

Análise Estatística no Diagnóstico de Doenças Cardíacas

Paulo Patrício da Silva¹, Kathya Silvia Collazos Linares², Cristian Mara M. M. Patrício³

Agência Estadual de Regulação de Serviços Públicos de Mato Grosso do Sul (AGEPAN), Rua Barão do Rio Branco 1092 - CEP: 79002-175 - Campo Grande – MS – Brasil

¹ paulosilva@agepan.ms.gov.br

Laboratório de Inteligência Artificial, Eletrônica de Potência e Eletrônica Digital - DEL/CCET/UFMS Cidade Universitária - CEP 79070-900 - Campo Grande MS - Brasil

^{1, 2, 3} katy@batlab.ufms.br cristian@batlab.ufms.br

Resumo

Na área da saúde, precisa-se de métodos matemáticos que permitam afirmar ou negar, a partir de dados clínicos, a influência ou não de determinado sintoma no desenvolvimento ou aparecimento de uma doença. O Teste de hipótese é um método estatístico que permite avaliar amostras ou comparar amostras de populações independentes. Neste trabalho apresenta-se a análise de um banco de dados sobre diagnóstico de doenças do coração utilizando o teste de hipótese visando determinar quais sintomas influenciam no percentual de estreitamento da artéria principal. Para tal utiliza-se a base de dados de doenças do coração disponibilizados pelo repositório de dados da UCI *Machine Learning Repository*.

Palavras-chave: Análise estatística, Teste de hipótese, Doenças do coração.

Abstract

In health area are needed mathematical methods that allow to affirm or to deny, starting from clinical data, the influence or not of certain symptom in the development or emergence of a disease. The hypothesis Test is a statistical method that allows to evaluate samples or to compare samples of independent populations. In this work is presented the analysis of a database on heart diseases diagnosis using the hypothesis test seeking to determine which symptoms influence in the percentage of artery narrowing. For that, the database of heart diseases, available in the repository of data of UCI Machine Learning Repository, is used.

Key-words: Statistical analysis, Hypothesis test, Hearth disease.

1 Introdução

A teoria epidemiológica tem de levar em consideração todos os fatores de terminantes de condições que possibilitem a ocorrência dos fenômenos que dizem respeito à saúde coletiva. E tal realidade, tanto pode ser objetiva como subjetiva, na realidade deverá ser a combinação dos dois conhecimentos. Assim sendo, a metodologia de estudo obedecerá a orientação que tem sido ditada pela experiência, propiciando respostas úteis para a solução de importantes problemas de saúde pública [5]. A epidemiologia tem dois ramos a descritiva e a analítica, esta última engloba os estudos de associação entre fatos observáveis, mediata ou imediatamente relacionáveis as doenças, e estas, vistas

sob o ponto de vista de sua incidência sobre grupos populacionais. Seu objetivo maior, que preside e dá sentido ao pensamento desenvolvido sob sua cobertura, é o conhecimento das causas determinantes das doenças, que afetam grupos significativos de pessoas [10, 11].

Estudos epidemiológicos de doenças cardíacas dizem respeito a fatores prognósticos e sobrevida [3, 6, 14] entre outros. Entretanto, a análise de dados de pacientes diagnosticados ainda não é definitivo, uma das dificuldades reside na diversidade genética. Há populações onde uma regra é válida e outras onde não o é. Um exemplo disto é a icterícia neonatal, quando o sintoma apresenta-se conjuntamente com escurecimento da urina e clareamento das fezes indica uma colestase neonatal. Sendo necessário um diagnóstico precoce para evitar o desenvolvimento de cirrose e insuficiência

hepática [13]. Entretanto na raça asiática a icterícia neonatal é freqüente e não necessariamente esta associada à colestase neonatal, enquanto que em outras raças somente o fato de apresentar icterícia já seria um indicador suficiente para o diagnóstico precoce.

O Teste de hipótese é um método estatístico que permite provar se uma afirmação dada é válida ou não dentro de um conjunto de dados. Durante a avaliação clínica de um paciente coletam-se dados que permitem ao clínico chegar a um diagnóstico. Entretanto muitas vezes, não é uma tarefa simples determinar as similaridades entre pacientes com um mesmo diagnóstico. A análise estatística de um conjunto de pacientes pode permitir avaliar quais atributos poderiam ser os determinantes em um diagnóstico clínico [7].

O presente artigo realiza uma avaliação estatística dos dados clínicos disponibilizados pela UCI *Machine Learning Repository* [12]. Tais dados foram coletados e organizados no Instituto Húngaro de Cardiologia, Budapeste, responsável Andras Janosi; Hospital Universitário de Zurique, Suíça, responsável William Steinbrunn, M.D.; Hospital Universitário de Basileia, Suíça, responsável Matthias Pfisterer, M.D.; V.A. Medical Center, Long Beach, Califórnia e Fundação Clínica de Cleveland, responsável Robert Detrano, M.D. eles publicaram o uso de um modelo de função discriminante para calcular probabilidades de doença coronária angiográfica tendo sucesso no diagnóstico dos pacientes com síndromes de dor de tórax e prevalência de doença intermediária [4]. Trabalhos, abordando a eficiência de algoritmos de aprendizado na presença de atributos irrelevantes também utilizaram estas bases de dados [1].

A determinação da relevância de atributos em uma base de dados pode ser realizada utilizando métodos estatísticos. Neste artigo utiliza-se o Teste F para determinar variâncias iguais e posteriormente aplica-se o teste t-Student para determinar se as médias das amostras são iguais. Comparam-se amostras do mesmo banco de dados, caracterizadas por ter estreitamento maior ou menor do que o 50% da artéria principal, com o objetivo de determinar quais dos atributos influenciam nessa característica.

2 Fundamentação Teórica

A análise de uma população é feita através do estudo de uma amostra da população devido muitas vezes à complexidade ou impossibilidade de se analisar a população completa. A análise feita utilizando métodos estatísticos inicia na amostragem da população, para posteriormente realizar o cálculo das estatísticas e inferência de parâmetros. Muitas vezes é necessário decidir se uma afirmação feita acerca de um parâmetro deve-se aceitar ou rejeitar. Esta afirmação é denominada de hipótese, e, o procedimento para a tomada de decisão sobre a hipótese é denominado Teste de hipótese. Este é um dos aspectos mais úteis da inferência estatística, dado que em diversas áreas e em particular na área da

saúde podem ser formulados alguns problemas como testes de hipóteses [2, 8, 9].

Quando se deseja realizar inferência estatística para duas populações independentes, a situação geral é mostrada na Figura 1. Onde, a população 1 tem média μ_1 e variância σ_1^2 , enquanto que a população 2 tem média μ_2 e variância σ_2^2 . Inferências serão baseadas em duas amostras de tamanhos n_1 e n_2 respectivamente. Ou seja, $x_{11}, x_{12}, \dots, x_{1n}$ é uma amostra aleatória de n_1 observações provenientes da população 1 e $x_{21}, x_{22}, \dots, x_{2n}$ é uma amostra aleatória de n_2 observações provenientes da população 2. Os testes de hipóteses que podem ser usados para analisar os parâmetros são os testes: t-Student e o teste F. Nesses dois testes devem se ter como suposições que:

- $x_{11}, x_{12}, \dots, x_{1n}$, é uma amostra aleatória proveniente da população 1;
- $x_{21}, x_{22}, \dots, x_{2n}$, é uma amostra aleatória proveniente da população 2;
- As duas populações representadas por x_1 e x_2 são independentes;
- Ambas as populações são normais ou se elas não forem normais, as condições do teorema central do limite se aplicam.

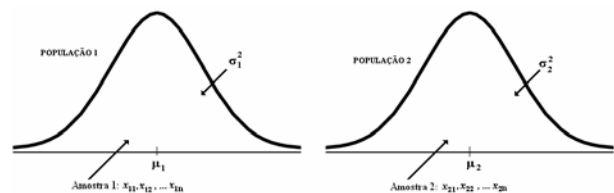


Figura 1 - Duas populações independentes

2.1 Teste de hipótese para a diferença nas médias com variâncias desconhecidas

Variâncias iguais, $\sigma_1^2 = \sigma_2^2 = \sigma^2$: o teste t-Student é utilizado para testar a hipótese da diferença entre as médias de duas populações. Supondo que se têm duas populações, com médias conhecidas μ_1 e μ_2 , e variâncias desconhecidas e iguais, tem-se:

Hipótese nula: $H_0 : \mu_1 = \mu_2$,

Hipótese alternativa: $H_1 : \mu_1 \neq \mu_2$

Teste estatístico:

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{S_a \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

Onde:

n_1 = tamanho da amostra 1;

n_2 = tamanho da amostra 2;

\bar{X}_1 = média da amostra 1;

\bar{X}_2 = média da amostra 2;

S_1^2 = variância da amostra 1;

S_2^2 = variância da amostra 2;

S_a = variância agregada das duas amostras, calculada por:

$$S_a^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (2)$$

O grau de liberdade esta dado por: $gl = n_1 + n_2 - 2$

O critério de rejeição, para o teste bilateral, esta dado por: $t_0 > t_{\alpha/2, n_1 + n_2 - 2}$ ou $t_0 < -t_{\alpha/2, n_1 + n_2 - 2}$, como mostrado na Figura 2. Se aceita a hipótese H_0 quando a área encontrada no teste for maior que $\alpha/2$, ocorrendo o contrário a evidência estatística aponta para a hipótese H_1 .

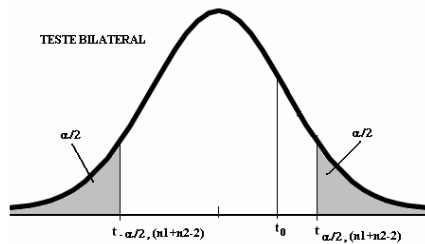


Figura 2 - Distribuição para o teste t-Student Bilateral

Variâncias diferentes, $\sigma_1^2 \neq \sigma_2^2$: quando não podemos assumir que as variâncias desconhecidas são iguais para testar a hipótese $H_0 : \mu_1 = \mu_2$, utiliza-se o seguinte teste estatístico:

$$t_0^* = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \quad (3)$$

O número de graus de liberdade dado por:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \quad (4)$$

Seguem-se os mesmos critérios de rejeição e aceitação da hipótese nula H_0 que no caso anterior, exceto pelo fato de que t_0^* é usada como estatística de teste e $n_1 + n_2 - 2$ é substituído por ν na determinação do grau de liberdade para o teste.

2.2 Teste de hipótese para a razão de duas variâncias

O procedimento para teste de hipótese da igualdade de duas variâncias baseia-se no seguinte resultado: Seja $x_{11}, x_{12}, \dots, x_{1n}$ uma amostra aleatória de uma população normal com média μ_1 e variância σ_1^2 , e seja $x_{21}, x_{22}, \dots, x_{2n}$, uma amostra aleatória de uma segunda

população normal com média μ_2 e variância σ_2^2 . Supõe-se que ambas as populações são independentes. Sejam S_1^2 e S_2^2 as variâncias das amostras. Então a razão:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \quad (5)$$

Têm uma distribuição F, com $n_1 - 1$ graus de liberdade no numerador e $n_2 - 1$ graus de liberdade no denominador.

Esse resultado é baseado no fato de que $(n_1 - 1)S_1^2/\sigma_1^2$ é uma variável aleatória qui-quadrado com $n_1 - 1$ graus de liberdade, que $(n_2 - 1)S_2^2/\sigma_2^2$ é uma variável aleatória qui-quadrado com $n_2 - 1$ graus de liberdade e que as duas populações normais sejam independentes. Isto sujeita à hipótese nula de $H_0 : \sigma_1^2 = \sigma_2^2$ à razão $F_0 = S_1^2/S_2^2$ com uma distribuição F_{n_1-1/n_2-1} . O procedimento de teste é como segue:

Hipótese nula: $H_0 : \sigma_1^2 = \sigma_2^2$

Hipótese alternativa: $H_1 : \sigma_1^2 \neq \sigma_2^2$

Teste estatístico:

$$f_0 = \frac{S_1^2}{S_2^2} \quad (6)$$

Onde:

S_1^2 = variância da amostra de n_1 elementos;

S_2^2 = variância da amostra n_2 elementos;

O grau de liberdade esta dado por:

$u = n_1 - 1$ = grau de liberdade no numerador;

$v = n_2 - 1$ = grau de liberdade no denominador.

O critério de rejeição para o teste bilateral esta dado por:

$f_0 > f_{\alpha/2, n_1-1, n_2-1}$ ou $f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$.

Onde:

$$f_1 = f_{1-\alpha/2, u, v} = \frac{1}{f_{\alpha/2, v, u}} \quad (7)$$

$$f_2 = f_{\alpha/2, u, v} \quad (8)$$

São os pontos percentuais $\alpha/2$ superior e inferior, respectivamente, da distribuição F, como mostrado na Figura 3.

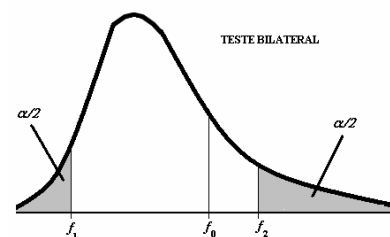


Figura 3 - Distribuição F para o teste bilateral

3 Metodologia

Os bancos de dados utilizados neste estudo foram: da Fundação Clínica de Cleveland com 303 pacientes, do Instituto Húngaro de Cardiologia, Budapeste com 294 pacientes e do V. A. Medical Center, Long Beach, Califórnia com 200 pacientes [UCI]. Cada arquivo contém 76 atributos, mas apenas 14 foram utilizados nas principais pesquisas publicadas. O atributo considerado para a decisão é o atributo Diagnóstico que se refere à presença da doença de coração no paciente. Este dado é um campo de valor inteiro que assume valores no intervalo de [0-4]. As análises completas dos pacientes levaram os médicos a distinguir, para cada um deles a presença de doença em níveis 1, 2, 3 e 4 da ausência valor 0. A seguir detalham-se a descrição completa dos 14 atributos mais citados nas publicações:

- Idade em anos;
- Sexo: 0 = fêmea, 1 = macho;
- Tipo de doença: 1= angina típica, 2= angina atípica, 3 = sem angina, 4 = assintomático;
- Pressão do Sangue em repouso (mmHg na admissão ao hospital);
- Colesterol-S (mg/dl);
- Açúcar no sangue: > (mg/dl) (1=verdadeiro; 0=falso);
- Resultado do eletrocardiograma em repouso: 0 =normal; 1 =anomalia da onda do segmento ST-T (inversão da onda T e/ou elevação do segmento ST ou depressão maior que 0,05 mV); 2 = mostrando hipertrofia ventricular esquerda (provável ou definitiva) através do critério ESTES;
- Taxa máxima de batimentos cardíacos alcançada;
- Angina induzida pelo exercício: 0 = não, 1 = sim;
- ST depressão, induzida pelo exercício em relação ao repouso;
- Inclinação da rampa no exercício ST: 1 = inclinado para cima, 2 = sem inclinação, 3 = inclinado para baixo;
- Número de artérias principais (0-3) coloridas pela fluoroscopia;
- Status do coração: 3= normal; 6 = problema permanente; 7 = problema reversível;
- Diagnóstico da doença de coração (*angiographic disease status*): 0=estreitamento < 50%, 1 = estreitamento > 50%.

A Tabela 1 apresenta o resumo para diagnóstico dos pacientes avaliados para cada Banco de Dados. Para este estudo foram escolhidos quatro atributos: Idade; Pressão Sangüínea em Repouso; Colesterol-S e Taxa Máxima de Batimentos Cardíacos. Optou-se por descartar o banco de dados Suíço por ter apenas 08 pacientes com estreitamento menor que 50% e o banco de dados de Long Beach por ter apenas 23 pacientes com estreitamento < 50% contra 79 > 50%. Também foram descartadas dos outros bancos de dados, as linhas de arquivo com falta de dados ou com valor igual à zero para algum dos atributos escolhidos. Estabeleceu-se

assim para cada banco, duas amostras de pacientes: uma com maior probabilidade de ataque cardíaco e outra com menor probabilidade.

Tabela 1 – Diagnóstico dos pacientes após completa avaliação médica

Banco de Dados	Diagnóstico dos pacientes com estreitamento da artéria principal					Total
	0	1	2	3	4	
Cleveland	164	55	36	35	13	303
Húngaro	188	37	26	28	15	294
Suíço	8	48	32	30	5	123
Long Beach	51	56	41	42	10	200

A tabela 2 mostra a nova configuração dos dados.

Tabela 2 – Configuração das Bases após a Limpeza de dados

Banco de Dados	Diagnóstico dos Pacientes com Estreitamento da Artéria Principal		Total
	< 50%	> 50%	
Cleveland	164	139	303
Húngaro	169	101	270

4 Resultados

Para o banco de dados Cleveland calcularam-se as médias e dispersões dos atributos de análise, como mostrado na Tabela 3.

Tabela 3 – Média e dispersão para o banco de dados Cleveland

Atributo	Diagnóstico dos Pacientes com Estreitamento da Artéria Principal	
	< 50%	> 50%
Idade	52.59 ± 9.51	56.63 ± 7.94
Pressão do Sangue	129.25 ± 16.20	134.57 ± 18.77
Colesterol-S	242.64 ± 53.45	251.47 ± 49.49
Frequência Cardíaca	158.38 ± 19.20	139.26 ± 22.59

Logo, foi aplicado o teste F com nível de significância de 5% ($\alpha = 0,05$) testando a variância das duas amostras, utilizaram-se as equações (6), (7) e (8), os resultados mostram-se na Tabela 4.

Tabela 4 – Teste F nos dados de Cleveland

	Idade	Pressão do Sangue	Colesterol-S	Frequência Cardíaca
f_0	1.4357	1.3415	1.1669	1.3848
P	0.0144	0.0356	0.1751	0.0228

Após a confirmação da hipótese nula para as variâncias ($H_0 : \sigma_1^2 = \sigma_2^2$); aplicou-se o teste t-Student

com nível de significância de 5% para testar a diferença entre as médias ($H_0: \mu_1 = \mu_2$). Para Pressão do Sangue e Colesterol-S, considerou-se variâncias iguais ($\sigma_1^2 = \sigma_2^2$) ver Tabela 5.

Tabela 5 – Teste t-Student para variâncias iguais

	Pressão do Sangue	Colesterol-S
t_0	-2.6470	-1.4829
P	0.0085	0.1391

Para Idade e Frequência cardíaca, as evidências estatísticas rejeitaram a hipótese nula, portanto, utilizou-se o teste t-Student para variâncias diferentes ($\sigma_1^2 \neq \sigma_2^2$) ver Tabela 6.

Tabela 6 – Teste t-Student para variâncias diferentes

	Idade	Frequência Cardíaca
t_0^*	3.9838	7.5990
P	0.0001	0.0001

Para o intervalo de confiança de 95% adotado, o teste bilateral mostrou que as evidências estatísticas não rejeitaram a hipótese nula apenas para o atributo “Colesterol-S”, ou seja, há uma diferença real entre as médias das amostras dos demais atributos.

Para o banco de dados Húngaro obteve-se a média e dispersão dos atributos de análise, como mostrado na Tabela 7, posteriormente fez-se o teste F para verificar se as amostras tinham variâncias iguais, os resultados do teste mostram-se na Tabela 8.

Tabela 7 – Média e dispersão para o banco de dados Húngaro

Atributo	Diagnóstico dos Pacientes com estreitamento da Artéria Principal	
	< 50%	> 50%
Idade	46.89 ± 7.96	49.27 ± 7.33
Pressão do Sangue	130.54 ± 16.61	135.74 ± 18.69
Colesterol-S	239.55 ± 56.62	269.19 ± 79.91
Frequência Cardíaca	145.18 ± 21.83	129.06 ± 22.78

Tabela 8 – Teste F no banco de dados Húngaro

	Idade	Pressão do Sangue	Colesterol-S	Frequência Cardíaca
f_0	1.1811	1.2656	1.9920	1.0884
P	0.1821	0.0897	0.0001	0.3118

Como mostrado na Tabela 8 para o atributo Colesterol rejeita-se a hipótese de variâncias iguais. Realizou-se o teste t-Student de forma similar ao caso anterior, os resultados mostram-se na Tabela 9 e 10.

O teste de hipótese comparativo para as médias de ambas as amostras estudadas (Tabelas 9 e 10), resultou na rejeição da hipótese nula para todos os atributos, ou seja, há uma diferença real entre as médias das amostras

de Idade, Pressão Sanguínea, Colesterol-S e Frequência Cardíaca.

Tabela 9 – Teste t-Student para variâncias iguais

	Idade	Pressão do Sangue	Frequência Cardíaca
t_0	-2.4469	-2.3734	5.7739
P	0.0151	0.0183	0.0000

Tabela 10 – Teste t-Student para variâncias diferentes

	Colesterol-S
t_0^*	-2.7606
P	0.0064

5 Discussão e Conclusões

O uso da análise estatística permitiu inferir a influência do atributo Colesterol-S para o banco de dados Cleveland. Observa-se que enquanto o teste F assinalava dispersões iguais para os atributos Colesterol-S e Frequência cardíaca, somente o atributo Colesterol-S passou o teste t-Student. Já para o banco de dados Húngaro o teste F assinalou o Colesterol-S como um atributo com dispersões diferentes, e no teste t-Student nenhum atributo passou o teste. Descartou-se a influência dos atributos em análise no maior ou menor percentual de estreitamento da artéria principal.

É possível que a distribuição das amostras não seja normal, considerando que existe uma classificação para diagnósticos confirmados em quatro graus [1-4]. Assim, uma nova análise utilizando testes não-paramétricos e considerando as amostras por grau de diagnóstico deverá ser feita.

Agradecimentos

Este trabalho foi parcialmente financiado pela CAPES.

Referências

- [1] AHA, D. W.; KIBLER, D.; ALBERT, M. K.: Instance-Based learning algorithms. *Machine Learning* 6, 1991, 37-66.
- [2] BARBETTA, P. A.: *Estatística Aplicada às Ciências Sociais*. Florianópolis: UFSC, 5ª edição, 2003.
- [3] DALY, C.; NORRIE J.; MURDOCH, D.L.; FORD, I.; DARGIE, H.J.; FOX, K. E TIBET study group: The value of routine non-invasive tests to predict clinical outcome in stable angina. *European Heart Journal* 24, 6, 2003, 532-540.
- [4] DETRANO, R.; JANOSI, A.; STEINBRUNN, W.; PFISTERER, M.; SCHMID, J.J.; SANDHU, S.; GUPPY, K. H.; LEE, S.; FROELICHER, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology* 64, 5,

- 1989, 304-310.
- [5] FORATTINI, O. P.: *Epidemiologia Geral*. São Paulo: Artes Médicas, 2ª edição, 1996.
- [6] GULATI, M.; PANDEY, D. K.; ARNSDORF, M. F.; LAUDERDALE, D. S.; THISTED, R. A.; WICKLUND, R. H.; AL-HANI, A. J.; BLACK, H. R.: Exercise Capacity and the Risk of Death in Women: The St James Women Take Heart Project. *Circulation Journal of the American Heart Association* 108, 13, 2003, 1554 – 1559.
- [7] MASSAD, E.; MENEZES, R. X.; SILVEIRA, P. S. P.; ORTEGA, N. R. S.: *Métodos Quantitativos em Medicina*. São Paulo: Manole, 2004.
- [8] MONTGOMERY, C. D. e RUNGER, G. C.: *Estatística Aplicada e Probabilidade para Engenheiros*. Rio de Janeiro: LTC, 2ª edição, 2003.
- [9] MOYA, R.; SARAIVIA, G.: *Probabilidade e Inferencia Estadística*. Lima-Perú: Editorial San Marcos, 2ª Edição, 1988.
- [10] NASSAR, S. M.: Sistema Estatístico Inteligente para Apoio a Pesquisas Médicas, Tese de Doutorado em Engenharia Elétrica, UFSC, 1995.
- [11] ROUQUAYROL, M. Z.: *Epidemiologia & Saúde*. Editado com auxílio do CNPq. Fortaleza: Impresso na Universidade de Fortaleza, 1983.
- [12] UCI Machine Learning Repository: Databases in heart-disease.
http://www.ics.uci.edu/~mlearn/databases/heart-dise_ase/ (12/out./2006)
- [13] WIKIPEDIA: Icterícia Neonatal.
<http://pt.wikipedia.org/wiki/Icterícia> (12/out./2006)
- [14] WYNS, W.; MUSSCHAERT-BEAUTHIER, E.; VAN DOMBURG, R.; LUBSEN, J.; ROUSSEAU, M. F.; COSYNS, J.; DETRY, J. M. R.: Prognostic value of symptom limited exercise testing in men with a high prevalence of coronary artery disease. *European Heart Journal* 6, 11, 1985, 939-945.