José Leonardo Ribeiro Macrini

Estimação do Risco de Recidiva em Crianças Portadoras de Leucemia Linfoblástica Aguda Usando Redes Neurais

TESE DE DOUTORADO

Departamento de Engenharia Elétrica

Programa de Pós-Graduação

em Engenharia Elétrica

Rio de Janeiro

Outubro de 2004

José Leonardo Ribeiro Macrini

Estimação do Risco de Recidiva em Crianças Portadoras de Leucemia Linfoblástica Aguda Usando Redes Neurais

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Carlos Eduardo Pedreira

Rio de Janeiro, Outubro de 2004

José Leonardo Ribeiro Macrini

Estimação do Risco de Recidiva em Crianças Portadoras de Leucemia Linfoblástica Aguda Usando Redes Neurais

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Dr. Carlos Eduardo Pedreira
Orientador
Departamento de Engenharia Elétrica - PUC-Rio

Dr. Marcelo Cunha Medeiros
Departamento de Economia - PUC-Rio

Dr. Nelson Spector
UFRJ – Faculdade de Medicina – Clinica Médica

Dr. Antonio Fernando Catelli Infantosi
UFRJ – Programa de Engenharia Biomédica

Dr. Marcelo Gerardin Poirot Land
UFRJ – Faculdade de Medicina - Pediatria

Dr. Eduardo Parente Ribeiro
UFPR – Departamento de Engenharia Elétrica

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 19 de Outubro de 2004

José Leonardo Ribeiro Macrini

Nascido no Rio de Janeiro-RJ em 1954. Graduou-se em Estatística (1979) pela Escola Nacional de Ciências Estatísticas, ENCE. Mestre em Eng. Elétrica em Teoria de Controle e Estatística (2000) pela Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio. Suas pesquisas de interesse incluem as áreas de análise estatística multivariada e sistemas inteligentes aplicados à classificação de padrões.

Ficha Catalográfica

Macrini, José Leonardo Ribeiro

Estimação do risco de recidiva em crianças portadoras de leucemia linfoblástica aguda usando redes neurais / José Leonardo Ribeiro Macrini ; orientador: Carlos Eduardo Pedreira. – Rio de Janeiro : PUC, Departamento de Engenharia Elétrica, 2004.

98 f. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui referências bibliográficas.

CDD: 621.3

Agradecimentos

Trata-se de uma tarefa difícil relacionar pessoas e instituições que contribuíram para a elaboração deste trabalho, sem correr o risco de cometer alguma injustiça. Porém, como não se vive sem “correr riscos”, vou me arriscar a fazê-lo.

Ao programa de Doutorado de Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio pela confiança em mim depositada.

Ao Conselho de Desenvolvimento Cientifico e Tecnológico – CNPq, pelo auxilio financeiro concedido.

Ao Professor Dr. Carlos Eduardo Pedreira, meu orientador, pela paciência, pelo compartilhar de seu conhecimento e, sobretudo, pela confiança, respeito pessoal e profissional e pela amizade que construímos ao longo desta caminhada.

Aos Professores integrantes da Banca examinadora, pela contribuição critica fundamental ao enriquecimento do trabalho.

Ao Professor Marcelo Land pela disponibilidade do Banco de Dados sem a qual esta tese não teria sido possível.

As Professoras Alice Azevedo e Cristiana Solza pelos casos adicionais incorporados ao Banco de Dados.

A Professora e amiga Elaine Sobral da Costa pela imensa paciência em rever o banco de dados utilizado nesse trabalho em diversas ocasiões e pelas contribuições teóricas ao entendimento correto do problema.

Na hora em que mais necessitei de auxilio, um grande amigo não me faltou. Agradeço ao Alexandre Zanini a inestimável ajuda, principalmente por ocasião da versão final do trabalho de tese.

As minhas filhas Camila e Carina meu profundo agradecimento pela paciência e por compartilhar seu computador e nas várias leituras das versões preliminares do trabalho.

Por fim, agradeço do fundo do meu coração, a minha esposa Márcia, pelo incentivo, ajuda, presença e paciência sempre constante em todas as etapas do trabalho e fora delas.

Resumo

Macrini, L. **Estimação do risco de recidiva em crianças portadoras de leucemia linfoblástica aguda usando redes neurais.** Rio de Janeiro, 2004. 98p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio.

Esta tese propõe uma metodologia, baseada em procedimentos quantitativos, para estimação do risco de evento adverso (recaída ou morte) em crianças portadoras de Leucemia Linfoblástica Aguda (LLA). A metodologia proposta foi implementada e analisada utilizando dados de grupo de crianças diagnosticadas no Setor de Hematologia do Instituto de Puericultura e Pediatria Martagão Gesteira (IPPMG) da UFRJ e no Serviço de Hematologia Hospital Universitário Pedro Ernesto (HUPE) da UERJ que constituem uma considerável parcela dos casos de LLA na infância registrados no Rio de Janeiro nos últimos anos.

A estimação do risco de recaída foi realizada através de um modelo de Redes Neurais após uma seqüência de procedimentos de pré-tratamento de variáveis e de refinamentos do método no que concerne a saída alvo da rede.

O tratamento das variáveis é fundamental uma vez que o número reduzido de amostras é uma característica intrínseca deste problema. Embora a LLA seja o câncer mais freqüente a infância, sua incidência é de aproximadamente 1 caso por 100 mil habitantes por ano.

Os resultados encontrados foram satisfatórios obtendo-se um percentual de acerto de 93% (fora da amostra) para os pacientes que recaíram quando comparados com o método classicamente utilizado na clínica médica para a avaliação do risco de recidiva (método do grupo BFM). Espera-se que os resultados obtidos possam vir a dar subsídios às condutas médicas em relação à estimativa do risco de recidiva dos pacientes, portanto, podendo vir a ser útil na modulação da intensidade da terapêutica.

Palavras-chave

Informação Mútua; Seleção de Variáveis; Redes Neurais; Classificação; Leucemia Linfoblástica Aguda; Risco.

Abstract

Macrini, L. **Relapse Risk Estimation in Children with Acute Lymphoblastic Leukemia by using Neural Networks.** Rio de Janeiro, 2004. 98p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, PUC-Rio.

In this it is proposed a methodology, based on quantitative procedure, to estimate the adverse event risk (relapse or death) in Acute Lymphoblastic Leukemia (ALL) in children. This methodology was implemented and analyzed in a dataset composed by children diagnosed and treated at the hematology service of the "Instituto de Puericultura e Pediatria Martagão Gesteira (IPPMG)" in the Federal University of Rio de Janeiro and of the "Hospital Universitário Pedro Ernesto (HUPE)" in the University of state of Rio de Janeiro. This group constitutes a considerable fraction of the ALL cases in childhood registered in the last few years in Rio de Janeiro.

The relapse risk was estimated by a Neural Networks model after a sequence of variable pre-treatment procedures. This treatment has a fundamental importance due to the small number of cases (an intrinsic characteristic of this problem). Although, the ALL is the most frequent cancer in childhood, it incidence is approximately just 1 case for 100 000 inhabitants by year.

The obtained results may be considered excellent when compared with the classical risk estimative method used in the medical clinics (BFM risk). A perceptual of successes of 93% (out-of-sample) in no-relapse patients was achieved. We expect that the obtained results may subsidize medical conduct concerning the risk of adverse event and so it could be useful in the treatment intensity modulation.

Keywords

Mutual information; Selection of Variables; Neural Networks; Classification; Acute Lymphocytic Leukemia; Risk.

Sumário

1. Introdução 13
1.1. Tratamento Direcionado ao Risco de Recidiva na Leucemia Linfoblástica Aguda na Infância 13
1.2. Objetivo 16
1.3. Contribuições da Tese 16
1.4. Organização da Tese 17

2. Uma Metodologia para Classificação do Risco de Eventos Adversos em LLA 19
2.1. Descrição e Tratamento da Base de Dados 19
2.1.1. Descrição da Base de Dados 19
2.1.2. Tratamento da Base de Dados 21
2.2. Estimativa do Risco de Evento Adverso 25
2.2.1. Estimativa pelo Critério BFM95 26
2.2.2. Uma Proposta de Modelagem para a Estimativa de Risco 27
2.2.2.1. Um Refinamento no Modelo: Inversão Dirigida da Saída-Alvo da Rede 30
2.2.2.2. Um Segundo Refinamento: Considerando o Acompanhamento do Tempo de Tratamento Decorrido 32

3 Resultados e Análises 33
3.1. Medidas de Performance 33
3.2. Experimentos e Análises Relativos a Otimização do Número de Variáveis 34
3.3. Experimentos e Análises Relativos a Modelagem para a Estimativa de Risco com a Inversão da Saída-Alvo da Rede 36
3.4. Experimentos e Análise Relativos a Modelagem para a Estimativa de Risco com a Inversão da Saída-Alvo da Rede Considerando o Tempo de Tratamento 42
3.5. Comparação Entre o Risco Estimado pelo Modelo e pelo Critério BFM95 47
3.6. Comparação com Metodologia Linear 48
3.6.1. Análise Discriminante 48
3.6.2. Regressão Logística 51

4 Considerações Finais 53

Referências Bibliográficas 56

Anexo A - Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U) 61
A.1 Introdução 61
A.2 Entropia e Informação Mútua 62
A.3 Algoritmo de Seleção de Variáveis 64
A.3.1 O Problema de FRn - k 64
A.3.2 Seleção de Variáveis sob Informação Mútua (MIFS) 65
A.3.3 Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U) 66

Anexo B – Redes Neurais Artificiais 71
B.1 Introdução 71
B.2 Variáveis Principais 72
B.2.1 Arquitetura 74
B.2.1.1 Redes *Feedforward* de Uma Única Camada 75
B.2.1.2 Redes *Feedforward* de Múltiplas Camadas 75
B.2.2 Métodos de Estimação 76
B.2.2.1 Aprendizagem Supervisionada 77
B.2.2.1.1 Aprendizagem com Regularização Bayesiana 78
B.2.2.1.2 Método *Leave-one-out* 80

Anexo C - Análise Discriminante 82
C.1 Introdução 82
C.2 Relação com o Modelo de Regressão Múltipla 83
C.3 Hipótese do Modelo de Análise Discriminante 83
C.4 Método de Fisher para 2 Grupos 83
C.5 Probabilidades a Priori e Função de Custo 84
C.6 Método *Stepwise* 85
C.7 Λ de Wilks 86
C.8 Estatística F 86

Anexo D – Regressão Logística 88

Anexo E – Relação de Variáveis do Banco de Dados 92

Lista de figuras

Figura 01. Representação do Desfecho em Função do Risco BFM95 26

Figura 02. Representação da Inversão da Saída-Alvo da Rede 322

Figura A1 – Relação entre Variáveis de Entrada e Classe de Saída 66

Figura B1. Redes Neurais tipo *feedforward* com uma Única Camada de Unidades Processadoras. (a) Arquitetura. (b) Sentido de Propagação do Sinal Funcional. 75

Figura B2. Redes Neurais tipo *feedforward* com Múltiplas Camadas. (a) Arquitetura. (b) Sentido de Propagação do Sinal Funcional e do Sinal de Erro. 76

Figura B3. Diagrama de Blocos do Processo de Aprendizagem Supervisionada. 77

Lista de quadros

Quadro 01 Classificação Prognostica do National Cancer Institute para Leucemia Linfoblástica Aguda 15

Quadro 02 Classificação Prognostica do Protocolo BFM95 para Leucemia Linfoblástica Aguda (LLA) 15

Quadro 03 Resultado Preliminar de Seleção e Ordenação de Variáveis 23

Quadro 04 Resultado de Seleção e Ordenação de Variáveis Após as Três Etapas do Tratamento 25

Quadro 05: Cruzamento Desfecho *versus* Risco – Avaliação pelo Critério BFM95 27

Quadro 06: Inversão da Saída Alvo da Rede 31

Quadro 07: Experimentos Relativos a Otimização do Numero de Variáveis e Percentual de Acerto de Recaída 35

Quadro 08: Resultados da Análise com a Inversão da Saída-alvo da Rede e Medidas de Performance 37

Quadro 09: Histogramas das Estimativas de Risco do Modelo (5 variáveis) 38

Quadro 10: Histogramas das Estimativas de Risco do Modelo (6 variáveis) 39

Quadro 11: Histogramas das Estimativas de Risco do Modelo (7 variáveis) 40

Quadro 12: Histogramas das Estimativas de Risco do Modelo (todas as variáveis) 41

Quadro 13: Resultados da Inversão da Saída-Alvo da Rede Considerando o Fator Tempo de Tratamento e Medidas de Performance 43

Quadro 14: Histogramas das Estimativas de Risco do Modelo (5 variáveis) 44

Quadro 15: Histogramas das Estimativas de Risco do Modelo (6 variáveis) 45

Quadro 16: Histogramas das Estimativas de Risco do Modelo (7 variáveis) 45

Quadro 17: Histogramas das Estimativas de Risco do Modelo (todas as variáveis) 46

Quadro 18: Cruzamento Desfecho Versus Risco Estimado pelo Modelo 47

Quadro 19: Cruzamento Entre a Estimativa de Risco do Modelo Proposto Versus BFM95 47

Quadro 20: Resultados da Análise de Inversão da Saída-Alvo Considerando o Fator Tempo de Tratamento na Análise Discriminante e Medidas de Performance 49

Quadro 21: Histogramas das Estimativas de Risco da Análise Discriminante (6 variáveis) 50

Quadro 22: Resultados da Análise de Inversão da Saída-Alvo Considerando o Fator Tempo de Tratamento na Regressão Logística e Medidas de Performance 51

Quadro 23: Histogramas das Estimativas de Risco da Regressão Logística (6 variáveis) 52

Quadro A1: Valor da Informação Mútua com o Desfecho 68

Quadro A2: Valor da Entropia 69

Quadro A3: Valor da Informação Mútua entre Variáveis 70

Quadro E1: Lista de variáveis 92