, na maioria das vezes, regressão linear múltipla (RLM), com o objetivo principal de investigar os determinantes de óbitos de crianças menores que um ano, em diferentes contextos.

O objetivo do presente trabalho é utilizar a simulação *bootstrap* e a análise de sensibilidade na avaliação da importância relativa das variáveis de entrada em uma rede neural desenvolvida para a modelagem da mortalidade infantil em municípios brasileiros.

2. Métodos

2.1 Estimativa *bootstrap* do erro de predição

Uma limitação aos métodos de modelagem estatística em saúde decorre do freqüentemente pequeno tamanho das amostras disponíveis e dos pressupostos necessários ao seu uso. A análise estatística baseada em simulações pode fornecer, como no caso da metodologia *bootstrap*, ferramentas adicionais, que auxiliem a contornar esses problemas.

A técnica do *bootstrap* pode ser aplicada à avaliação de diferentes estimativas estatísticas, como erros-padrão, *bias*, resíduos, intervalos de confiança, testes de hipóteses, dentre outros, em contextos diversos [5,6]. O marco fundamental na sua difusão é o trabalho de Efron [5], que propiciou um reestudo e síntese de antigas idéias sobre reamostragem, e, ao mesmo tempo, estabeleceu novos paradigmas para a análise estatística baseada em simulações.

A idéia-chave do *bootstrap* consiste em promover a reamostragem dos dados originais – direta/aleatória ou através de um modelo – para criar conjuntos de dados replicados, a partir dos quais a variabilidade das estatísticas de interesse possa ser avaliada.

Uma amostra *bootstrap*, $x^* = (x_1^*, x_2^*, ..., x_n^*)$, é obtida por amostragem aleatória *n* vezes, com reposição, a partir de *n* observações, $x_1, x_2, ..., x_n$, de um conjunto de dados. Por exemplo, seja *n* = 8, poder-se-ía obter uma das *B* amostras *bootstrap*, como:

$$x^1 = (x_5, x_3, x_7, x_3, x_1, x_4, x_4, x_2)$$

A partir deste princípio, torna-se possível a análise da variabilidade de estatísticas em um modelo. Por exemplo, uma boa medida da capacidade preditiva de um modelo é representada pelo *Erro Quadrático Médio EQM*:

$$EQM = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}, \qquad (1)$$

onde y_i é o valor observado da variável dependente *y* para o caso *i*, $\hat{y}_i$ o valor estimado e *n* o número de casos disponíveis. O EQM de um modelo pode ser obtido através da replicação *bootstrap*, tomando-se *B* amostras *bootstrap* dos dados originais, ajustando-se um modelo para cada amostra, aplicando-se a cada modelo os dados originais, e, então, obtendo-se *B* estimativas de EQM, a partir das quais obtém-se o *Erro Quadrático Médio de Predição Bootstrap, EQM_B*, como sua média [5].

Por outro lado, o Coeficiente de Determinação, R^2, pode ser definido como:

$R^2 = 1 - \frac{S_{(\hat{y}-y)}}{S_{yy}}$, onde $S_{(\hat{y}-y)}$ representa a variância do erro de estimação, enquanto S_{yy} corresponde à variância da variável dependente observada.

Similarmente, uma estimativa *bootstrap* do coeficiente de determinação de um modelo (percentual de variância explicada) pode ser também obtida como a média dos R^2 das B amostras *bootstrap*.

2.2. Dados, subconjuntos e modelagem

A importância relativa das variáveis de entrada foi estudada em uma modelagem RNA da mortalidade infantil a partir de seus determinantes demográficos, sócio-econômicos e ambientais em municípios do país. Os dados utilizados representam valores médios de 43 variáveis (sócio-econômicas, ambientais e de oferta de serviços de saúde, Tabela 1) tomados sobre o período 1978-82, referentes a 59 municípios brasileiros com mais de 100.000 habitantes. Estas variáveis representam as possíveis entradas da rede, enquanto a taxa de mortalidade infantil nos municípios selecionados foi escolhida como variável de saída.

A RNA foi desenvolvida orientando-se a escolha dos preditores a partir de uma Análise Fatorial (AF)[7,8] prévia, interpretada com base na experiência profissional e na literatura referente à mortalidade infantil. Para tal, as variáveis de entrada foram subdivididas em dois grupos, um relativo aos setores educacional, econômico, e de condições de moradia e meio ambiente (21 variáveis), e outro aos setores social, de produção e oferta de serviços de saúde (22 variáveis). Na AF do primeiro grupo, identificaram-se três fatores comuns: *menor escolaridade* e *renda* (31,55% da variância total), *maior renda* (26,68%) e *maior escolaridade* (18,94%), totalizando aproximadamente 77% da variância. No segundo grupo foi possível explicar 75,85% da variância, com fatores interpretados como representando a *estrutura agrária* dos municípios analisados (24,11% da variância), *oferta de emprego e serviços públicos de saúde* (15,50%), *oferta de serviços particulares de saúde* (14,80%), a *prestação de serviços* (11,37%), e a *seguridade social e estrutura industrial* (10,07%) desses municípios. A partir dessa análise, uma variável por fator foi selecionada como preditora, com auxílio da vasta literatura na área [6,9,10,11,12,13,14]. Assim, correspondendo aos fatores relacionados, foram selecionadas como entradas da RNA: número de pessoas ocupadas em estabelecimentos agropecuários - *agrária*; número de estabelecimentos comerciais - *estcom;* número de leitos em estabelecimentos particulares - *leitos;* número de pessoas ocupadas na prestação de serviços - *serviços;* número de pessoas ocupadas em estabelecimentos industriais - *indústria;* número de pessoas alfabetizadas - *alfabet;* número de aparelhos telefônicos - *telefone;* e número de pessoas com 8 ou mais anos de estudos – *escolaridade*.

No estudo foram utilizados os programas *Statistica* 5.1B (para implementar a Análise Fatorial), *S-plus 4.0* (gerar as amostras *bootstrap* e realizar as regressões lineares), e *Statistica Neural Networks 3.0* (simulações com redes neurais). A arquitetura da RNA constitui-se, genericamente, de oito entradas correspondentes às variáveis preditoras da mortalidade infantil, uma camada oculta com quatro neurônios com função de transferência logsigmóide e uma camada de saída com um neurônio também logsigmóide (arquitetura: 8x4x1). O número de neurônios na camada oculta foi estabelecido de modo a que a razão entre o número de variáveis de entrada e o número de conexões (neurônios na camada oculta) estivesse entre 1,8 e 2,2, minimizando, assim, a possibilidade de super-especialização da rede [15]. Para facilitar o processo de convergência da rede, todas as variáveis foram normalizadas para a faixa 0-1 [1,4]. O algoritmo de treinamento utilizado foi o *Conjugate Gradient Descent*.

Também com vistas a reduzir as chances de super-especialização da rede, procedeu-se à interrupção do treinamento após um determinado número de ciclos [1,4]. Para essa estimativa, os dados relativos às *k* variáveis de entrada de cerca de 80% dos municípios, selecionados por sorteio [15], foram utilizados para treinamento da RNA, ficando os demais municípios para a validação cruzada. Adotou-se como critério de

interrupção do treinamento o número de ciclos em que o EQM da amostra de validação apresentou estabilidade ou tendência a aumentar por cinco ciclos consecutivos [1], tendo resultado em 80 ciclos.

A seguir, foram geradas, a partir do conjunto original de dados, B=50 amostras *bootstrap*. Procedeu-se ao treinamento com cada uma dessas amostras, obtendo-se estimativas da mortalidade infantil, dos *EQM* e coeficientes de determinação (R^2) tomando-se os dados originais como entrada desta nova RNA. Finalmente, obteve-se EQM_B e R_B^2.

Oito experimentos foram então realizados, mantendo-se a arquitetura original da rede (número de camadas e de neurônios), porém tendo a variável "de teste", cuja importância para o modelo desejava-se estimar, sido substituída, em cada amostra *bootstrap*, pelo seu valor esperado para os 59 municípios. A importância relativa desta variável foi, então, avaliada como a diferença entre a média dos R_B^2 assim obtidos e o R_B^2 referente à replicação *bootstrap* com todas as variáveis originais do ensaio.

3. Resultados

Com a utilização das oito variáveis de entrada (determinadas através da AF) a RNA original foi capaz de prever aproximadamente 80% da variância total da mortalidade infantil. A avaliação *bootstrap* desta RNA resultou em coeficiente de determinação $R_B^2 = 0{,}65$ e $EQM_B = 12{,}97 \times 10^3$. Substituindo-se cada uma das oito variáveis de entrada pelo seu valor esperado (Tabela 2), nota-se que a intervenção na variável *alfabet*, relativa ao número de pessoas alfabetizadas, produz a maior redução (quase 21%) em R_B^2 e uma elevação de cerca de 40% em EQM. Ao se intervir na variável *estcom* e *agrária*, a redução percentual do coeficiente de determinação *bootstrap* é superior a 17%, enquanto que para as variáveis *escolaridade*, *leitos*, *telefone*, *indústria* ou *serviços*, o impacto é inferior a cerca de 8%.

4. Discussão

Devido ao grande número de variáveis (43) a ser considerado na camada de neurônios de entrada da RNA e a possível existência de colinearidade, a Análise Fatorial foi aplicada com vistas à redução da dimensionalidade. Desta resultou a identificação de oito fatores comuns, a partir dos quais foram selecionadas as oito variáveis preditoras para a modelagem RNA, sendo duas delas relativas a níveis educacionais, *número de pessoas alfabetizadas* e *número de pessoas com oito ou mais anos de estudo*. Essas variáveis posicionaram-se na AF em fatores diferentes, representando, portanto, conceitos distintos, a primeira podendo ser interpretada como um indicador de condições de vida das populações mais pobres, e, a segunda, de parcelas mais favorecidas. A RNA obtida para a mortalidade infantil mostrou melhor desempenho que modelos lineares desenvolvidos a partir do mesmo conjunto de dados [16].

A aplicação da técnica *bootstrap* permitiu avaliar a capacidade preditiva da RNA quando submetida a outros conjuntos de amostras, porém estatisticamente oriundos da mesma população. O método é uma alternativa útil para a situação na qual não se dispõe de uma base de dados em número suficiente para a validação do modelo. Tal aspecto é de particular importância em epidemiologia, onde, em geral, não se tem como realizar replicação do experimento, sendo este o caso do presente estudo.

Além disto, a técnica *bootstrap* foi também empregada para se investigar a importância relativa de cada uma das oito entradas da RNA. Assim, as variáveis *pessoas alfabetizadas* (*alfabet*) e *pessoas ocupadas nos estabelecimentos comerciais (estcom) e agropecuários* (*agrária)* mostraram-se como as de maior impacto na modelagem da mortalidade infantil (redução do coeficiente de determinação superior a cerca de 17%). Estes achados confirmam aspectos bem estabelecidos, em especial para *alfabet* e *agrária*. Para a primeira, sabe-se que alfabetização elevada tem enorme influência nos aspectos preventivos e curativos em saúde, levando a uma melhor compreensão da etiologia da doença, à capacidade de evitá-la e oportunidade de tratá-la. A segunda, relacionada à estrutura agrária, tem sido inclusive utilizada como indicador da capacidade das famílias terem acesso a alimentação adequada [6,9,10,11].

Em resumo, este trabalho permitiu estabelecer a importância relativa das variáveis de entrada em um modelo RNA, com o auxílio de técnicas de simulação de reamostragem. Desta forma, pode-se apontar a potencialidade do uso de redes neurais artificiais como metodologia de modelagem da mortalidade infantil, com vistas à predição e, até mesmo, diagnóstico, planejamento e avaliação.

5. Referências

[1] J.V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, *J. Clin. Epid*., 49(11), 1225-31, 1996.

[2] M. Buscema, A general presentation of artificial neural networks, *Substance Use & Misuse*, 32(1), 97-112, 1997.

[3] P.D. Wasserman, *Advanced Methods in Neural Computing*. 1st Ed. New York, Van Nostrand Reinhold, 1993.

[4] T.A. Hammad, M.F. Abdel-Wahab, N. DeClaris et al. Comparative evaluation of the use of artificial neural networks for modelling the epidemiology of schistossomiasis mansoni, *Trans R Soc Trop Med Hyg*, 90(4), 372-6, 1996.

[5] B. Efron, Bootstrap methods: another look at the jackknife. *Ann Stat*; 7, 26, 1979

[6] C.G. Victora, P.G. Smith, J.P. Vaughan, Social and environmental influences on child mortality in Brazil:

logistic regression analysis of data from census files, *Biosoc Sci* 16, 87-101, 1986.

[7] KIM, J., MUELLER, C.W. *Introduction to factor analysis.* Beverly Hills, Sage Publications, 1978.

[8] KIM, J., MUELLER, C.W. *Factor analysis-statistical methods and practical issues.* Beverly Hills, Sage Publications, 1985.

[9] R.J. Waldman, Income distribution and infant mortality , *The Quarterly J Economics.* November, 1283-1302, 1992.

[10] K.P. Gbsemete, D.A. Jonsson, A comparison of empirical models on determinants of infant mortality: a cross-sectional study on Africa, *Health Policy* 24, 155-74, 1993.

[11] L. Loslier, Ambiocontrol as a primary factor of health, *Soc Sci Med* 37(6), 735-743, 1993.

[12] E. Hertz, J.R. Herbert, J. Landon, Social and environmental factors and life expectancy, infant mortality, and maternal mortality rates: results of a cross-national comparison, *Soc Sci Med* 39(1), 105-114, 1994.

[13] R.M.V.R. Almeida, M. Thamer, E.O. Attinger, Characterization of health and social development, *J Biosoc Sci* 24(1) 1-9, 1992.

[14] A.G.G. Vasconcelos, R.M.R.V. Almeida, F.F. Nobre. The Path Analysis approach for infant mortality modelling, *Annals of Epidemiology* 8(4), 262-275 1998.

[15] H. Demuth, M. Beale, *Neural Network Toolbox–Matlab*. 3rd Ed. Natick (USA), The MathWorks, Inc. 1994.

[16] D.J. Livingstone, D.T. Manallack, I.V Tetko, 1997, Data modelling with neural networks: advantages and limitations, *J. Computer-Aided Molecular Design*, 11, 135-142.

[17] Infantosi AFC, Almeida RMVR, Gismondi RC. Modelagem da Mortalidade Infantil em Municípios Brasileiros usando Redes Neurais e Regressão Linear. *CBEB 2000 – Congresso Brasileiro de Engenharia Biomédica 2000*, 968-973. Florianópolis, SC.

Agradecimentos: este trabalho foi parcialmente financiado pelos projetos PRONEX 41.96.0937.00 e CNPq 521444/96-9.

Tabela 1. Variáveis inicialmente utilizadas e resultados da Análise Fatorial (variáveis selecionadas em negrito)

Grupo *educacional, econômico, moradia e meio ambiente:* (*variância explicada: 77%)*	Grupo *social, de produção e oferta de serviços de saúde:* *(variância explicada: 75,85%)*
Fator 1: menor escolaridade e renda (31,55%)	***Fator 1: estrutura agrária (24,11%)***
Pessoas sem rendimentos	**Pessoas ocupadas estabelecimentos agropecuários**
Pessoas economicamente ativas	Estabelecimentos agropecuários
Moradias com algum sistema de esgoto	Valor da produção agrícola
Pessoas alfabetizadas	Total de terra cultivada
Pessoas c/renda mensal de 1 a 3 SM	Valor da produção agropecuária
Domicílios com iluminação elétrica	Valor do rebanho
Pessoas c/renda mensal de 3 a 5 SM	Total de terra arável
Pessoas com 1º grau completo	
Pessoas c/renda mensal l < 1 SM	***Fator 2: oferta de emprego e serv. públ. de saúde***
Moradias com água encanada	Hospitais públicos
	Leitos estabelecimentos públicos
Fator 2: maior renda (26,68%)	**Estabelecimentos comerciais**
Pessoas c/renda mensal de 10 a 20 SM	Médicos estabelecimentos públicos
Pessoas c/renda mensal > 20 SM	Valor produção industrial
Salários estabelecimentos de serviços	
Pessoas c/renda mensal 5 a 10 SM	***Fator 3: oferta de serv. particulares de saúde***
Aparelhos telefonicos	**Leitos estabelecimentos particulares**
Salários pagos comércio	Internações
Automóveis particulares e de trabalho	Médicos estabelecimentos particulares
	Hospitais particulares
Fator 3: maior escolaridade (18,94%)	***Fator 4: prestação de serviços (11,37%)***
Pessoas com segundo grau completo	Receita c/a prestação de serviços
Pessoas com curso superior completo	**Pessoas ocupadas serviços**
Pessoas com 8 ou mais anos de estudo	
Pessoas com 11 ou mais anos de estudo	***Fator 5: seguridade social e estrutura industrial***
	Auxílios-doença concedidos
	Aposentados por invalidez
	Pessoas ocupadas estabelecimentos industriais
	Estabelecimentos industriais

Tabela 2. Estimativas *bootstrap* de R_B^2 e *EQM*, estimativas de redução do R_B^2 e de elevação do EQM_B. Desvio-padrão *bootstrap* entre parêntesis. ($R_{B\ \mathrm{RNA}}^2 = 0{,}65$, $EQM_{B\ \mathrm{RNA}} = 12{,}97 \times 10^3$)

Variável excluída	R_B^2 (replicação *bootstrap* do R^2)	redução R_B^2 (valor absoluto)	Estimativa de redução R_B^2 (%)	EQM_B (x 10^3) (replicação b*ootstrap EQM*)	elevação EQM_B (valor absoluto x 10^3)	elevação EQM_B (%)
alfabet	0,51 (0,10)	0,14	21,54	18,34 (3,83)	5,37	41,40
estcom	0,53 (0,08)	0,12	18,46	17,20 (3,19)	4,23	32,61
agrária	0,54 (0,08)	0,11	16,92	17,04 (3,13)	4,07	31,38
escolaridade	0,60 (0,05)	0,05	7,69	14,84 (1,78)	1,87	14,42
leitos	0,62 (0,11)	0,03	4,62	14,63 (2,90)	1,66	12,80
telefone	0,64 (0,08)	0,01	1,54	13,28 (3,08)	0,31	2,39
leitos	0,62 (0,11)	0,03	4,62	14,63 (2,90)	1,66	12,80
indústria	0,64 (0,08)	0,01	1,54	13,39 (2,99)	0,42	3,24
serviços	0,64 (0,08)	0,01	1,54	13,44 (3,02)	0,47	3,62

alfabet – número de pessoas alfabetizadas (por habitantes com 5 ou mais anos de idade).
agrária – pessoas ocupadas nos estabelecimentos agropecuários (por 1.000 habitantes entre 20 e 65 anos).
estcom – número de estabelecimentos comerciais (por 1.000 habitantes).
telefone – número de aparelhos telefônicos (por 1.000 habitantes).
escolaridade – número de pessoas com 8 ou mais anos de estudos (por 1.000 habitantes com 20 anos ou mais).
leitos – número de leitos em estabelecimentos particulares (por 1.000 habitantes).
indústria – número de pessoas ocupadas nos estabelecimentos industriais (por habitantes entre 20 e 65 anos).
serviços – número de pessoas ocupadas nos estabelecimentos prestadores de serviços (por habitantes entre 20 e 65 anos).