Referências Bibliográficas

Agrawal, R., Imielinski, T and Swami, A.. Database minimng: A performance perspective, *IEEE Trans. Knowledge Data Eng.*,v. 5, Dec. 1993.

Aitchison, J. & Dunsmore, I. R.. *Statistical Prediction Analysis*. Cambridge University Press, 1975.

Anderson, J. A.. Diagnosis by logistic discriminant function: further practical problems and results. *Appl. Statist*., 23, 1974, p. 397-404.

Azevedo A M.B.. Análise de sobrevida de crianças tratadas com leucemia linfoblástica aguda durante dez anos em duas instituições universitárias. Dissertação de Mestrado, Escola de Medicina da UFRJ. 2003.

Battiti, R.. Using mutual information for selecting features in supervised Neural net learning. *IEEE Trans. Neural Networks*, v. 5, 1994, p. 537-550.

Bishop, C. M.. Neural Networks for Pattern Recognittion. Oxford. Clarendon Press. 1995.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C.. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

Costa, E.S.., Implementação da detecção de doença residual mínma na Leucemia Linfoblástica Aguda na infância por citometria de fluxo, Pós-graduação em clinica médica UFRJ. Tese de Metrado. Orientadores: Marcelo Land e Adriana Bonomo, 2003.

Cover, T. M., J. Thomas, A.. *Elements of Information Theory*. New York: Wiley, 1991.

Cybenco, G..Approximation by superposition of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, v.2, 1989, p. 303-314

Daberllay, G., Klan, P.. An information-theoretic Adaptive Method for Time Series Forecasting. *Neural Networks World*, 1997, p. 227-238

Daberllay, G., Slama, M.. Forecasting the Short-Term Demand for Electricity: Do Neural Networks Stand a Better Chance? *International Journal of Forecasting*, v. 16, 2000, p. 71-83

De Castro, L. N. & Von Zuben, F. J.. In Improving Pruning Technique with Restart for the Kohonen Self_Organizing Feature Map, Proc. Do IJCNN, 3, (1999a), pp.1916-1919.

De Castro, L. N. & Von Zuben, F. J., "Neural Networks with Adaptive Activation Functions: A Second Order Approach", Proc. do SCI/ISAS'99, 3, (1999b), pp. 574-581.

De Castro, L. N., Iyoda, E. M., Santos, E. P. & Von Zuben F. J.. "Redes Neurais Construtivas: Uma Abordagem Comparativa", Anais do IV CBRN, 1999, pp. 102-107.

Djavan et all. Novel Artificial Neural Network for Early detection of Prostate Cancer – *Journal of Clinical Oncology*, v.20, n. 4 (February 15), 2002, p. 921-929.

Draper, N. R. And Smith, H.. Applied Regression Analysis. 2nd. New York: Wiley, 1981.

Duda, R.O. & Hart, P.E.. *Pattern Classification and Scene Analysis*. New York. n Wiley, 1973.

Fisher, R. A.. "The use of measurements in taxonomic problems" – *Annals of Eugenics*, v. 7:176-184, 1936.

Foresee, F. D., and M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization," *Proceedings of the 1997 International Joint Conference on Neural Networks*, 1997.

Fraser, A. M. & Swinney, H. L.. "Independent coordinates for strange attractors from mutual information", *Phys Rev.*, v. 33, n. 2, 1986.

Fukunaga, K.. *Introduction to Statistical Pattern Recognition*. Academic Press, New York. 1972.

Greaves, M.. A natural history for pediatric acute leukemia. Blood. 1993, 82, p. 1043-1051.

Gujarati, D. N.. *Econometria Básica*, Makron Books, 2000.

Hagan, M. T., Menhaj, M.. "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, v. 5, n. 6, 1994, p.989-993.

Haykin, S.. *Neural Networks: a comprehensive foundadtion*, Prentice-Hall, 1998.

Hermans, J. & Habbema, J. D. F.. Comparison of five methods to estimate posterior probabilities. *EDV in Medizin und Biologie*, 6, 1975, p. 14-19.

Hosmer, D. & Lemeshow, S.. *Applied Logistic Regression. Wiley series in probability and mathematical statistics*. New York. Jonh Wiley & Sons. 1989

Johnson A. R. & Wichern D. W. *Applied multivariate statistical analysis*, 4th. Ed., Prentice Hall, 1998.

Joliffe, I. T.. *Principal Component Analysis.* New York: Springer-Verlag, 1986.

Kattan, M.. Statistical Prediction Models, Artificial Neural Networks, and the Sophism "I Am A Patient, Not a Statistic". *Journal of Clinical Oncology*, v.20, n. 4(February 15), 2002, p. 885-887.

Kersey J. H.. Fifty years of studies of biology and therapy of childhood leukemia. Blood. 1997, 90, p. 4243-4251.

Krusinska, E. & Liebhart, J.. Robust discriminant functions in assisting medical diagnosis: application to the chronic obturative lung disease data. *Biometrical Journal*, 32, 1990, p. 915-929.

Kwak, N. and Choi, C.. Input Feature Selection for Classification Problems. *IEEE Trans. Neural Networks*, v. 13, no.1, 2002, p. 143-159.

Kwok, T. Y. & Yeung, D. Y.. Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems, *IEEE Trans. On Neural Networks*, 8(3), 1997, pp. 630-645.

Larson, H. J.. *Introduction to Probability Theory and Statistical Inference*, Wiley, 1982.

MacKay, D. J. C., "Bayesian interpolation," *Neural Computation*, v. 4, n. 3, 1992, p. 415-447.

Macrini, L., Pedreira C.E., Costa E.S, Land M.. Variable selection and Neural networks applied to the classification of risk of adverse event in childhood leukemia, *Proceedings of 5th IFAC 2003 Symposium on Modelling and Control of Biomedical Systems*, Melbourne, Australia, 2003a.

Macrini, J.L.R.; Pedreira, C.E.; Sobral, E.; Land M.. Seleção de Variáveis e RedesNeurais: Uma Aplicação de Classificação de Risco de Evento Adverso em Leucemia Infantil. *VI Congresso Brasileiro de RedesNeurais*. Sao Paulo, Anais, 2003b.

Marquardt, D.. "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *SIAM J. Appl. Math.* v. 11, 1963, p. 431-441.

Medeiros, M.C. & Pedreira, C.E.. "What are the effects of forecasting linear time series with neural networks?", *Engineering Intelligent Su]ystems*, v.9, p.237-242, 2001.

McCulloch, W. S. & Pitts, W.. "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics.* v. 5, 1943, p. 115-133.

Nguyen, D., B. Widrow, "The truck backer-upper: An example of self-learning in neural networks," *Proceedings of the International Joint Conference on Neural Networks*, vol 2, 1989, pp. 357-363.

Ohno-Machado, L.. Methodological Review Modelong Medical Prognosis:Survival Analysis Techniques. *Journal of Biomedical Informatics* 34, 2002, p. 428-439

Ohno-Machado, L. & Musen, M. A. Modular Neural Networks for Medical Prognosis: Quantifying the Benefits of Combining Neural Networks for Survival Prediction. Knowledge Systems Laboratory, *Medical Computer Science*, February, 1996.

Principe, J. C., Euliano, N. R. and Lefebvre, W. C.. Neural and Adaptive Systems: Fundamentals Through Simulations, John Wiley. 2000.

Quinlan, R.. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

Riehem H.. *Estudo terapêutico Multicêntrico da Sociedade da "Gesellschaft für Pädiatrische Onkologie und Hämatologie – BFM95 – para o Tratamento de Crianças e Adolescentes com Leucemia Linfoblástica Aguda"*. Direção do Estudo: Prof. Dr.Riehem h.c.- Tradução Dra L. Laun em 18/01/1996. 1995.

Rivera, G. K., Pinkel, D., Simone, J. V., Hancock, M. L., Crist, M. L.. Treatment of Acute Lymphoblastic Leukemia - 30 Years' Experience at St Jude Children' research Hospital. N Engl. J Med., 1993, 329, p. 1289-1295.

Schrappe, M., Beier, R., Bartram, B.. New treatment Strategies in Childhood Acute Leukemia. Best Practice and Research Clinical Hematology. 2003, 15(4), p. 729-740

Setiono, R., Liu, H.. Neural network feature selector. *IEEE Trans. Neural Networks*, v. 8, 1997, p. 654-661.

Shannon, C. E. & Weaver, W.. *The Mathematical Theory of Communication*. Urbana, IL: Univ. Illinois Press, 1949.

Smith, M., Arthur D., Camitta B., et al. Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. Clin Oncol., 14, 1996, p. 18-24.

Stone, M.. "Cross-validacion: A review" – Mathematische Operationsforschung Statistishen, *Serie Statistics*, v. 9, 1978, p. 127-139.

Stone, M.. "Cross-validacion choice and assessment of statistical predictions"– *Journal of the Royal Statistical Society*, v. B36, 1974, p. 111-133.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A.M., Habemma, J. D. F. & Gelpke, G. J.. Comparison of discrimination techniques applied to a complex data set of head injuried patients (com discussão). J.R. *Statist. Soc*. A, 144, 1981, p. 145-175.

Yao, X. and Liu, Y. (1999). Neural networks for breast cancer diagnosis, *Proc. of the 1999 Congress on Evolutionary Computation, Vol. 3*, IEEE Press, Piscataway, NJ, USA, 1999, p. 1760-1767.

Anexo A - Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U)

A.1
Introdução

A seleção de variáveis de entrada constitui uma das fases mais importantes em problemas de classificação. As variáveis de entrada podem ser classificadas como pertinentes, irrelevantes ou redundantes, e o que se pretende é selecionar somente aquelas que sejam pertinentes (Kwak & Choi, 2002). Nesta tese utilizou-se o algoritmo de Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U). O algoritmo proposto é aplicado ao problema de estimação do risco de crianças portadoras de Leucemia Linfoblástica Aguda submetidas a diferentes protocolos.

Seleção de variáveis tem um papel fundamental na classificação de sistemas como Redes Neurais. Problemas de seleção de variáveis foram pesquisados por vários autores como Battiti (1994), Joliffe (1986) e Agrawal *et al* (1993). Um dos métodos mais populares para lidar com este problema é a análise de componentes principais (PCA) (Joliffe, 1986). Porém, caso se queira preservar os dados originais, este método não é desejável. Recentemente, uma das contribuições mais importantes trata do método de árvore de decisão. Os atributos pertinentes são descobertos um a um iterativamente (Quinlan, 1993) (Breiman *et al*, 1984). Setiono e Lui propuseram um algoritmo de seleção de variáveis baseado em uma árvore de decisão excluindo a variável de entrada da Rede Neural uma a uma e treinando novamente a Rede repetidamente (Setiono&Lui, 1997). O classificador com poda dinâmica (CDP) (Agrawal *et al.*,1993) também se baseia numa árvore de decisão a qual faz uso da informação mútua da entrada com a saída. Este método é eficiente e encontra regras mapeando entrada e saída, mas requer muita memória. O seletor de variáveis (Battiti, 1994) também usa a informação mútua entre entrada e saída como o CDP. A regressão stepwise (Drapper&Smith, 1981) é também considerada uma técnica padrão na seleção de variáveis fazendo uso do teste F como critério de parar a seleção.

O algoritmo de Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U) investiga a limitação do seletor de variáveis proposto por (Battiti,1994) e se propõe superar esta limitação e melhorar o desempenho no processo de seleção de variáveis.

Feitas estas considerações, nas seções seguintes serão introduzidos alguns conceitos básicos da teoria da informação que serão usados na aplicação do algoritmo usado nesta tese.

A.2
Entropia e Informação Mútua

Sistemas de classificação em Redes Neurais mapeiam variáveis de entrada em classes de saída. Neste processo, existem variáveis que são importantes e variáveis irrelevantes, isto é, com pouca informação relativa à saída. Para resolver o problema de seleção de variáveis, tem-se que achar entradas que contenham muita informação sobre a saída e é necessária uma ferramenta para medir essa informação. A teoria da informação fornece um método para medir a informação de variáveis aleatórias: a entropia e a informação mútua (Shannon *et al*, 1949, Cover *et al*, 1991).

A entropia é uma medida de incerteza de variáveis aleatórias. Seja uma variável aleatória discreta com função de densidade de probabilidade (pdf) p(x). A entropia de X é definida como:

$$H(X) = -\sum_{x \in c} p(x) \log\, p(x) \qquad \text{(A.1)}$$

Para duas variáveis aleatórias discretas e com pdf conjunta p(x,y), a entropia é definida como:

$$H(X,Y) = -\sum_{X \in c} \sum_{y \in c} p(x,y) \log\, p(x,y) \qquad \text{(A.2)}$$

Quando certas variáveis são conhecidas e outras não, a incerteza é medida pela entropia condicional:

$$H(Y \mid X) = \sum_{x \in c} p(x) H(Y \mid X = x) = -\sum_{x \in c} p(x) \sum_{y \in c} p(y \mid x) \log p(y \mid x) = \qquad \text{(A.3)}$$

$$= -\sum_{x \in c} \sum_{y \in c} p(x,y) \log\, p(y \mid x)$$

A entropia conjunta e a entropia condicional têm a seguinte relação:

$$\begin{aligned} H(X,Y) &= H(X)+H(Y|X) \\ &= H(Y)+H(X|Y) \end{aligned} \tag{A.4}$$

Esta relação é conhecida como regra da cadeia e implica que a entropia total das variáveis aleatórias X e Y é a entropia de X mais a entropia restante de Y dado X.

A informação contida em duas variáveis aleatórias é definida como a informação mútua entre duas variáveis aleatórias.

$$I(X,Y) = \sum_{x \in c} \sum_{y \in g} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{A.5}$$

Se a informação mútua entre duas variáveis aleatórias é grande (pequena), significa que as duas variáveis são muito (pouco) relacionadas. Se a informação mútua é próxima de zero, as duas variáveis aleatórias são independentes.

A informação mútua e a entropia têm as seguintes relações:

$$\begin{aligned} I(X,Y) &= H(X)-H(X|Y) \\ I(X,Y) &= H(Y)-H(Y|X) \\ I(X,Y) &= H(X)+H(Y)-H(X,Y) \\ I(X,Y) &= I(Y,X) \\ I(X,X) &= H(X) \end{aligned} \tag{A.6}$$

Para variáveis aleatórias contínuas, a entropia diferencial e a informação mútua são definidas como:

$$\begin{aligned} H(X) &= -\int p(x) \log\, p(x)dx \\ I(X,Y) &= \int p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}\,dxdy \quad (7) \end{aligned} \tag{A.7}$$

Salienta-se que é praticamente impossível achar exatamente esta função de densidade de probabilidade e executar sua integração. Por esse motivo, dividi-se o

espaço da variável de entrada contínua em várias partições discretas e calcula-se a entropia e a informação mútua usadas nas definições para os casos discretos. O erro inerente ao processo de conversão de variáveis contínuas para variáveis discretas é um valor constante que depende do número de partições em que se divide o espaço contínuo (Fraser *et al,* 1986).

A.3
Algoritmo de Seleção de Variáveis

A.3.1
O Problema de FRn - k

No processo de selecionar variáveis de entrada, é desejável reduzir o número de entradas excluindo variáveis irrelevantes ou redundantes dos dados. Este conceito é formalizado selecionando k variáveis de um conjunto de n variáveis chamado de problema de "redução de variável" (Battiti, 1994). O processo será apresentado a seguir:

[**FRn - k**]: Dado um conjunto inicial de n variáveis, encontre o subconjunto com k<n variáveis que são "a máxima informação" sobre a classe de saída.

Como visto na seção anterior, a informação mútua entre duas variáveis aleatórias é a quantidade de informação comum entre essas variáveis. O problema de selecionar variáveis de entradas pode ser resolvido calculando a informação mútua (IM) entre variáveis de entrada e classes de saídas. Se a informação mútua entre variáveis de entrada e classes de saída pudesse ser obtida com precisão, o problema FRn - k poderia ser reformulado como segue:

[FRn - k]: Dado um conjunto inicial F com n variáveis e C classes de saída, ache o subconjunto $S \subset F$ com k variáveis que minimizam H(C|S), isto é, que maximizam a informação mútua I(C;S).

O algoritmo de seleção que usa informação mútua é como segue:

1) (inicialização) conjunto F $\leftarrow$ "conjunto inicial com n variáveis", S $\leftarrow$"conjunto vazio."

2) (calculo da IM com a classe de saída), $\forall \phi_i \in F$, compute $I(C; \phi_i)$.

3) (seleção da primeira variável) ache a variável que maximiza $I(C; \phi_i)$, faça $F \leftarrow F-\{\phi_i\}$, $S \leftarrow \phi_i$.

4) repita até que o número desejado de variáveis seja selecionado.

a) (Cálculo da IM conjunta entre variáveis), $\forall \phi_i \in F$, compute $I(C; \phi_i,S)$.

b) (Seleção da próxima variável) escolha a variável $\phi_i \in F$ que maximiza $I(C; \phi_i: S)$ e faça $F \leftarrow F-\{\phi_i\}$, $S \leftarrow \phi_i$.

5) saída do conjunto S contém as variáveis selecionadas.

A realização desse algoritmo de seleção é praticamente impossível de ser executado. Isto é, este algoritmo é inviabilizado pelo tamanho do vetor de variáveis no cálculo de $I(C; \phi_i; S)$.

A.3.2
Seleção de Variáveis sob Informação Mútua (MIFS)

O algoritmo MIFS é similar ao algoritmo de seleção anterior com exceção do Passo 4. Em vez de se calcular $I(C;\phi_i,S)$, usa-se somente as seguintes informações mútuas: $I(C, \phi_i)$ e $I(\phi_i; \phi_i)$ (Battiti,1994). No MIFS, o passo 4 do algoritmo de seleção é substituído como segue:

4) repita até o número desejado de variáveis a serem selecionadas.

(Calculo da IM entre variáveis) para todos os pares de variáveis $(\phi_i; \phi_s)$ com $\phi_i \in F$, $\phi_s \in S$, calcule $I(\phi_i; \phi_s)$

(Seleção da próxima variável) escolha a variável com $\phi_i \in F$ que maximiza $I(C, \phi_i) - \beta \sum_{\phi s \in S} I(\phi_i; \phi_s)$ e faça $F \leftarrow F-\{\phi_i\}$, $S \leftarrow \phi_i$.

Nesse ponto, β é o parâmetro de redundância. Se $\beta=0$, o algoritmo seleciona variáveis na ordem da informação mútua entre variáveis de entrada e saída. A redundância entre as variáveis de entrada nunca é refletida. Quando $\beta>0$ a redundância é reduzida.

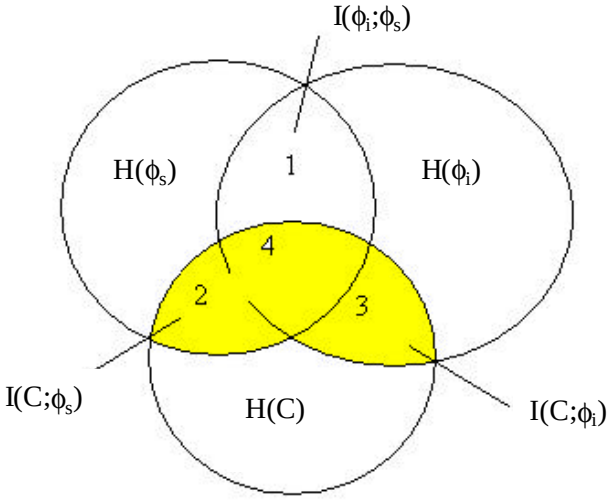

Figura A1 – Relação entre Variáveis de Entrada e Classe de Saída

A relação entre variáveis de entrada e saída pode ser representada na figura A1. O algoritmo de seleção usa a informação mútua para escolher a variável ϕ_i que maximiza a informação mútua conjunta $I(C; \phi_i; \phi_s)$ que são as áreas 2, 3, e 4. Como $I(C; \phi_s)$ (área 2 e 4) é comum para todas as variáveis não selecionadas ϕ_i, calculando-se a informação mútua conjunta $I(C; \phi_i; \phi_s)$, o algoritmo seleciona a variável que maximiza a área 3. Por outro lado, o algoritmo MIFS seleciona a variável que maximiza $I(C; \phi_i) - \beta\, I(\phi_i; \phi_s)$. Para β=1, isto corresponde a área 3 subtraída da área 1.

Entretanto, se uma variável a ser selecionada é fortemente relacionada com alguma variável já selecionada, a área 1 é grande e isto pode degradar o desempenho do algoritmo. Por isto, o MIFS pode não trabalhar bem em problemas não lineares.

A.3.3
Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U)

O algoritmo maximiza a informação mútua $I(C;\phi_i, \phi_s)$ (áreas 2, 3, e 4 na figura A1) através da seguinte expressão:

$$I(C; \phi_i, \phi_s) = I(C; \phi_s) + I(C; \phi_i \mid \phi_s) \qquad (A.8)$$

Onde $I(C; \phi_i \mid \phi_s)$ representa a informação mútua restante entre a classe de saída C e a variável ϕ_i para um dado ϕ_s. Isto corresponde a área 3 na figura A1,

onde a área 2 mais a área 4 representa I(C; ϕ_s). Como I(C; ϕ_s) é comum para todas as variáveis candidatas a serem selecionadas pelo algoritmo, não há nenhuma necessidade de se calcular essa informação mútua. Assim o algoritmo tenta achar a variável que maximiza I(C; ϕ_i | ϕ_s) (área 3). Porém, calcular I(C; ϕ_i | ϕ_s) requer tanto trabalho quanto calcular I(C; ϕ_i, ϕ_s). O algoritmo então faz uma aproximação de I(C; ϕ_i | ϕ_s) com I(ϕ_i, ϕ_s) que são relativamente fáceis de se calcular. A informação mútua condicional I(C; ϕ_i | ϕ_s) pode ser representada como:

$$I(C; \phi_i | \phi_s) = I(C; \phi_i) - \{ I(\phi_s, \phi_i) - I(\phi_s; \phi_i | C)\} \quad (A.9)$$

Onde I(ϕ_s, ϕ_i) corresponde as áreas 1 e 4 e I(ϕ_s ; ϕ_i | C) corresponde a área 1. Assim o termo I(ϕ_s , ϕ_i) - I(ϕ_s ; ϕ_i | C) corresponde a área 4 na figura A1. O termo I(ϕ_s ; ϕ_i | C) significa a informação mútua entre a variável já selecionada ϕ_s e o candidato ϕ_i para uma determinada classe. Se condicionando pela classe C a razão entre a entropia de ϕ_s e a informação mútua entre ϕ_s e ϕ_i não mudar, a seguinte relação pode ser escrita:

$$\frac{H(f_s | C)}{H(f_s)} = \frac{I(f_s, f_i | C)}{I((f_s, f_i)} \quad (A.10)$$

onde I(ϕ_s ; ϕ_i | C) pode ser representada por:

$$I(f_s, f_i | C) = \frac{H(f_s | C)}{H(f_s)} I(f_s; f_i) \quad (A.11)$$

Usando a equação acima e A9 tem-se que:

$$I(C; f_i | f_s) = I(C; f_i) - \left(1 - \frac{H(f_s | C)}{H(f_s)}\right) I(f_s; f_i) = \\ = I(C; f_i) - \frac{I(C; f_s)}{H(f_s)} I(f_s; f_i) \quad (A.12)$$

A condição A10 é mais significativa quando sua distribuição é uniformemente distribuída ao longo da região de H(ϕ_s) da figura A1. Por essa razão refere-se ao algoritmo como MIFS-U.

Sendo assim, o passo (4) do algoritmo revisto segue:

4) repetir até que o número desejado de variáveis seja selecionado.

a) (Calculo da entropia), $\forall f_s \in S$ compute H(ϕ_s).

b) (Calculo da IM entre variáveis) para todos os pares de variáveis (ϕ_i, ϕ_s) com $f_i \in F \; e \; f_s \in S$, calcule I(ϕ_s; ϕ_i).

c) (Seleção da próxima variável) escolhe uma variável $f \in F$ que maximiza $I(C, f_i) - b \sum_{f_s \in S} (I(C; f_s) / H(f_s)) I(f_i; f_s)$; e faça F←F-{$\phi_i$}, S ← ϕ_i.

Se β=0, o algoritmo seleciona variáveis na ordem da informação mútua entre variáveis de entrada e saída. A redundância entre as variáveis de entrada nunca é refletida. Quando β>0, o algoritmo exclui as variáveis redundantes mais eficazmente. Em geral nós podemos fixar b =1 (Breiman *et al*, 1984). Para todas as experiências desta tese fixou-se b =1.

Os valores encontrados para informação mútua com o desfecho, a entropia e informação mútua entre as variáveis podem ser visualizadas nos quadros A1, A2 e A3 a seguir.

Quadro A1: Valor da Informação Mútua com o Desfecho

Variáveis	IM
Blastos	0.203
Leucócitos	0.197
Idade	0.086
Hemoglobina	0.069
Figado	0.063
RCD8	0.036
Baço	0.033
Imunofenotipo T	0.032
Sexo	0.024
Raça	0.012
G. Cervical	0.010
Febre	0.004
Astenia	0.004
Sangramento	0.002
Dor Ossea	0.002
G. Virilha	0.001
Ganglio	0.001
Dor	0.001
Infecção	0.001
FAB	0.000

Quadro A2: Valor da Entropia

Variáveis	Entropia
Sexo	1.038
Astenia	1.043
Dor	1.041
Dor Ossea	1.048
Figado	3.013
Baço	2.407
Gânglio	0.936
G. Cervical	0.983
G. Virilha	1.041
Sangramento	1.049
Febre	0.889
Infecção	1.022
Leucócitos	2.950
RC8D	0.814
Idade	3.113
FAB	0.670
Raça	1.016
Blastos	2.551
Hemoglobina	3.274
Imunof. T	0.738

Quadro A3: Valor da Informação Mútua entre Variáveis

Variáveis	Sexo	Astenia	Dor	Dor Ossea	Figado	Baço	Gânglio	G. Cervical	G. Virilha	Sangramento	Febre	Infecção	Leucócitos	RC8D	Idade	FAB	Raça	Blastos	Hemoglobina	Imunof. T
Sexo	0.000	0.004	0.017	0.017	0.063	0.023	0.016	0.009	0.012	0.005	0.000	0.000	0.203	0.008	0.125	0.004	0.041	0.131	0.051	0.000
Astenia	0.004	0.000	0.000	0.000	0.083	0.066	0.001	0.010	0.004	0.000	0.029	0.005	0.139	0.002	0.039	0.006	0.001	0.141	0.085	0.003
Dor	0.017	0.000	0.000	0.623	0.100	0.026	0.004	0.000	0.006	0.001	0.001	0.004	0.193	0.004	0.095	0.009	0.001	0.221	0.123	0.008
Dor Ossea	0.017	0.000	0.623	0.000	0.081	0.030	0.003	0.000	0.004	0.004	0.005	0.000	0.158	0.008	0.063	0.002	0.001	0.182	0.085	0.020
Figado	0.063	0.083	0.100	0.081	0.000	0.564	0.043	0.044	0.045	0.098	0.113	0.059	0.655	0.180	0.528	0.020	0.042	0.706	0.634	0.034
Baço	0.023	0.066	0.026	0.030	0.564	0.000	0.015	0.055	0.051	0.089	0.028	0.026	0.463	0.140	0.374	0.026	0.064	0.509	0.331	0.015
Gânglio	0.016	0.001	0.004	0.003	0.043	0.015	0.000	0.564	0.261	0.032	0.000	0.005	0.171	0.004	0.099	0.003	0.000	0.212	0.061	0.001
G. Cervical	0.009	0.010	0.000	0.000	0.044	0.055	0.564	0.000	0.213	0.034	0.001	0.003	0.169	0.000	0.084	0.002	0.000	0.202	0.051	0.005
G. Virilha	0.012	0.004	0.006	0.004	0.045	0.051	0.261	0.213	0.000	0.016	0.001	0.001	0.157	0.001	0.054	0.012	0.002	0.143	0.051	0.012
Sangramento	0.005	0.000	0.001	0.004	0.098	0.089	0.032	0.034	0.016	0.000	0.043	0.004	0.206	0.008	0.066	0.001	0.008	0.194	0.091	0.016
Febre	0.000	0.029	0.001	0.005	0.113	0.028	0.000	0.001	0.001	0.043	0.000	0.112	0.124	0.005	0.083	0.000	0.000	0.137	0.026	0.021
Infecção	0.000	0.005	0.004	0.000	0.059	0.026	0.005	0.003	0.001	0.004	0.112	0.000	0.150	0.004	0.076	0.002	0.010	0.193	0.074	0.001
Leucócitos	0.203	0.139	0.193	0.158	0.655	0.463	0.171	0.169	0.157	0.206	0.124	0.150	0.000	0.283	0.482	0.097	0.145	1.908	0.618	0.164
RC8D	0.008	0.002	0.004	0.008	0.180	0.140	0.004	0.000	0.001	0.008	0.005	0.004	0.283	0.000	0.043	0.001	0.000	0.277	0.083	0.002
Idade	0.125	0.039	0.095	0.063	0.528	0.374	0.099	0.084	0.054	0.066	0.083	0.076	0.482	0.043	0.000	0.045	0.033	0.518	0.599	0.083
FAB	0.004	0.006	0.009	0.002	0.020	0.026	0.003	0.002	0.012	0.001	0.000	0.002	0.097	0.001	0.045	0.000	0.000	0.077	0.047	0.000
Raça	0.041	0.001	0.001	0.001	0.042	0.064	0.000	0.000	0.002	0.008	0.000	0.010	0.145	0.000	0.033	0.000	0.000	0.177	0.051	0.002
Blastos	0.131	0.141	0.221	0.182	0.706	0.509	0.212	0.202	0.143	0.194	0.137	0.193	1.908	0.277	0.518	0.077	0.177	0.000	0.552	0.152
Hemoglobina	0.051	0.085	0.123	0.085	0.634	0.331	0.061	0.051	0.051	0.091	0.026	0.074	0.618	0.083	0.599	0.047	0.051	0.552	0.000	0.085
Imunof. T	0.000	0.003	0.008	0.020	0.034	0.015	0.001	0.005	0.012	0.016	0.021	0.001	0.164	0.002	0.083	0.000	0.002	0.152	0.085	0.000

Anexo B – Redes Neurais Artificiais

B.1
Introdução

Desde o final da década de 80, as Redes Neurais artificiais são uma metodologia, na fronteira da estatística com a inteligência artificial, eficiente e capaz de resolver uma gama de problemas importantes. Na área médica podem ser encontradas diversas aplicações, tais como Kattan (2002), Djavan *et al* (2002), Ohno-Machado (2002), Ohno-Machado & Musen (1996) e Yao & Liu (1999) dentre outros. Na literatura, alguns livros se destacam como de suma importância: Haykin (1998), Bishop (1995), Duda and Hart (1973) dentre outros.

A motivação original desta metodologia[1] foi a tentativa de modelar a Rede de neurônios humanos visando compreender o funcionamento do cérebro. Portanto, como o próprio nome da metodologia revela, sua motivação inicial foi a de realizar tarefas complexas que o cérebro executa com elevada efetividade (por exemplo: reconhecimento de padrões, percepção e controle motor) através da simulação de seu funcionamento.

Segundo Haykin (1998), uma Rede Neural artificial (RNA) é um sistema de processamento massivamente paralelo, composto por unidades simples com capacidade natural de armazenar conhecimento e disponibilizá-lo para uso futuro.

Do ponto de vista neurofisiológico, muito pouco se conhece sobre o funcionamento dos neurônios e suas conexões o que compromete a fidelidade destes modelos em fisiologia. As RNAs assemelham-se ao cérebro em dois aspectos:

- Elas extraem conhecimento do ambiente através de um processo de *aprendizagem* ou *treinamento*; e
- Os pesos das conexões entre os neurônios, conhecidos como *pesos sinápticos*, são utilizados para armazenar o conhecimento adquirido.

[1] Haykin (1998).

A natureza das RNAs faz com que seu estudo seja multidisciplinar, envolvendo pesquisadores de diversas áreas, como neurofisiologia, psicologia, física, computação, engenharia, estatística, entre outras.

Cientistas da área de computação têm em vista a construção de computadores dotados de processamento paralelo e distribuído, buscando superar as limitações impostas pelos computadores atuais, que realizam processamento serial simbólico.

Inspirados na habilidade apresentada pelos seres humanos e outros animais no desempenho de funções como o processamento de informação sensorial e a capacidade de interação com ambientes pouco definidos, os engenheiros, por exemplo, estão preocupados em desenvolver sistemas artificiais capazes de desempenhar tarefas semelhantes. Habilidades como capacidade de processamento de informação incompleta ou imprecisa e generalização são propriedades desejadas em tais sistemas.

McCulloch & Pitts (1943) projetaram a estrutura que é conhecida como a unidade básica de uma Rede Neural. Estes pesquisadores propuseram um modelo de neurônio como uma unidade de processamento binária e provaram que estas unidades são capazes de executar várias operações lógicas (OU, AND, etc.). Este modelo, apesar de muito simples, fornece uma grande contribuição para as discussões sobre a construção dos primeiros computadores digitais, permitindo a criação dos primeiros modelos matemáticos de dispositivos artificiais que buscavam analogias biológicas. Matematicamente, o neurônio da Figura B2 pode ser expresso por:

$$y = f(u) = f(x_1 w_1 + x_2 w_2 + ... + x_n w_n) = f(w^T x) \qquad \text{(B.1)}$$

onde y é a saída do neurônio, u é a ativação do neurônio, f(.) sua função de ativação, x_i (i = 1,...,n) é o i-ésimo componente do vetor $\boldsymbol{x}$ de entradas, e w_i (i = 1,...,n) é o i-ésimo componente do vetor $\boldsymbol{w}$ de pesos do neurônio.

B.2
Variáveis Principais

A comparação com a neurofisiologia foi apenas uma motivação original da qual pouco sobrou além do nome da ferramenta. Desde o final da década de 80, as Redes Neurais artificiais são uma metodologia, na fronteira da estatística com a

inteligência artificial, eficiente e capaz de resolver uma gama de problemas importantes. No exterior, em especial nos E.U.A, já se encontrou grande aplicabilidade fora dos muros acadêmicos sendo que, aqui no Brasil, começa-se a perceber seu grande potencial. Conceitualmente, uma Rede Neural artificial é um dispositivo tanto capaz de processar informação de forma distribuída quanto de incorporar conhecimento através de exemplos. Trata-se, portanto, de um processador capaz de extrair conhecimento experimental disponibilizando-o para uso prático (tomada de decisões, por exemplo).

As Redes Neurais artificiais têm sido desenvolvidas como generalizações de modelos matemáticos de cognição humana ou neurobiologia, assumindo que:

- O processamento da informação ocorre com o auxílio de vários elementos chamados *neurônio*s;
- Os sinais são propagados de um elemento a outro através de *conexõe*s;
- Cada conexão possui um *peso* associado, que, em uma Rede Neural típica, pondera o sinal transmitido; e
- Cada neurônio (ou unidade) possui uma *função de ativação* (geralmente não-linear), que tem como argumento a soma ponderada dos sinais de entrada, para determinar sua saída.

Uma grande vantagem de usar-se uma Rede Neural é a capacidade de resolver problemas sem a necessidade de definição de listas de regras ou de modelos explícitos. Isto possibilita tratar de situações onde é difícil criar modelos adequados da realidade ou situações com freqüentes mudanças no ambiente.

Atenta-se que grande parte desta sua adequabilidade funcional deve-se à sua capacidade em inferir relações não lineares complexas. Frente a estas suas propriedades, hoje, pode-se observar sua aplicabilidade principalmente nas áreas de classificação de padrões (em um sentido amplo) e de previsão.

Uma Rede Neural caracteriza-se pela capacidade de extrair conhecimento experimental e por disponibilizar este conhecimento para uso prático. Apesar da plausibilidade biológica ter sido apenas uma motivação original cabe aqui uma comparação. O cérebro desenvolve a função de, a partir da observação de dados (*input*), extrair informação disponibilizando-a para a tomada de decisões.

Sabe-se que o conhecimento é adquirido através de um processo de aprendizado. O mesmo acontece com as Redes Neurais artificiais. A informação é

armazenada em “densidades de conexão” conhecidas como “pesos sinápticos” (ou simplesmente pesos). O processo de aprendizado de uma Rede se dá através de um algoritmo que deve ser capaz de ajustar iterativamente os pesos de modo que se atinja o objetivo proposto.

A Rede Neural aprende, então, o ambiente através de um processo iterativo de modificação dos pesos de interligação, a partir de estímulos fornecidos pelo ambiente. O tipo de aprendizado é determinado pelo modo com que se promove a adaptação dos parâmetros e isso pode ser feito de dois modos:

1) Aprendizado Supervisionado – usa-se um conjunto de pares, entrada e saída, previamente conhecidos que representam a realidade;

2) Aprendizado Não Supervisionado – não se usa um conjunto de exemplos previamente conhecidos. Uma medida da qualidade da representação do ambiente pela Rede é estabelecida e os parâmetros são modificados de modo a otimizar esta medida. Este tipo de aprendizado é muito utilizado na área de reconhecimento de padrões.

Salienta-se que, nesta tese, usou-se o aprendizado supervisionado, ou seja, escolheu-se as variáveis referentes a sintomas e ou sinais, dados clínicos e laboratoriais como uma medida da representação do ambiente em estudo.

Em síntese, uma Rede Neural pode ser caracterizada por três aspectos principais: (1) padrão de conexões entre as unidades *(arquitetura* ou *estrutur*a), (2) método de determinação dos pesos das conexões *(algoritmo de treinamento* ou *aprendizage*m) e (3) *função de ativaçã*o.

B.2.1
Arquitetura

A forma pela qual os neurônios de uma RNA estão estruturados (interconectados) está intimamente relacionada ao algoritmo de aprendizagem a ser utilizado para treiná-la. Em geral é possível distinguir três classes fundamentais de arquiteturas: *Redes feedforward de uma única camad*a, *Redes feedforward de múltiplas camadas* e *Redes recorrente*s, sendo de interesse desta tese os dois primeiros casos.

B.2.1.1
Redes *Feedforward* de Uma Única Camada

No caso mais simples de Redes em camadas *(layer*s), tem-se uma camada de entrada com neurônios cujas saídas alimentam a última camada da rede. Geralmente, os neurônios de entrada são propagadores puros, ou seja, eles simplesmente repetem o sinal de entrada em sua saída distribuída. Por outro lado, as unidades de saída costumam ser unidades processadoras, como apresentado na Figura B1. A propagação de sinais nesta Rede é puramente unidirecional *(feedforwar*d): os sinais são propagados apenas da entrada para a saída, e nunca vice-versa. Esta arquitetura está ilustrada na Figura B1(a) e a direção de propagação dos sinais na Figura B1(b).

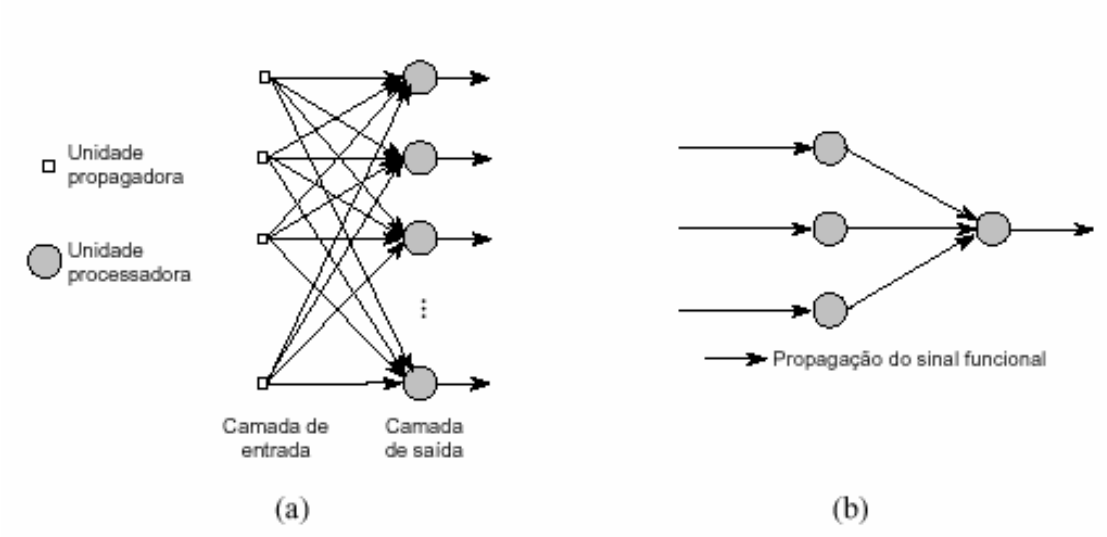

Figura B1. Redes Neurais tipo *feedforward* com uma Única Camada de Unidades Processadoras. (a) Arquitetura. (b) Sentido de Propagação do Sinal Funcional.

B.2.1.2
Redes *Feedforward* de Múltiplas Camadas

A segunda classe de Rede feedforward se distingue pela presença de uma ou mais camadas intermediárias ou escondidas (camadas em que os neurônios são efetivamente unidades processadoras, mas não correspondem à camada de saída). Adicionando-se uma ou mais camadas intermediárias, aumenta-se o poder computacional de processamento não-linear e armazenagem da rede. O conjunto de saídas dos neurônios de cada camada da Rede é utilizada como entrada para a

camada seguinte. A Figura B2(a) ilustra uma Rede feedforward de múltiplas (duas) camadas intermediárias.

As Redes feedforward de múltiplas camadas, são geralmente treinadas usando o algoritmo de retro-propagação do erro *(error backpropagatio*n), embora existam outros algoritmos de treinamento. Este algoritmo requer a propagação direta *(feedforwar*d) do sinal de entrada através da rede, e a retro-propagação (propagação reversa, ou *backpropagatio*n) do sinal de erro, como ilustrado na Figura B2(b).

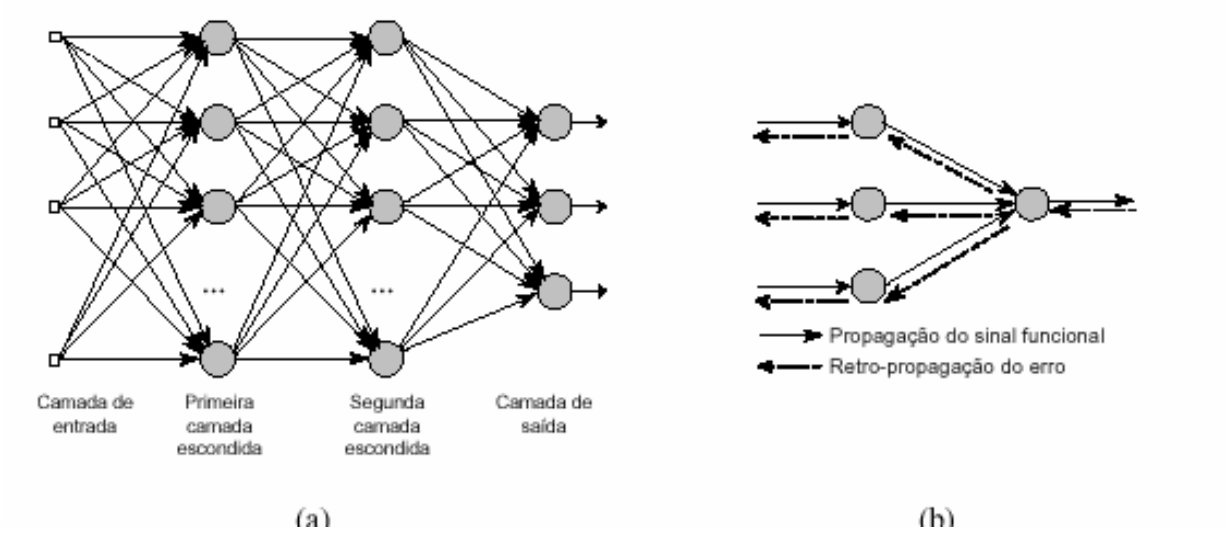

Figura B2. Redes Neurais tipo *feedforward* com Múltiplas Camadas. (a) Arquitetura. (b) Sentido de Propagação do Sinal Funcional e do Sinal de Erro.

B.2.2
Métodos de Estimação

A capacidade de *aprendizagem* é uma das variáveis marcantes das RNAs. Uma Rede Neural aprende, basicamente, através de um processo iterativo de ajuste de pesos e limiares (*bias*). Atualmente, existem processos mais sofisticados de aprendizagem (ou *treinament*o), que são capazes de ajustar não apenas os pesos da rede, mas também sua arquitetura e as funções de ativação dos neurônios (Kwok & Yeung, 1997, de Castro *et al*., 1999a,b; de Castro *et al*, 1999).

Segundo Haykin (1998), *Aprendizagem* (ou *treinament*o) é o processo pelo qual os parâmetros livres de uma Rede Neural são adaptados, através de um mecanismo de apresentação de estímulos fornecidos pelo ambiente no qual a Rede está inserida. O tipo de treinamento é definido pela forma na qual os parâmetros são modificados.

Esta definição de aprendizagem implica na seguinte seqüência de eventos:

- Apresentação de estímulos à RedeNeural;
- Alteração dos parâmetros livres da rede; e
- Novo padrão de resposta ao ambiente.

Os principais paradigmas de aprendizagem são: (1) supervisionada, (2) não-supervisionada, e (3) por reforço.

B.2.2.1
Aprendizagem Supervisionada

Trata-se de um paradigma de aprendizagem, no qual um *supervisor* possui conhecimento sobre o ambiente em que a Rede está inserida. Este conhecimento está representado sob a forma de um conjunto de amostras de *entrada-saída*. O ambiente, por sua vez, é desconhecido. A Figura B3 ilustra esta abordagem. Os parâmetros da Rede são ajustados pela combinação do sinal de entrada com um sinal de erro, que é a diferença entre a saída desejada e a fornecida pela rede.

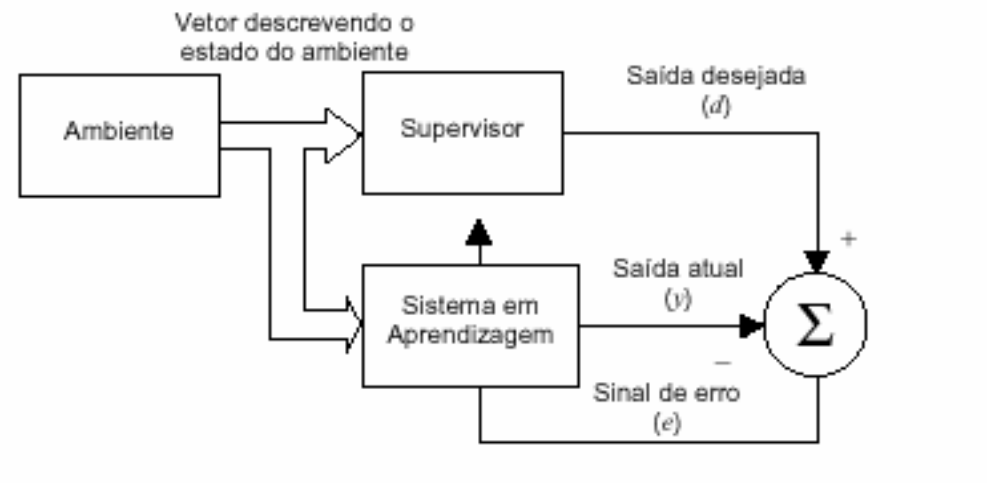

Figura B3. Diagrama de Blocos do Processo de Aprendizagem Supervisionada.

Seja *t* o índice que denota tempo discreto ou, mais precisamente, o intervalo de tempo do processo iterativo responsável pelo ajuste de pesos do neurônio k. O único sinal de saída y_k(t), do neurônio k, é comparado com uma *saída desejada*, denominada d_k(t).

Consequentemente, um sinal de erro e_k(t) é produzido:

$$e_k(t) = d_k(t) - y_k(t) \qquad \text{(B.2)}$$

B.2.2.1.1
Aprendizagem com Regularização Bayesiana

Um *perceptron* calcula a combinação linear dos dados de entrada de uma rede e os submete a uma função de ativação (linear ou não) produzindo uma saída. Um perceptron de múltiplas camadas (MLP), pode ser definido como um modelo não linear que aproxima as realizações de um processo estocástico por uma função $G: XxΨ \to \Re$ onde $X \subset \Re^n$ e Ψ é um subconjunto compacto de dimensão finita de $\Re^p$, sendo p o número de pesos da rede. Estas definições são atendidas pela especificação de uma rede com uma única camada oculta de neurônios (White & Racine, 2000):

$$y_i = G(x,y) + e_i = a_0 + \sum_{h=1}^{H} a_h F(g_0 + \sum_{i=1}^{I} g_{hi} x_i) + e_i \qquad \text{(B.3)}$$

onde $(x,y) \in XxΨ$ sendo $x = [x_1, x_2, ...x_I]$ vetores de variáveis independentes e y o vetor de parâmetros $y = [a', g']$, composto pelos vetores de pesos da camada de saída e da camada oculta respectivamente. Os parâmetros a_0 e g_0 são respectivamente o bias para a camada de saída (intercepto) e o *bias* para a camada oculta. A aplicação $F(x,y) \to \Re$ contínua para todo $x \in X$, chamada função de ativação é a função logística:

$$F(x) = (1 + e^{-x})^{-1} \qquad \text{(B.4)}$$

Os MLP são modelos não lineares que para um dado número de neurônios na camada oculta e um tamanho suficiente da amostra podem aproximar qualquer função com grau de precisão, em outras palavras, um MLP é um aproximador universal (Cybenko, 1989).

O aprendizado ou treinamento de uma rede neural tem tipicamente por objetivo reduzir a soma dos quadrados dos erros (Foresee & Hagan, 1977):

$$\hat{y} = \arg\min_y Q_1(y) = \arg\min_y \sum_{t=1}^{N} (y_t - G(x,y))^2 \qquad \text{(B.5)}$$

Assim como outros modelos flexíveis não lineares, as RNA podem sofrer de overfitting. Este problema ocorre quando utilizamos um número excessivo de neurônios na camada oculta, que levarão a uma perda da capacidade de generalização (fora da amostra). Em contrapartida, se reduzirmos o número de neurônios em excesso, teremos perda da capacidade de aproximar o processo gerador dos dados (Medeiros e Pedreira, 2001).

Atualmente, diversas metodologias são utilizadas para solucionar o problema de overfitting (Haykin, 1998). Nesta tese utilizaremos o procedimento desenvolvido por Mackay (1992) chamado de Regularização Bayesiana, que consiste em adicionar um termo de penalização (regularização) à função objetivo, de forma que o algoritmo de estimação faça com que os parâmetros irrelevantes convirjam para zero, reduzindo assim o número de parâmetros efetivos utilizados no processo.

Seguindo a notação utilizada por Medeiros e Pedreira (2001), o problema de estimação passa a ser definido como:

$$\hat{y} = \arg\min_{y} Q_T(y) = \arg\min_{y} \sum_{t=1}^{N} (\eta Q_1(y) - \phi Q_2(y))^2 \qquad \text{(B.6)}$$

onde a função de penalização é a soma do quadrado dos parâmetros:

$$Q_2(y) = \sum_{h=0}^{H} \alpha_h^2 + \sum_{h=0}^{H} \sum_{i=0}^{I} \gamma_{hi}^2 \qquad \text{(B.7)}$$

O problema de regularização é otimizar a função objetivo de forma a encontrar valores para os parâmetros ϕ e η. Este problema de otimização requer o cálculo da matriz Hessiana como pode ser visto em Mackay (1992). O algoritmo desenvolvido por Foresee e Hagan (1997) propõe a aproximação da matriz Hessiana pelo algoritmo de Levenberg-Marquardt (1963), reduzindo o custo computacional.

A aproximação é feita utilizando os seguintes passos (Foresee e Hagan, 1997)

i) Faça $\phi = 0$ e $\eta = 1$ e utilize o método de Nguyen-Widrow para inicializar os parâmetros;

ii) Faça uma estimativa (um passo) do algoritmo de Levenberg-Marquardt minimizando $Q_1(y)$;

iii) Calcule o numero efetivo de parâmetros $\hat{d} = \dim(y) - 2\hat{h}.tr(\hat{H})^{-1}$ onde $\hat{H}$ é aproximação da matriz Hessiana feita pelo algoritmo de Levenberg-Marquardt: $\hat{H} = \nabla^2 Q_T(y) \approx 2\hat{h}J^T J + 2\hat{f}I_N$ onde J é matriz jacobiana dos erros;

iv) Calcule as novas estimativas para f e h, $\hat{h} = \frac{T - \hat{d}}{2Q_1(y)}$ e $\hat{f} = \frac{\hat{d}}{2Q_2(y)}$, onde T é o numero de observações;

v) Repita os passos ii, iii e iv até a convergência.

Uma discussão detalhada do uso da Regularização Bayesiana, em combinação com o treinamento de Levenberg-Marquardt (Marquardt, 1963; Hagan, 1994), pode ser achado em Foresee & Hagan (1997).

A Regularização Bayesiana está implementada na função *trainbr* encontrada no *toolbox* do Matlab. Uma variável deste algoritmo é que ele fornece uma medida de quantos parâmetros da Rede (pesos e vieses) estão sendo, efetivamente, usados pela rede. Quando usamos a função *trainbr*, é importante deixar o algoritmo trabalhar até o efetivo número de parâmetros convergir. O treinamento pode parar com a mensagem "Máximo MU alcançado". Isto é uma indicação de que o algoritmo verdadeiramente convergiu. Você também pode saber que o algoritmo convergiu quando a soma dos quadrados dos erros (MSE) e a soma dos quadrados dos pesos (MSW) são relativamente constantes em várias repetições.

B.2.2.1.2 Método *Leave-one-out*

A essência da aprendizagem por retropopagação do erro é codificar um mapeamento de entrada-saída (representado por um conjunto de exemplos rotulados) nos pesos sinápticos e limiares de um perceptron de múltipla camadas. Espera-se que a Rede torne-se bem treinada de modo que aprenda o suficiente do passado para generalizar no futuro. Desta perspectiva, o processo de

aprendizagem se transforma em uma escolha de parametrização da Rede para esse conjunto de dados (Haykin, 1998).

Neste contexto, uma ferramenta padrão da estatística conhecida como **validação cruzada** fornece um princípio orientador atraente (Stone, 1974, 1978). Primeiramente o conjunto de dados disponível é dividido aleatoriamente em um conjunto de treinamento e em um conjunto de teste. O conjunto de treinamento é dividido adicionalmente em dois subconjuntos distintos: subconjunto de estimação (usado para selecionar o modelo) e subconjunto de validação (usado para testar ou validar o modelo).

A motivação aqui é validar o modelo com um conjunto de dados diferente daquele usado para estimar os parâmetros. Há, entretanto, uma possibilidade considerável de que o modelo assim selecionado, com os valores de parâmetros com melhor desempenho, possa acabar ajustando excessivamente o subconjunto de validação. Para evitar esta possibilidade, o desempenho de generalização do modelo selecionado é medido sobre o conjunto de teste, que é diferente do subconjunto de validação.

Existem outras variantes da validação cruzada, particularmente quando há uma escassez de exemplos rotulados. Nessa situação pode-se usar a validação cruzada múltipla dividindo o conjunto disponível de N exemplos em K subconjuntos (onde K>1). O modelo é treinado com todos os subconjuntos, exceto um, e o erro de validação é medido testando-se com este subconjunto deixado de lado no treinamento. Este procedimento é repetido para um total de K tentativas, cada vez usando um subconjunto diferente para a validação. O desempenho do modelo é avaliado pela média do erro quadrado obtido na validação sobre todas as tentativas do experimento.

Quando o número de exemplos rotulados disponíveis, N, for severamente limitado, pode-se usar a forma extrema de validação cruzada múltipla conhecida como o *"método deixe um de fora"* (*leave-one-out method*). Neste caso, N-1 exemplos são usados para treinar o modelo, e o modelo é validado testando-o sobre o exemplo deixado de fora. O experimento é repetido para um total de N vezes, cada vez deixando de fora um exemplo diferente para a validação. O erro quadrado na validação é então a média sobre as N tentativas do experimento (Haykin, 1998, Bishop, 1995).

Anexo C - Análise Discriminante

C.1
Introdução

A Análise Discriminante, como método estatístico multivariado, compõe um conjunto de técnicas destinadas a tratar problemas de classificação. Esta técnica surgiu com o objetivo de se distinguir estatisticamente entre dois ou mais grupos de indivíduos, previamente definido a partir de variáveis conhecidas para todos os membros dos grupos. Isto é, pretende-se discriminar grupos de indivíduos definidos a priori com base num critério pré-definido, a partir da informação recolhida sobre os indivíduos desses grupos.

Esta técnica de análise multivariada é empregada para descobrir as variáveis que distinguem os membros de um grupo dos de outro, de modo que, conhecida às variáveis de um novo indivíduo, se possa prever a que grupo ele pertence. Neste sentido, a Análise Discriminante tem um importante campo de aplicação em problemas de diagnóstico médico. Podem ser encontrados exemplos em Hermans & Habbema (1975), Anderson (1974), Aitchison & Dunsmore (1975), Titteringtn *et al* (1981) e Krusinska *et al* (1990), dentre outros.

Em geral, o objetivo da Análise Discriminante é encontrar a separação máxima entre os grupos através da maximização da diferença entre as médias dos grupos relativamente aos desvios padrão no interior de cada grupo. A idéia central é substituir as variáveis originais, em geral numerosas e correlacionadas, por uma combinação linear cujos valores diferenciem ao máximo os indivíduos de seus grupos. Para classificar um indivíduo, calcula-se o valor da combinação linear para seus atributos e verificam-se se este é menor ou maior que um valor limite calculado, de forma a minimizar a probabilidade de erro de classificação (Johnson & Wichern , 1998).

C.2
Relação com o Modelo de Regressão Múltipla

Do ponto de vista algébrico, o modelo de Análise Discriminante para 2 grupos é um caso especial de um modelo de regressão múltipla, onde a variável dependente assume valores discretos. Considere o modelo abaixo,

$$g_{(i)} = b^T X_{(i)} + b_0 + e_{(i)} \qquad \text{(C.1)}$$

onde $\mathbf{g_{(i)}}$ é a classe do individuo i (0 ou 1), **b** é o vetor de coeficientes da regressão**,** $\mathbf{b_0}$ é o intercepto e $\mathbf{e_{(i)}}$ é o erro do modelo. Em condições ideais, a estimação por mínimos quadrados de **b** produz um vetor na mesma direção do eixo determinado pelo método de Fisher (1936), descrito mais adiante.

C.3
Hipótese do Modelo de Análise Discriminante

Para que a regra de classificação fornecida pela Análise Discriminante seja ótima, tornando a probabilidade de erro de classificação mínima, é necessário que os dados atendam as seguintes condições:

- As variáveis explicativas ($\mathbf{X_{(i)}}$) tenham distribuição normal multivariada.
- A matriz de covariância dos grupos seja a mesma. Por outro lado, quanto mais diferentes as médias, mais fácil será discriminar os grupos.

C.4
Método de Fisher para 2 Grupos

O método de Fisher (1936) busca a combinação linear que maximiza a razão da distância ao quadrado entre as projeções dos centróides[1] dos dois grupos para a variância das projeções. Seja **b** o vetor com os pesos da combinação linear. Então, $\mathbf{y_{(i)}}$, o valor da projeção de $\mathbf{X_{(i)}}$ sobre **b**, é dado por:

$$y_{(i)} = b^T X_{(i)} \qquad \text{(C.2)}$$

[1] Centróide é um ponto do espaço p-dimensional cujas coordenadas são as médias aritméticas das variáveis discriminantes para os indivíduos pertencentes ao mesmo grupo.

Sejam $\overline{X}_1$ *e* $\overline{X}_2$ os centróides dos grupos 1 e 2, respectivamente. Então, as médias das projeções sobre **b** são dadas por:

$$\overline{y}_1 = b^T \overline{x}_1 \quad \text{e} \quad \overline{y}_2 = b^T \overline{x}_2 \tag{C.3}$$

Supondo que as matrizes de covariância de ambos os grupos são iguais e denotadas por $\mathbf{C_x}$, tem-se que a variância das projeções para qualquer dos grupos é dada por:

$$S_y^2 = b^T C_x \tag{C.4}$$

O objetivo do método é encontrar **b** que maximiza a proporção entre a diferença das médias de $\overline{y}_1$ *e* $\overline{y}_2$ ao quadrado a variância de **y**, dada por:

$$\Delta = \frac{\left(b^T(\overline{x}_1 - \overline{x}_2)\right)^2}{b^T C_x b} \tag{C.5}$$

A solução deste problema é dada por:

$$b = S^{-1}(\overline{x}_1 - \overline{x}_2) \tag{C.6}$$

onde **S** é a matriz de covariância (comum aos dois grupos) estimada por:

$$S = \frac{1}{n_1 + n_2 - 2}(x_1^T x_1 + x_2^T x_2) \tag{C.7}$$

e onde x_1 e x_2 são as matrizes de dados referentes aos grupos 1 e 2, com, respectivamente, n_1 e n_2 indivíduos.

A chamada função discriminante de Fisher é dada por:

$$Y_{(i)} = b^T X_{(i)} = (\overline{x}_1 - \overline{x}_2)^T S^{-1} X_{(i)} \tag{C.8}$$

A função discriminante de Fisher pode ser usada para construir uma regra de decisão: uma observação $\mathbf{X_{(i)}}$ será classificada no grupo cuja média esta mais próxima. Isto é, se $| y_{(i)} - \overline{y}_1 | < | y_{(i)} - \overline{y}_2 |$ então $\mathbf{X_{(i)}}$ é classificado no grupo 1.

C.5
Probabilidades a Priori e Função de Custo

É comum em situações práticas que se tenha uma probabilidade *a priori* do indivíduo pertencer a um ou outro grupo. Ao mesmo tempo, também ocorre que as conseqüências de um erro de classificação sejam diferentes segundo o tipo de erro.

A função discriminante de Fisher pode incorporar estes parâmetros. Para isso determina-se **b** que minimiza o custo esperado médio dos erros de classificação.

Se $\mathbf{p_i}$ é a probabilidade *a priori* do individuo pertencer ao grupo **i** e $\mathbf{C_{ji}}$ é o custo de classificar um individuo da classe **j** como do grupo **i**, então, a regra ótima de classificação é: se $| b^T (X_{(i)} - \bar{x}_1) | < k | b^T (X_{(i)} - \bar{x}_2) |$ então $X_{(i)}$ é classificado no grupo 1, onde $k=p_2C_{12}/p_1C_{21}$.

C.6
Método *Stepwise*

Por vezes, o investigador vê-se confrontado com um conjunto de informações, sob a forma de variáveis, superior ao necessário para se obter uma distinção satisfatória. Neste caso é possível utilizar um método discriminante *stepwise* (Drapper & Smith, 1981), o qual começa por selecionar as variáveis que mais contribuem para a distinção entre grupos, e em seguida vai incluindo e/ou retirando variáveis nas funções discriminantes, uma a uma, de acordo com um critério[2] que pode ser definido pelo próprio analista.

O método *stepwise* começa por escolher a variável que mais diferencia os grupos de acordo com o critério pré-estabelecido. A segunda variável a ser escolhida é a que, juntamente com a primeira, maximiza o aumento do critério discriminante, e assim por diante. De acordo com este processo, variáveis já escolhidas nas etapas anteriores podem ser retiradas e novas introduzidas, se tais variáveis contribuírem para um aumento do critério definido.

No passo final, ou se verifica que todas as variáveis foram selecionadas ou, então, que as que foram rejeitadas não teriam qualquer contribuição adicional para a distinção entre os grupos.

[2] Os critérios mais comuns são as estatísticas F e o λ de Wilks.

C.7
L de Wilks

A estatística Λ de Wilks é definida como a razão entre a dispersão das médias dos grupos e a variância total e é expressa por:

$$l_{(i)} = \frac{1}{\text{var}(X_{(i)})} \left[\frac{\sum_{j=1}^{k} n_j \, \text{var}(X_{j_{(i)}})}{n} \right] \qquad \text{(C.9)}$$

Por ser uma estatística inversa, a primeira variável a ser escolhida é a que produz o menor valor de Λ. É possível aproximar esta estatística a um teste para a diferença de médias entre os grupos com distribuição F. Depois de feita esta aproximação, a variável a ser introduzida é a que provocar um maior acréscimo no valor de F.

C.8
Estatística F

Uma outra medida, mais formal, da importância de uma variável para a diferenciação entre dois grupos é dada pela estatistica F-parcial. A interpretação desta estatística está ligada ao problema de determinar se o grupo a que pertence o indivíduo influência o valor da variável $\mathbf{X_{(i)}}$, descontadas as contribuições de outras variáveis.

A estatística F-parcial é calculada em função do coeficiente de determinação[3], R^2, da regressão associada ao modelo:

$$F_{p-1,n-1} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \qquad \text{(C.10)}$$

[3] É uma medida que diz o quanto o modelo de regressão da amostra se ajusta aos dados (Gujarati, 2000)

Esta estatística é o critério de entrada e saída de variáveis mais utilizado. Os valores ideais do nível de significância de entrada e saída de uma variável do modelo devem ser definidos pelo usuário[4]. Tradicionalmente, o F-parcial para incluir uma variável deve ser maior do que para excluí-la.

[4] Em geral, utiliza-se o nível de significância como 5% ou 1%.

Anexo D – Regressão Logística

A regressão logística tornou-se uma técnica padrão, sobretudo na área médica, por relacionar um conjunto de variáveis independentes a uma única variável resposta binária. Este modelo pode ser estendido quando a variável resposta qualitativa tem mais do que duas categorias; por exemplo, a pressão sanguínea pode ser classificada como alta, normal e baixa.

Em muitos estudos a variável resposta qualitativa tem duas possibilidades e, assim, pode ser representada pela variável indicadora, recebendo os valores 0 (zero) e 1 (um).

Por simplicidade, considere-se o seguinte modelo:

$$Y_i = b_0 + b_1 X_i + e_i \qquad \text{(D.1)}$$

onde Y_i assume os valores 0 ou 1. Desta relação pode-se escrever:

$$P_i = E(Y_i = 1 | X_i) = b_0 + b_1 X_i \qquad \text{(D.2)}$$

Porém tem-se uma relação linear onde, na realidade, espera-se que P_i se relacione de forma não linear com as variáveis X_i. Dessa forma precisa-se de um modelo (de probabilidade) que tenha duas características: (1) à medida que X_i aumenta, P_i também aumenta, mas nunca saia do intervalo 0-1; e (2) a relação entre P_i e X_i seja não linear e aproxime-se de 1 mais lentamente a medida que X_i fique maior.

Pode-se então representar o modelo através da seguinte expressão:

$$P_i = E(Y_i = 1 | X_i) = \frac{1}{1 + e^{-Z_i}} \qquad \text{(D.3)}$$

onde

$$Z_i = b_0 + b_i X_i \tag{D.4}$$

A equação acima é conhecida como função distribuição logística (acumulada). É fácil verificar que quando Z_i varia de -∞ a ∞ , P_i varia entre 0 e 1 e P_i não se realciona linearmente com Z_i.

Da relação D.4 Pode-se escrever:

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \tag{D.5}$$

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i}$$

Aplicando-se o logaritmo natural obtem-se a seguinte expressão:

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = b_0 + b_1 X_i \tag{D.6}$$

ou seja, L_i, o logaritmo da razão de probabilidades é não somente linear em X, mas também é linear nos parâmetros. L é chamado de Logit, daí o nome modelo logit.

Para fins de estimativa, podemos escrever (D.6) como segue:

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = b_0 + b_1 X_i + u_i \tag{D.7}$$

Pode-se demonstrar que se N_i for suficientemente grande e se cada observação em uma determinada classe X_i se distribuir independentemente como uma variável binomial, então:

$$u_i \sim N(0, \frac{1}{N_i P_i (1 - P_i)} \tag{D.8}$$

e os parâmetros são estimados pelo método da máxima verossimilhança.

Um primeiro passo para obtenção do modelo final de regressão logística consiste então, após estimar os coeficientes da equação de regressão, em verificar se cada variável é significativamente relacionada com a variável resposta do modelo. De acordo com Hosmer & Lemeshow (1989), isto é usualmente realizado através da formulação de testes de hipóteses estatísticas, que avaliam o modelo com a variável inclusa com o modelo sem a variável. Um método para tanto usa a estatística G, a qual, sob a hipótese que ßt é igual a zero, seguirá uma distribuição χ^2 (qui-quadrado) com numero de graus de liberdade igual ao número de parâmetros da equação de regressão:

$$G = -2\{L(b_t) - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]\} \quad \text{(D.9)}$$

Onde $n_1 = \sum y_t$, $n_0 = \sum (1 - y_t)$ e $n = n_0 + n_1$

Desse modo G é uma medida interna do modelo, posto que depende do número de observações n, não podendo ser comparada entre modelos de diferentes universos de dados.

Uma outra técnica usada para se estimar a significância de uma variável ao modelo é o teste de Wald, o qual é obtido pela comparação do coeficiente estimado e seu desvio padrão estimado. Segundo Hosmer e Lemeshow (1989), a estatística do teste de Wald segue uma distribuição χ^2 ..

$$W = \frac{\hat{b}_i}{\hat{SE}(\hat{b}_i)} \quad \text{(D.10)}$$

Estes testes, contudo, são meras afirmativas de evidências da significância estatística entre variáveis, devendo se considerar também outros fatores como a importância de cada variável ao fenômeno a ser modelado e a influencia conjunta de outras variáveis importantes. Portanto, a idéia básica por traz de um modelo de regressão logística consiste em minimizar o número de variáveis para que o

modelo resultante seja mais estável numericamente e facilmente generalizado, dado que quanto mais variáveis incluídas no modelo, mas ele se torna dependente dos dados. Por isso, faz-se também necessário o uso de técnicas *stepwise* na regressão logística, processo pela quais variáveis são incluídas ou excluídas do modelo, baseado somente em critérios estatísticos tais como estatística G e teste de Wald.

Por último, o modelo ajustado deve ser testado em relação à sua eficiência para descrever a variável resposta. Isto é conhecido como teste de qualidade do ajuste. De acordo com Hosmer e Lemeshow (1989), são usados para isto os testes de Pearson c^2 e o *Deviance*, os quais avaliam H_0: O modelo está bem ajustado. Nestes testes os graus de liberdade são iguais a j-(p-1), onde j é número de padrões covariados e p é o número de parâmetros do modelo. Ainda, um outro método conhecido como Homer-Lemershow, apresentado pelos mesmos autores, testa esta mesma hipótese com graus de liberdade igual a g, onde g equivale ao número de grupos que as n observações foram subdivididas. No presente caso em dois grupos.

Anexo E – Relação de Variáveis do Banco de Dados

Quadro E1: Lista de variáveis

	Nome da variável	Tipo	Descrição
1	ROWNAMES	Discreta	Identificação do Paciente
2	INSTITUI	Categórica	Instituição
3	SEXO	Categórica	Sexo
4	IDADE	Discreta	Idade
5	ANOSE	Discreta	
6	DATADENA	Data	Data de Nascimento
7	DATADEEN	Data	Data de Entrada - Início da Quimioterapia
8	PRONTUAR	Discreta	Prontuário
9	PALIDEZ	Categórica	Palidez
10	ASTENIA	Categórica	Astenia
11	EMAGRECI	Categórica	Emagrecimento
12	RASH	Categórica	
13	DOROSSEA	Categórica	Dor Óssea
14	ARTRITE	Categórica	Artrite
15	ARTRALGI	Categórica	Dor na Articulação
16	AUMENTOA	Categórica	
17	DORABDOM	Categórica	Dor Abdominal
18	DORTORAC	Categórica	Dor torácica
19	NODULOCU	Categórica	Nódulo cutâneo
20	COURO	Categórica	
21	APLASIAD	Categórica	
22	USOPREVI	Categórica	Uso prévio de Corticóide
23	FIGADO	Contínua	Fígado
24	BACO	Contínua	Baço
25	ADENOMEG	Categórica	Gânglio
26	OCCIPITA	Categórica	Gânglio

Quadro E1: Lista de variáveis (continuação)

27	MENTONIA	Categórica	Gânglio mandíbula
28	RETROAUT	Categórica	Gânglio Auricular
29	SUBMANDI	Categórica	Gânglio mandíbula
30	CERVICAL	Categórica	Gânglio
31	SUPRACLA	Categórica	Gânglio clavícula
32	AXILAR	Categórica	Gânglio axilas
33	MASSAMED	Categórica	Gânglio
34	ABDOMINA	Categórica	Gânglio
35	INGUINAL	Categórica	Gânglio na virilha
36	LESAOLIT	Categórica	Lesão dos ossos
37	ALTERACA	Categórica	
38	SANGRAME	Categórica	Sangramento
39	EQUIMOSE	Categórica	Sangramento
40	PETEQUIA	Categórica	Sangramento
41	EPISTAXE	Categórica	Sangramento
42	GENGIVAL	Categórica	Sangramento Gengiva
43	GASTROEN	Categórica	Sangramento
44	GENITOUR	Categórica	Sangramento
45	FEBRE	Categórica	Febre
46	INFECCAO	Categórica	Infecção
47	PULMONAR	Categórica	Infecção pulmonar
48	URINARIA	Categórica	Infecção urinaria
49	VIARESPI	Categórica	Infecção vias respiratórias
50	PELE	Categórica	Infecção pele
51	ORAL	Categórica	Infecção oral
52	OUVIDO	Categórica	Infecção ouvido
53	SEMSITIO	Categórica	Infecção não determinada
54	FATORDER	Continua	
55	MENORQUE	Categórica	
56	N08MENOR	Categórica	
57	N12MENOR	Categórica	
58	MAIOROUI	Categórica	

Quadro E1: Lista de variáveis (continuação)

59	HEMOGLOB	Discreta	Hemoglobina
60	MAIOROU8	Categórica	
61	L1	Categórica	Classificação FAB
62	L2	Categórica	Classificação FAB
63	CONTAGEM	Contínua	Leucócitos
64	BLASTOS	Contínua	Blastos
65	N10MENOR	Categórica	
66	MAIOROU5	Categórica	
67	LDH	Contínua	
68	ACOMETIM	Categórica	
69	CD19	Categórica	
70	CD20	Categórica	
71	CD10	Categórica	
72	CD21	Categórica	
73	CD22	Categórica	
74	DR	Categórica	
75	CYIG	Categórica	
76	CD3	Categórica	
77	CD7	Categórica	
78	CD5	Categórica	
79	CD11	Categórica	
80	CD34	Categórica	
81	CD33	Categórica	
82	CD14	Categórica	
83	CD4	Categórica	
84	CD8	Categórica	
85	CD1	Categórica	
86	CD2	Categórica	
87	CD301	Categórica	
88	SIG	Categórica	
89	K	Categórica	
90	CD15	Categórica	

Quadro E1: Lista de variáveis (continuação)

91	CD13	Categórica	
92	CD3DC	Categórica	
93	CD38	Categórica	
94	L	Categórica	
95	PREPREBL	Categórica	
96	CLLA	Categórica	
97	PREBLLA	Categórica	
98	TLLA	Categórica	
99	LINDAGUD	Categórica	
100	LINHAGEM	Categórica	
101	NAODETER	Categórica	
102	ANTIGENO	Categórica	
103	POSITIVO	Categórica	
04	NEGATIVO	Categórica	
05	NAOEXAMI	Categórica	
06	T922	Categórica	Translocação cromossomial
07	T411	Categórica	Translocação cromossomial
08	T119	Categórica	Translocação cromossomial
09	OUTROS	Categórica	
10	NORMAL	Categórica	
11	HIPOPLOI	Categórica	
12	HIPERDIP	Categórica	
113	N4750	Categórica	
114	MAIOR50	Categórica	
115	TRIPLOID	Categórica	
116	TETRAPLO	Categórica	
117	PAS	Categórica	
118	FA	Categórica	
119	SUDAN	Categórica	
120	SINDROME	Categórica	Síndrome de Down
121	MAIOROU1	Categórica	mais de 1000 blastos no 8o. dia
122	PROTOCOL	Categórica	Protocolo

Quadro E1: Lista de variáveis (continuação)

123	SRG	Categórica	Risco Padrão
124	MRG	Categórica	Risco intermediário
125	HRG	Categórica	Alto Risco
126	MORTEPRE	Categórica	Morte prematura
127	DATAMP	Data	data da Morte Prematura
128	FALHADER	Categórica	Falha de Remissão
129	DATAFR	Data	Data da Falha de Remissão
130	ABANDONO	Categórica	Abandono
131	DATAABD	Data	Data do Abandono
132	REMISSAO	Categórica	Remissão
133	MEDULANO	Categórica	
134	MORTEEMR	Categórica	Morte em Remissão
135	DATAMRC	Data	Data da Morte em remissão
36	MALIGNID	Categórica	
37	DATAMS	Data	
38	RECAIDA	Categórica	Recaída
39	PRECOCE	Categórica	Precoce
40	TARDIA	Categórica	tardia
41	DATARC	Data	Data da Recaída
42	SNC	Categórica	Sistema Nervoso Central
43	MO	Categórica	Medula Óssea
44	TESTICUL	Categórica	Testículos
145	OUTROS01	Categórica	
146	RECAIDAC	Categórica	
147	SNCMO	Categórica	
148	MOTESTIC	Categórica	
149	OUTRAS01	Categórica	
150	DOENCAAT	Categórica	
151	MORTE	Categórica	Morte
152	DATAMORT	Data	Data da Morte
153	PERDADEF	Categórica	Perda definitiva
154	DATAPERD	Data	Data da perda definitiva

Quadro E1: Lista de variáveis (continuação)

155	TRANSFER	Categórica	Transferido
156	DATATRAN	Data	Data da transferência
157	PACREC	Categórica	Paciente recaída
158	DAREC	Data	Data da recaída
159	PACCENS	Categórica	Paciente censurado
160	DACENS	Data	Data da censura
161	PROTOC	Categórica	Protocolo
162	IDA1	Contínua	Idade
163	IDA2	Categórica	Idade
164	LEUC50	Categórica	Leucócitos
165	RAC	Categórica	Raça
166	LDH1	Categórica	
167	PLAQ	Categórica	Plaquetas
68	RSICO	Categórica	Risco
69	T	Categórica	Imunofenotipo t
70	INTIDAT	Categórica	
71	DOR	Categórica	Dor
72	PALIDO	Categórica	Pálido
73	FEVER	Categórica	
74	ADENO	Categórica	
75	SPLENO	Categórica	
76	HEPATO	Categórica	
177	LEUCOMET	Categórica	Leucometria
178	PLAQUET	Categórica	Plaqueta
179	SOB	Contínua	
180	FALHA	Categórica	Falha
181	MOR	Categórica	Morte
182	IDACAT	Categórica	Idade
183	WBCCAT	Categórica	
184	BLAST	Contínua	Blastos
185	FR	Contínua	Fator de Risco
186	FRCAT	Categórica	Fator de Risco

Quadro E1: Lista de variáveis (continuação)

187	NCI	Categórica	
188	H	Categórica	
189	HL	Categórica	