versão impressa ISSN 0101-7438 / versão online ISSN 1678-5142

SELEÇÃO DE VARIÁVEIS E CLASSIFICAÇÃO DE PADRÕES POR REDES NEURAIS COMO AUXÍLIO AO DIAGNÓSTICO DE CARDIOPATIA ISQUÊMICA

Thiago Baptista Rodrigues*
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro – RJ, Brasil
thiagobr@cepel.br

José Leonardo Ribeiro Macrini
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro – RJ, Brasil
macrini@metrologia.ctc.puc-rio.br

Elisabeth Costa Monteiro
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro – RJ, Brasil
beth@puc-rio.br

* *Corresponding author* / autor para quem as correspondências devem ser encaminhadas

Recebido em 05/2007; aceito em 04/2008 após 1 revisão
Received May 2007; accepted April 2008 after one revision

Resumo

Este estudo propõe uma metodologia, baseada em procedimentos quantitativos, para auxiliar o diagnóstico de indivíduos portadores de doença cardíaca. Os resultados obtidos neste estudo, utilizando "Redes Neurais", foram comparados aos resultados de outros autores. Um percentual de acerto médio de 91,0 % foi atingido, enquanto outros estudos utilizando a mesma base de dados atingem até 83,5 %. Foram utilizadas também outras técnicas de classificação de padrões conhecidas na literatura, denominadas "Análise Discriminante" e "Algoritmo C4.5", de forma a estabelecer comparações com os resultados aqui obtidos utilizando "Redes Neurais". A metodologia de divisão dos conjuntos de treinamento/generalização sugerida promoveu melhorias em todas as três técnicas de classificação de padrões utilizadas, tendo a Rede Neural apresentado o melhor desempenho.

Palavras-chave: informação mútua; redes neurais; doença cardíaca.

Abstract

This study proposes a methodology, established in quantitative procedures, to assist the diagnostic of individuals with heart disease. The results obtained in this study using "Neural Networks" had been compared with the results of other authors. A percentage of average correct diagnosis of 91.0 % was reached, whereas other studies using the same database had reached until 83.5 %. Others techniques of classification of standards known in literature had also been used, called "Discriminate Analysis" and "C4.5 Algorithm", to establish comparisons with the results obtained here using "Neural Networks". The methodology of division of the sets of training/generalization suggested promoted improvements in all the three used techniques of classification of standards, with the "Neural Networks" showing the best performance.

Keywords: mutual information; neural networks; heart disease.

1. Introdução

Este estudo busca implementar e analisar uma metodologia, baseada em procedimentos quantitativos, para auxiliar o diagnóstico de indivíduos com doença cardíaca, através da investigação e seleção das variáveis que possuem maior grau de importância na determinação deste evento. A metodologia foi implementada em um grupo de indivíduos do banco de dados público intitulado *"Heart Disease Database"* (Base de Dados pública de Doença Cardíaca) (Aha, 2001), diagnosticados nas cidades de *Cleveland* e *Long Beach*, nos Estados Unidos. São ao todo 303 indivíduos, que foram avaliados com relação a 76 atributos, porém, a maioria das pesquisas publicadas refere-se ao uso de um subconjunto de 14 deles (13 atributos mais o desfecho), já identificados, a priori, como os mais relevantes. Os resultados obtidos neste estudo foram comparados aos resultados de outros autores encontrados na literatura, de forma a se ter uma medida da qualidade dos resultados aqui obtidos.

O diagnóstico de doença cardíaca foi realizado através de um modelo de Redes Neurais após uma seqüência de procedimentos de pré-tratamento e seleção de variáveis. Este tratamento é fundamental devido ao número relativamente reduzido de amostras no banco de dados.

O sedentarismo, o colesterol alto e o estresse são alguns dos piores inimigos do coração. Eles se encaixam dentro dos fatores chamados modificáveis, ou seja, que são passíveis de reparo a partir de mudanças de hábito da população. Algumas pessoas têm chance maior de desenvolver doenças cardíacas, principalmente levando em conta idade e histórico familiar. Mas isso não quer dizer que os mais jovens não devam se preocupar. Se a pessoa tem na família irmãos ou pais que já sofreram um infarto na faixa dos 30 anos, deve redobrar ainda mais os cuidados, caso se identifique com algum dos fatores modificáveis citados.

Angina e infarto são algumas das doenças que mais assustam. Ambas têm como importante fator de risco o excesso de colesterol no sangue. O colesterol está associado a aproximadamente 4,4 milhões de mortes anuais no mundo.

As informações sobre doenças cardíacas apresentadas aqui foram fornecidas pelo cardiologista Carlos Vicente Serrano e pelo clínico geral João Carlos Coluço.

1.1 Angina

A angina é uma das manifestações da doença arterial coronária, que é a formação de placas de gordura por deposição nas artérias do coração. No caso da angina, a placa de gordura bloqueia o fluxo de sangue quase totalmente. Com isso, o músculo cardíaco recebe menor fluxo sangüíneo, que pode ser insuficiente para a execução de sua atividade contrátil.

Os sintomas da angina são os mesmos do infarto, só que com menor intensidade e duração (menos de 20 minutos). O mais comum é o indivíduo referir uma sensação de aperto no peito que irradia para o braço esquerdo, geralmente acompanhado por outros sintomas como falta de ar, sudorese, palidez e náuseas. A pessoa pode sentir também dor no braço direito, mandíbula, pontadas no peito e falta de ar sem desconforto. Ter uma dieta leve e sem colesterol, evitar o tabagismo, a obesidade e o estresse, fazer avaliações periódicas, são algumas formas de se prevenir da angina.

1.2 Infarto Agudo do Miocárdio

O Infarto Agudo do Miocárdio é uma das conseqüências da doença arterial coronária, que é a formação de placas de gordura (colesterol e triglicerídios) por deposição nas artérias do coração. A gordura interrompe uma ou mais artérias, o que barra o fluxo sanguíneo no músculo cardíaco e provoca o infarto. Dependendo do comprometimento do músculo cardíaco, pode ocorrer uma insuficiência cardíaca. Há vários fatores que deixam o indivíduo mais vulnerável a um infarto como a idade, sexo, histórico familiar, tabagismo, hipertensão, diabetes, nível de colesterol sanguíneo, estresse, obesidade e sedentarismo.

Sinais da doença como aperto no peito que irradia para o braço esquerdo, geralmente vêm acompanhados por outros sintomas como falta de ar, sudorese, palidez e náuseas. A dor, intensa e prolongada, dura cerca de 20 minutos. Os sintomas também se manifestam de formas atípicas, como dor no braço direito, mandíbula, pontadas e apenas falta de ar sem desconforto.

Para prevenir essa doença, é importante a pessoa reconhecer que se encaixa no grupo de risco e ficar atenta, sobretudo, a um novo sintoma, que geralmente surge com o esforço físico ou emocional. A partir dos 35 anos, a pessoa deve fazer um check-up para avaliar colesterol, pressão, etc. Com o resultado da primeira avaliação, o médico determinará se os controles terão de ser anuais ou bianuais. O indivíduo deve também evitar o tabagismo, elevados níveis de colesterol sanguíneo e obesidade. É importante também cuidar do diabetes e da hipertensão, manter uma alimentação saudável e a prática de exercícios físicos. A chance de um infarto nos homens aumenta aos 55 anos; nas mulheres, a idade é 65 anos (a idade entre as mulheres é mais alta por causa da proteção dos hormônios), mas isso não impede que uma pessoa mais jovem tenha infarto. É preciso ficar ainda mais atento ao histórico familiar, sobretudo se pai ou mãe tiveram um ataque cardíaco quando jovens.

2. Materiais e Métodos

2.1 Descrição da Base de Dados

A base de dados de doenças do coração utilizada neste estudo (*"Heart Disease Database"*) é formada originalmente por quatro subconjuntos diferentes (1 da Hungria, 1 da Suíça, e 2 dos Estados Unidos), mas somente os dados de Cleveland são utilizados, pois os demais subconjuntos apresentam muitos dados incompletos. Os dados de Cleveland foram obtidos no *V.A. Medical Center*, em *Long Beach*, e no *Cleveland Clinic Foundation*, em *Cleveland*, ambas nos Estados Unidos, por Robert Detrano. A base de dados contém 303 amostras, das quais 297 são amostras completas e 6 amostras apresentam dados incompletos (desta forma, foram utilizadas somente as 297 amostras neste estudo). Do total de amostras utilizado no estudo, 160 são amostras de indivíduos não-doentes, e o restante, 137, de indivíduos doentes. Originalmente, o banco de dados contém 76 atributos. Contudo, devido ao grande número de dados incompletos na maioria dos atributos, todas as pesquisas publicadas referem-se à utilização de somente 13 destes, os quais são listados a seguir:

1 → Idade (AGE) – variando de 29 a 77 anos;

2 → Sexo (SEX) – masculino ou feminino, sendo representados por 1 ou 0, respectivamente;

3 → Tipo de dor no peito (CP) – quatro tipos de dor no peito:
- Valor 1: angina típica;
- Valor 2: angina atípica;
- Valor 3: sem dor anginal;
- Valor 4: assintomático.

4 → Pressão arterial em repouso (TRESTBPS) – medida em mm Hg;

5 → Colesterol no soro sangüíneo (CHOL) – medido em mg/dl;

6 → Concentração de açúcar no sangue (FBS) > 120 mg/dl – verdadeiro (1) ou falso (0);

7 → Resultado da eletrocardiografia em repouso (RESTECG):
- Valor 0: Normal;
- Valor 1: Com onda ST-T anormal;
- Valor 2: Mostrando provável (ou definida) hipertrofia do ventrículo esquerdo.

8 → Máxima taxa de batimento cardíaco atingida (THALACH);

9 → Angina induzida por exercício (EXANG):
- Valor 1: Sim;
- Valor 0: Não.

10 → Depressão ST induzida por exercício relativamente sossegado (OLDPEAK);

11 → Inclinação da extremidade do segmento ST no exercício (SLOPE):
- Valor 1: Inclinado para cima;
- Valor 2: Plano;
- Valor 3: Inclinado para baixo.

12 → Número de vasos coloridos pela fluoroscopia (CA) – valor de 0 a 3;

13 → Talassemias (THAL):
- Valor 3: normal;
- Valor 6: defeito fixo (irreparável);
- Valor 7: defeito reversível (reparável).

Existem duas classes de saída para diagnóstico de 'doença cardíaca' (*Heart Disease*): menos de 50 %, e 50 % ou mais de estreitamento do diâmetro do vaso sangüíneo, sendo representadas por "0" e "1", respectivamente. Convencionou-se que indivíduos com valor "0" na classe de saída seriam chamados de "indivíduos não-doentes", e aqueles com valor "1", de "indivíduos doentes".

O estudo envolveu um tratamento criterioso das variáveis de forma a selecionar aquelas relevantes e não-redundantes para o diagnóstico de doença cardíaca pelo modelo proposto.

2.2 Tratamento da Base de Dados

O tratamento da base de dados utilizada neste estudo foi baseado na seleção de variáveis através de *Informação Mútua*. A informação mútua entre duas variáveis aleatórias pode ser dada pela seguinte equação:

$$I(X|Y) = H(X) - H(X|Y) \qquad 2.1$$

Onde X e Y são as variáveis aleatórias. A entropia de X (H (X)) e a entropia de X depois de observados os valores de Y (H (X|Y)), podem ser dadas, respectivamente, por:

$$H(X) = -\sum_i P(x_i) \cdot \log(P(x_i)) \qquad 2.2$$

$$H(X|Y) = -\sum_j P(x_j) \sum_i P(x_i \mid y_j) \cdot \log\left(P(x_i \mid y_j)\right) \qquad 2.3$$

Onde:

$P(x_i)$ ➔ Probabilidades a priori para todos os valores de X;

$P(x_i \mid y_j)$ ➔ Probabilidades a posteriori de X dados os valores de Y.

Sobre o conjunto de 13 variáveis do banco de dados, utilizou-se o algoritmo de Seleção de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação (MIFS-U) (Kwak & Choi, 2002). Trabalhos anteriores utilizando informação mútua para seleção de variáveis no contexto de Redes Neurais incluem Battiti (1994), Daberllay (1997, 2000) e Setiono & Liu (1997). O resultado desta etapa é uma seleção das variáveis por ordem de importância. Este procedimento é resumido a seguir:

Seja F o conjunto de variáveis de entrada do banco (no caso, as 13 variáveis iniciais) e S o conjunto de variáveis a serem selecionadas por ordem de importância. S é inicialmente um conjunto vazio.

Calcula-se então, a informação mútua (IM) entre cada uma dessas variáveis com a variável desfecho. A informação mútua é uma medida de dependência entre variáveis aleatórias. O cálculo da IM apresenta sempre um valor maior ou igual a "0", e quando este valor for igual a "0", isto significará independência estatística entre as variáveis aleatórias em questão (Cover, 1991). Um valor alto, ou pequeno, de IM significa que as variáveis são muito, ou pouco relacionadas, respectivamente.

Posteriormente, seleciona-se (fi), a variável que apresenta a maior informação mútua com o desfecho. Então, a variável selecionada (fi) sai do conjunto F ($F \leftarrow F-\{fi\}$) e entra no conjunto S ($S \leftarrow \{fi\}$). O processo de cálculo da informação mútua deve então se repetir considerando-se agora (1) o desfecho, (2) a variável selecionada e (3) as variáveis candidatas a serem selecionadas. Este processo torna-se rapidamente complexo, uma vez que se teria de calcular uma distribuição conjunta de três variáveis aleatórias, o que implicaria na necessidade de se estimar a função densidade de probabilidade conjunta destas 3 variáveis. Este procedimento fica ainda mais difícil no caso presente, por causa da pequena quantidade de dados. O algoritmo MIFS-U (Kwak & Choi, 2002), e os resultados obtidos, reduzem o problema a cálculos de informação mútua entre duas variáveis aleatórias.

Obtém-se então uma ordenação por importância das 13 variáveis do banco de dados em questão. Esta ordenação está apresentada na Tabela 2.1. É importante salientar que os dados apresentados na última coluna da Tabela 2.1 correspondem aos valores de informação mútua entre cada uma das variáveis e o desfecho, que foi utilizado para selecionar a primeira variável na ordem de importância.

É importante salientar, que somente as variáveis relevantes e não-redundantes são de interesse para que haja uma determinação otimizada do desfecho, ou seja, do diagnóstico de doença cardíaca. Logo, as variáveis irrelevantes e redundantes devem ser eliminadas do conjunto de variáveis.

Tabela 2.1 – Resultado da Seleção e Ordenação das Variáveis.

Ordem de Importância	Variáveis	IM
1	THAL	0,2102
2	CA	0,1846
3	CP	0,1972
4	EXANG	0,1323
5	SLOPE	0,1088
6	SEX	0,0579
7	RESTECG	0,0235
8	FBS	0,0000
9	OLDPEAK	0,1682
10	THALACH	0,1674
11	AGE	0,0955
12	CHOL	0,0610
13	TRESTBPS	0,0526

2.3 Proposta de Modelagem para Diagnóstico de Cardiopatia Isquêmica

O diagnóstico de doença cardíaca isquêmica foi estimado utilizando uma Rede Neural *feedforward*. Este tipo de modelo não-linear tem sido utilizado com sucesso em uma gama extensiva de aplicações desde o final da década de 80. Referências clássicas em Redes Neurais incluem Haykin (1998), Bishop (1995) e Príncipe *et al*. (2000).

Rede Neural Artificial (ou simplesmente "Rede Neural") é um modelo distribuído composto por unidades (chamadas na literatura de "neurônios") constituídas de funções não-lineares (tipicamente sigmóides e tangentes hiperbólicas). A combinação destas unidades, através de parâmetros estimados a partir dos dados, é o que confere a capacidade deste modelo de inferir relações não-lineares de complexidade arbitrária. Na forma utilizada neste estudo, estas unidades são arrumadas em camadas, incluindo uma camada oculta, que não está diretamente conectada à saída do modelo. Estas conexões entre as unidades, ou neurônios, são chamadas de pesos (originalmente a terminologia era "pesos sinápticos"). Estes pesos são os parâmetros do modelo que são ajustados por um algoritmo iterativo através dos dados. Uma vez ajustados os pesos, a rede tem a capacidade de representar a relação dos dados de entrada com a variável de saída, neste caso o diagnóstico de doença. A capacidade de aprender através de "exemplos" ou dados (na-amostra) e de generalizar (fora-da-amostra) informação gerada em ambientes não-lineares complexos, é sem dúvida a grande vantagem das Redes Neurais.

As variáveis relevantes e não-redundantes são utilizadas como entrada da Rede Neural, e, após o processo de treinamento, tem-se como saída da Rede o diagnóstico de doença cardíaca. Assim, na fase de treinamento, a saída da rede assume valores 0 ou 1 para cada uma de duas possibilidades, não-doente ou doente, respectivamente. Utiliza-se uma função de ativação sigmóide na unidade de saída de forma que a saída da Rede varie sempre entre 0 e 1, já que esta função satura nestes valores. Já na fase de teste, e posteriormente de utilização do modelo, valores próximos a 1 indicam grande possibilidade do indivíduo ser doente e próximos de 0 indicam pequena possibilidade deste ser doente. O ponto de corte adotado para diferenciar grande de pequena possibilidade foi 0,5 e desta forma considerou-se

apenas dois tipos de saída, doente ou não-doente, para valores acima de 0,5 ou valores abaixo de 0,5, respectivamente.

Na Rede implementada, utilizou-se o algoritmo de Regularização Bayesiana (Mackay, 1992). Neste algoritmo, assume-se que os parâmetros da Rede são variáveis aleatórias com distribuições especificadas. Os parâmetros de regularização são variâncias desconhecidas associadas a estas distribuições e pode-se calcular estes parâmetros utilizando então técnicas estatísticas. Portanto o modelo não é especificado de uma forma arbitrária.

O aprendizado ou treinamento de uma rede neural tem tipicamente por objetivo reduzir a soma dos quadrados dos erros (Foresee & Hagan, 1997), conforme a seguinte equação:

$$\hat{\psi} = \arg\min_{\psi} Q_1(\psi) = \arg\min_{\psi} \sum_{t=1}^{N} (y_t - G(x,\psi))^2 \qquad 2.4$$

Onde $(x,\psi) \in XX\Psi$, sendo $x = [x_1, x_2, ..., x_I]$ vetores de variáveis independentes e ψ o vetor de parâmetros $\psi = [\alpha, \gamma]$, composto pelos vetores de pesos da camada de saída e da camada oculta respectivamente; y_t é a saída alvo da Rede e $G(x,\psi)$ é a saída estimada pela Rede. Assim como outros modelos flexíveis não-lineares, as Redes Neurais podem sofrer de *overfitting*. Este problema ocorre quando é utilizado um número excessivo de neurônios na camada oculta, que levarão a uma perda da capacidade de generalização (fora-da-amostra). Em contrapartida, se o número de neurônios em excesso for reduzido, ocorre a perda da capacidade de aproximar o processo gerador dos dados (Medeiros & Pedreira, 2001).

Atualmente, diversas metodologias são utilizadas para solucionar o problema de *overfitting* (Haykin, 1998). Neste estudo, será utilizado 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 & Pedreira (2001), o problema de estimação passa a ser definido como:

$$\hat{\psi} = \arg\min_{\psi} Q_T(\psi) = \arg\min_{\psi} \sum_{t=1}^{N} (\eta Q_1(\psi) - \phi Q_2(\psi))^2 \qquad 2.5$$

Onde ϕ e η são parâmetros de regularização, $Q_1(\psi)$ pode ser deduzido da equação 2.4, e $Q_2(\psi)$ é a função de penalização, que é dada pela soma do quadrado dos parâmetros (α, γ), vetores de pesos da camada de saída e da camada oculta, respectivamente, conforme a seguinte equação:

$$Q_2(\psi) = \sum_{h=0}^{H} \alpha_h^2 + \sum_{h=0}^{H} \sum_{i=0}^{I} \gamma_{hi}^2 \qquad 2.6$$

O problema de regularização é otimizar a função objetivo de forma a encontrar valores para os parâmetros de regularização 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 & Hagan, 1997) propõe a aproximação da matriz Hessiana pelo algoritmo de Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963), reduzindo o custo computacional.

Todos os modelos utilizados neste estudo tiveram como arquitetura da rede neural uma camada de entrada, uma camada escondida com dez neurônios e uma camada de saída com um

neurônio. A função de ativação sigmóide foi utilizada em todos os neurônios, inclusive no de saída. Os pesos e os *bias* foram inicializados através do algoritmo de Nguyen-Widrow (1989).

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*, Haykin, 1998 e Bishop, 1995). 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). No caso do banco de dados deste estudo, o uso do critério de *leave-one-out* seria o mais indicado. Porém, para que fosse possível estabelecer uma comparação com pesquisas previamente realizadas e encontradas na literatura, e também aplicar a metodologia proposta neste estudo, o conjunto de dados disponível foi dividido aleatoriamente em um conjunto de treinamento e em um conjunto de teste, na proporção de 2/3 das amostras para treinamento e 1/3 para teste (generalização). Desta forma, foi utilizada a mesma proporção das pesquisas previamente realizadas encontradas na literatura, às quais foram comparados os resultados aqui obtidos. Para que se tivesse uma boa amostragem estatística, foram utilizados 50 conjuntos de treinamento/generalização diferentes nos experimentos aqui realizados e a média dos resultados desses 50 conjuntos foi adotada como resultado final. Ou seja, do número total de amostras, 2/3 foram sorteadas aleatoriamente e utilizadas para o treinamento, sendo o 1/3 restante utilizadas para a generalização; este procedimento foi repetido 50 vezes de forma a se ter sempre conjuntos diferentes de treinamento/generalização, adotando-se a média desses 50 conjuntos como resultado final.

Devido à quantidade reduzida de amostras do banco de dados, a validação do método será feita utilizando o próprio conjunto de teste.

3. Resultados e Análises

Neste item serão apresentados os resultados obtidos com a aplicação da metodologia ao banco de dados descrito anteriormente.

As possibilidades de medidas para avaliar o desempenho da metodologia proposta são descritas a seguir:

(i) – Percentual de acerto dos não-doentes (PND):

É definido como:

$$PND = \frac{ND_R}{ND} x100 \qquad 3.1$$

Sendo:

ND_R ➔ número de indivíduos que a Rede estima como não-doente quando efetivamente o indivíduo é não-doente;
ND ➔ número total de indivíduos não-doentes.

(ii) – Percentual de acerto dos doentes (PD):

É definido como:

$$PD = \frac{D_R}{D} x100 \qquad 3.2$$

Sendo:

D_R ➔ número de indivíduos que a Rede estima como doente quando efetivamente o indivíduo é doente;
D ➔ número total de indivíduos doentes.

(iii) – Percentual de acerto médio (PM):

É definido como:

$$PM = \frac{PND \cdot ND + PD \cdot D}{ND + D} \qquad 3.3$$

De acordo com os valores de PND, PD e PM, obtidos nos experimentos aqui realizados, pode-se avaliar a qualidade do método proposto, estabelecendo-se comparações dos valores aqui obtidos com valores encontrados em pesquisas realizadas por outros autores e encontradas na literatura.

3.1 Analisando os Resultados da Rede Neural

Na Tabela 2.1, tem-se a ordenação das 13 variáveis presentes no banco e seus respectivos valores de informação mútua (IM) com relação ao desfecho. É importante salientar que o fato de uma variável ter IM nula com relação ao desfecho não significa que conjuntamente com outras variáveis ela não possa vir a ter importância. É por isso que a oitava variável em ordem de importância (FBS) também foi utilizada nas simulações da rede neural, conforme pode ser visto na Tabela 2.1, de forma que se pudesse verificar sua contribuição para a classificação na presença das outras variáveis. As variáveis ordenadas da nona à décima terceira posição não apresentam uma contribuição significativa ao percentual de acerto quando utilizadas nas simulações da rede neural. Assim, foram utilizadas as 8 primeiras variáveis da Tabela 2.1 como entrada da rede neural, da seguinte forma: o número de entradas da rede foi variado de 1 a 8, seguindo a ordem de importância na qual as variáveis foram ordenadas. Desta forma, a Rede Neural foi executada inicialmente com a primeira variável como entrada (THAL), depois com as duas variáveis mais importantes como entrada (THAL e CA), depois com as três variáveis mais importantes como entrada (THAL, CA e CP), e assim sucessivamente, até ter sido executada com as 8 variáveis ordenadas. Os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, foram os seguintes:

Tabela 3.1 – Percentual de Acerto na Generalização.

Variáveis de entrada	PND	PD	PM
Primeira	79,7	70,3	75,4
2 Primeiras	76,9	79,2	78,0
3 Primeiras	85,9	76,3	81,4
4 Primeiras	82,8	76,0	79,6
5 Primeiras	85,1	75,6	80,7
6 Primeiras	79,9	78,2	79,1
7 Primeiras	80,7	77,8	79,4
8 Primeiras	79,0	79,3	79,2

Como o que se deseja é analisar a qualidade da classificação feita pelo modelo com relação aos indivíduos que não pertenciam ao conjunto de treinamento, deve-se analisar então os resultados do acerto na generalização (fora-da-amostra). O resultado mais representativo encontrado foi então aquele obtido quando se utilizou 3 variáveis de entrada (THAL, CA e CP), obtendo-se um percentual de acerto médio (PM) fora-da-amostra de 81,4 %. É importante que os percentuais de acerto fora-da-amostra para os casos de não-doente (PND = 85,9 %) e doente (PD = 76,3 %), e na amostra (ou seja, no treinamento) para os casos de não-doente e doente (89,5 % e 81,7 %, respectivamente), também apresentem bons resultados, o que foi constatado neste caso.

Apesar de ter sido encontrado um bom resultado, percebeu-se ao longo das simulações da rede neural que, para os melhores resultados obtidos, classificavam-se erroneamente quase sempre os mesmos indivíduos. Ou seja, existia um patamar ótimo de percentagem de acerto, do qual não se conseguia ir além, pois existia um grupo de indivíduos que continuamente era classificado erroneamente, mesmo que o número de variáveis de entrada da rede fosse alterado. Por isso, decidiu-se separar e estudar com mais atenção este grupo de indivíduos e fazer simulações na rede sem esses indivíduos para ver se o resultado melhorava e o quanto ele melhorava. Para isso foram excluídos 10 indivíduos não-doentes e 21 indivíduos doentes (31 indivíduos no total), permanecendo agora 150 indivíduos não-doentes e 116 indivíduos doentes para serem utilizados em uma nova simulação (266 indivíduos no total). Antes havia um total de 160 indivíduos não-doentes e 137 indivíduos doentes no banco de dados.

Feito isto, uma nova simulação foi realizada somente com os 266 indivíduos que restaram após a exclusão dos 31 que eram sempre classificados erroneamente. Mantendo-se a mesma proporção de amostras treinamento/generalização da simulação anterior, utilizou-se então 2/3 das amostras para treinamento (177 indivíduos) e 1/3 para generalização (89 indivíduos). Os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, foram os seguintes:

Tabela 3.2 – Percentual de Acerto na Generalização (sem os 31 indivíduos).

Variáveis de entrada	PND	PD	PM
Primeira	83,9	80,6	82,4
2 Primeiras	83,2	88,2	85,3
3 Primeiras	94,4	92,7	93,7
4 Primeiras	93,2	90,5	92,0
5 Primeiras	91,6	88,7	90,4
6 Primeiras	91,0	90,8	90,9
7 Primeiras	89,9	89,5	89,7
8 Primeiras	91,0	88,7	90,0

Desta forma, pode-se perceber o quanto estes 31 indivíduos interferem na qualidade da classificação, visto que com eles inseridos no banco, o resultado de PM na generalização e usando as 3 variáveis (THAL, CA e CP) era 81,4 %, passando para 93,7 % quando estes foram excluímos do banco.

Utilizando-se novamente todo o banco de dados (297 indivíduos), foram feitas novas simulações na rede da seguinte forma: com a mesma proporção de 2/3 do banco de dados para treinamento e 1/3 para generalização. Os indivíduos de cada grupo foram escolhidos de forma a se garantir que os 31 que foram classificados como pertencentes ao grupo de **"indivíduos sempre classificados erroneamente", pertençam sempre ao conjunto de treinamento**. Os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, foram os seguintes:

Tabela 3.3 – Percentual de Acerto na Generalização (com os 31 indivíduos no treinamento).

Variáveis de entrada	PND	PD	PM
Primeira	82,0	82,7	82,3
2 Primeiras	83,9	90,8	87,1
3 Primeiras	91,9	89,8	91,0
4 Primeiras	87,3	87,7	87,5
5 Primeiras	85,6	83,7	84,7
6 Primeiras	83,6	86,4	84,9
7 Primeiras	82,7	82,6	82,7
8 Primeiras	80,8	85,7	83,1

Assim, pode-se perceber o quanto o resultado da classificação dos indivíduos melhora quando se compara os resultados da Tabela 3.1 com os resultados mostrados na Tabela 3.3. Utilizando-se as 3 variáveis (THAL, CA e CP) e todo o banco de dados, quando se escolhe os indivíduos de cada grupo de forma a garantir que os 31 indivíduos pertençam sempre ao conjunto de treinamento, o resultado de PM na generalização passou de 81,4 % (Tabela 3.1) para 91,0 % (Tabela 3.3).

A idéia é que, se este grupo de 31 indivíduos (ou parte deste grupo) apresentar um padrão de comportamento especial, este padrão será assimilado (aprendido) pela rede neural durante o treinamento, e possíveis indivíduos que apresentem este padrão de comportamento especial no conjunto de generalização (teste) poderão ser classificados corretamente.

A análise dos gráficos de colunas das duas primeiras variáveis na ordem de importância, THAL (Figura 3.1) e CA (Figura 3.2), permite concluir que o comportamento dos 31 indivíduos é praticamente o oposto ao restante dos indivíduos. O padrão que deveria ser apresentado para os não-doentes, está sendo apresentado para os doentes, e o padrão que deveria ser apresentado para os doentes, está sendo apresentado para os não-doentes. Possivelmente esses indivíduos apresentam um padrão de comportamento especial, padrão este que poderá ser assimilado (aprendido) pela Rede Neural durante o treinamento, e possíveis indivíduos que apresentem este padrão de comportamento especial no conjunto de generalização (teste) poderão ser classificados corretamente.

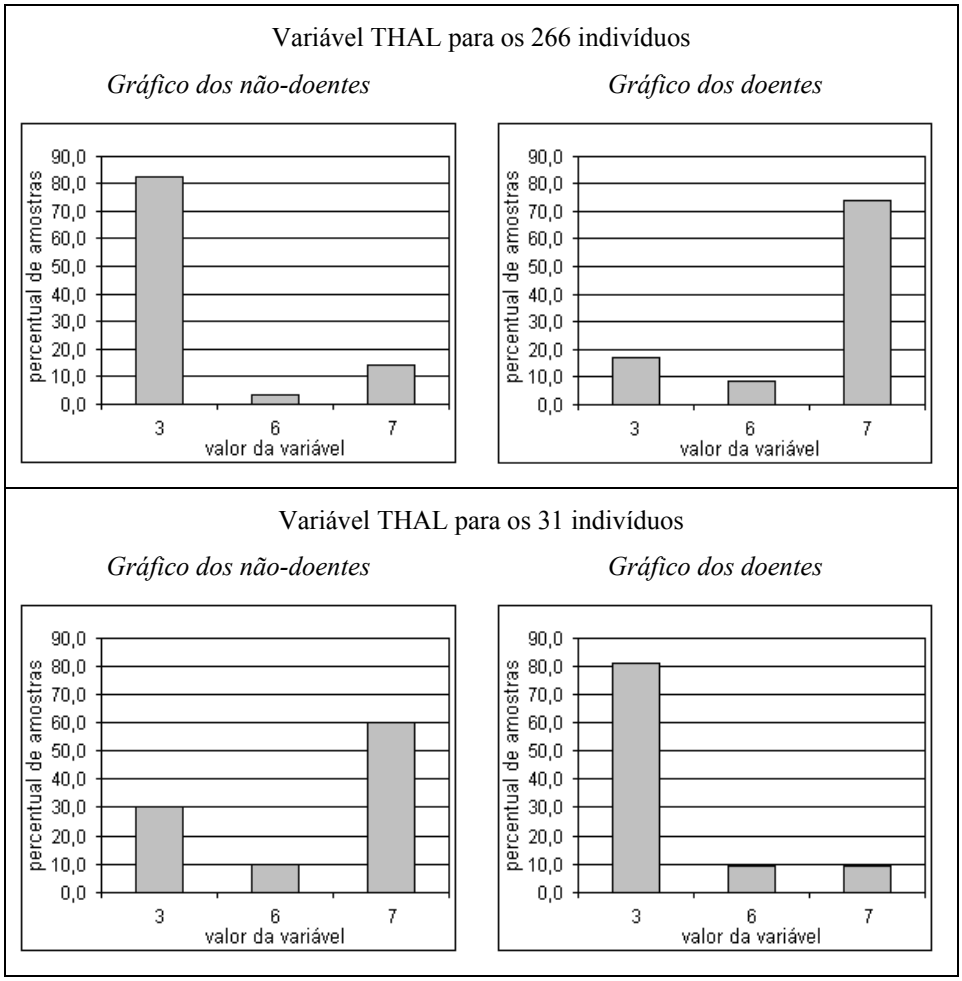

Figura 3.1 – Gráfico de colunas da variável THAL para os 266 e para os 31 indivíduos.

Quando se compara o resultado obtido neste experimento, a partir da metodologia proposta, com resultados obtidos por outros autores encontrados na literatura, percebe-se que o resultado obtido neste estudo apresenta um melhor percentual de acerto. Os resultados obtidos por Ho & Chou (2001) e por Hu, Li, Cai & Xu (2004), onde foi utilizada a mesma proporção aqui escolhida para dividir o banco de dados em conjunto de treinamento e conjunto de generalização, são da ordem de PM = 83,0 % e PM = 83,5 %, respectivamente, apresentados para o conjunto de generalização. O resultado obtido neste experimento com a metodologia aqui proposta, apresentou PM = 91,0 %, para o conjunto de generalização. Portanto, um ganho significativo no percentual médio de acerto, demonstrando a qualidade da classificação segundo a metodologia adotada. Cabe salientar que os outros autores utilizaram todas as 13 variáveis previamente selecionadas, enquanto nesta metodologia bastaram apenas 3 variáveis.

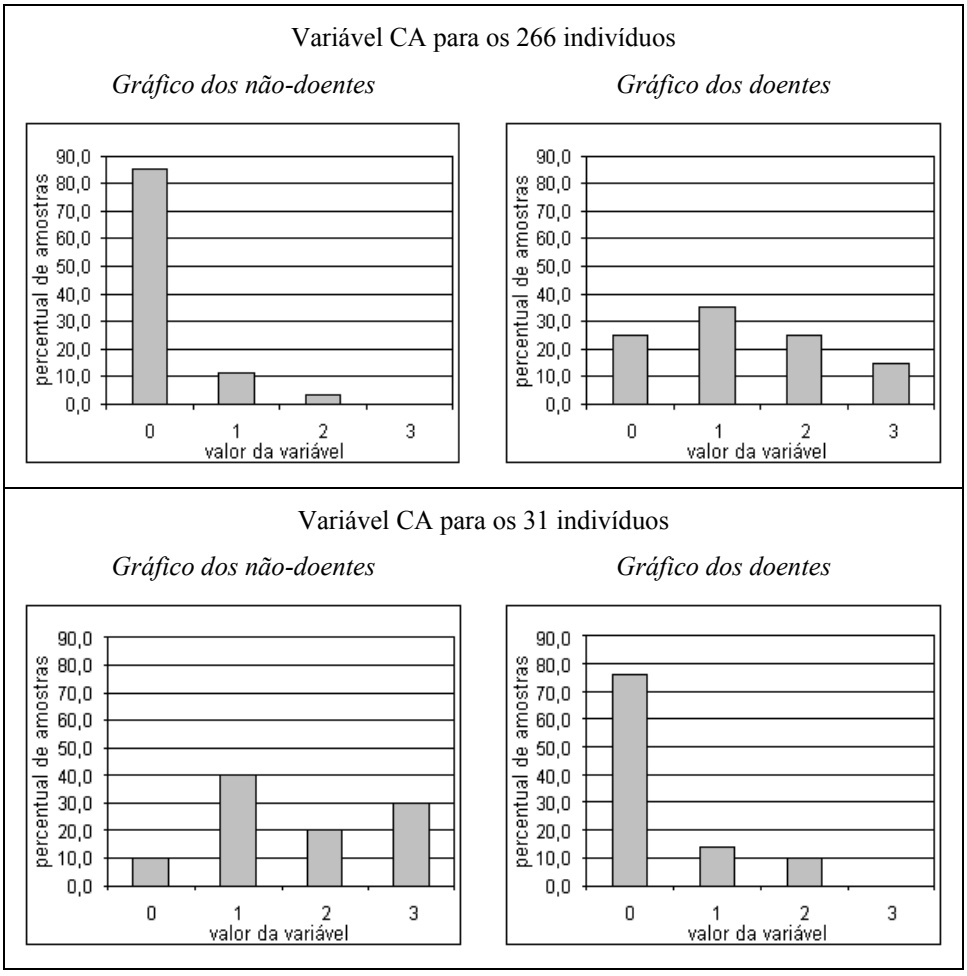

Figura 3.2 – Gráfico de colunas da variável CA para os 266 e para os 31 indivíduos.

3.2 Comparando os Resultados da Rede Neural aos de outros Métodos de Classificação de Padrões

É interessante também, utilizar-se outras técnicas de classificação de padrões conhecidas na literatura de forma a se estabelecer comparações com os resultados obtidos neste estudo, e também para avaliar o impacto sobre estas técnicas de se utilizar a metodologia de divisão dos conjuntos de treinamento/generalização aqui sugerida, ou seja, utilizando-se 2/3 do banco para treinamento e 1/3 para generalização e, da mesma forma, considerando os "indivíduos sempre classificados erroneamente" pertencendo sempre ao conjunto de treinamento.

Foram escolhidas duas técnicas de classificação de padrões encontradas na literatura, que são a **Análise Discriminante** (Haykin, 1998) e o **Algoritmo C4.5** (Quinlan, 1993). A Análise Discriminante é útil para as situações onde se quer construir um modelo preditivo de um

grupo de amostras baseado em características observadas de cada caso. O procedimento gera uma função discriminante (ou, para mais de dois grupos, um conjunto de funções discriminantes) baseada em combinações lineares das variáveis preditas que fornecem a melhor discriminação entre as amostras. As funções são geradas a partir de uma parte dos casos para os quais a classe das amostras é conhecida; as funções podem então ser aplicadas aos casos novos, para os quais a classe das amostras é desconhecida. Já o Algoritmo C4.5 gera um classificador na forma de uma árvore de decisão, com uma estrutura composta por: 1) uma folha, indicando uma classe; 2) um nó de decisão que especifica um teste a ser realizado no valor de um atributo, com um galho para cada resposta possível do teste, que levará para uma sub-árvore ou uma folha. Em uma árvore de decisão a classificação de um caso se inicia pela raiz da árvore, e esta árvore é percorrida até que se chegue a uma folha. Em cada nó de decisão será feito um teste que irá direcionar o caso para uma sub-árvore. Este processo irá guiar-se para uma folha. A classe do caso pressupõe-se que seja a mesma que está armazenada nesta folha.

Segundo as simulações realizadas na rede neural, cujos resultados estão apresentados nas tabelas 3.1, 3.2 e 3.3, pôde-se comprovar que as 3 primeiras variáveis ordenadas (THAL, CA e CP) são as que permitem uma melhor classificação dos indivíduos com relação à determinação do desfecho. Por isso, de forma a se estabelecer comparações, o número de variáveis de entrada nesta etapa foi variado de 1 a 3, seguindo-se a ordem de importância nas quais as variáveis foram colocadas (primeiro com THAL, depois com THAL e CA, e por fim com THAL, CA e CP). Utilizando-se 2/3 do banco para treinamento e 1/3 para generalização, sem nenhum outro refinamento, e aplicando no algoritmo C4.5 e na análise discriminante, os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, foram os seguintes:

Tabela 3.4 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (sem refinamento).

Número de Variáveis	C4.5			Análise Discriminante		
	PND	PD	PM	PND	PD	PM
1	79,4	71,1	**75,5**	79,2	72,7	**76,2**
2	79,6	71,4	**75,8**	77,4	78,6	**78,0**
3	76,9	74,8	**75,9**	84,0	76,8	**80,6**

A Tabela 3.5 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM, e utilizando-se o banco sem nenhum refinamento:

Tabela 3.5 – Tabela Comparativa, Valores de PM (sem refinamento).

Número de Variáveis	C4.5	Análise Discriminante	Rede Neural
1	75,5	76,2	75,4
2	75,8	78,0	78,0
3	75,9	80,6	81,4

Pela análise da Tabela 3.5, pode-se perceber que a Rede Neural, com a configuração aqui utilizada, fornece a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas.

Mantendo-se a mesma proporção de amostras treinamento/generalização da simulação anterior, faz-se agora uma nova simulação, com o algoritmo C4.5 e a análise discriminante, somente com os 266 indivíduos que restam após serem excluídos os 31 indivíduos. Os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, são os seguintes:

Tabela 3.6 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (sem os 31 indivíduos).

Número de Variáveis	C4.5			Análise Discriminante		
	PND	PD	PM	PND	PD	PM
1	82,8	82,3	**82,6**	83,0	82,3	**82,7**
2	82,5	80,9	**81,8**	81,4	91,2	**85,7**
3	77,2	82,1	**79,3**	90,6	87,1	**89,0**

A Tabela 3.7 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM, e utilizando-se o banco sem os 31 indivíduos classificados sempre incorretamente:

Tabela 3.7 – Tabela Comparativa, Valores de PM (sem os 31 indivíduos).

Número de Variáveis	C4.5	Análise Discriminante	Rede Neural
1	82,6	82,7	82,4
2	81,8	85,7	85,3
3	79,3	89,0	93,7

Pela análise da Tabela 3.7, onde os 31 indivíduos não são utilizados no banco, pode-se perceber que a rede neural, com a configuração aqui utilizada, continua fornecendo a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas.

Finalmente, utilizando-se novamente todo o banco de dados, as simulações são feitas garantindo-se que os 31 indivíduos que foram classificados como pertencentes ao grupo de **"indivíduos sempre classificados erroneamente" pertençam sempre ao conjunto de treinamento**. Aplicando-se estes conjuntos no algoritmo C4.5 e na análise discriminante, os resultados obtidos, para 50 conjuntos de treinamento/generalização diferentes, são os seguintes:

Tabela 3.8 – Percentual de Acerto no Algoritmo C4.5 e na Análise Discriminante (com os 31 indivíduos no treinamento).

Número de Variáveis	C4.5			Análise Discriminante		
	PND	PD	PM	PND	PD	PM
1	82,3	83,4	**82,8**	82,6	83,0	**82,8**
2	82,8	81,6	**82,2**	82,2	91,7	**86,6**
3	76,5	81,7	**78,9**	88,5	89,6	**89,0**

A Tabela 3.9 a seguir, apresenta os resultados para os 3 métodos de classificação de padrões que estão sendo comparados, considerando-se apenas os valores de PM:

Tabela 3.9 – Tabela Comparativa, Valores de PM (com os 31 indivíduos no treinamento).

Número de Variáveis	C4.5	Análise Discriminante	Rede Neural
1	82,8	82,8	82,3
2	82,2	86,6	87,1
3	78,9	89,0	91,0

Pela análise da Tabela 3.9, pode-se perceber que a Rede Neural ainda continua fornecendo a melhor classificação entre os 3 métodos avaliados, quando as 3 variáveis (THAL, CA e CP) são utilizadas. Uma das conclusões importantes que podem ser tiradas desses resultados, ao serem comparadas as tabelas 3.7 e 3.9, é que os valores de PM praticamente não se alteram ou se alteram pouco, principalmente os resultados do algoritmo C4.5 e da análise discriminante.

Quando se mantém os 31 indivíduos no conjunto treinamento, busca-se uma alternativa de extrair informação útil deste conjunto de indivíduos. A conclusão que se pode tirar das análises realizadas, é que ter o "grupo de indivíduos sempre classificados erroneamente" dentro do conjunto de treinamento, pode auxiliar na classificação correta de indivíduos presentes no conjunto de generalização. Isto provavelmente pode ser estendido para qualquer banco de dados.

4. Considerações Finais e Conclusões

Os resultados das simulações realizadas no desenvolvimento deste estudo indicaram que a utilização em conjunto de três variáveis (THAL, CA e CP) na rede neural para descrever e estudar o evento em questão (diagnóstico de doença cardíaca), é o que permite uma melhor identificação do evento. A variável "THAL" ("Talassemias") refere-se a uma doença genética que resulta na alteração da quantidade produzida de subunidades de hemoglobina. É um tipo de anemia hereditária. A variável "CA" ("Número de vasos coloridos pela fluoroscopia") refere-se a um tipo de exame denominado "fluoroscopia", que é uma técnica de imagem comumente usada por médicos para obter imagens em tempo real de estruturas

internas de um indivíduo através do uso de um fluoroscópio. Em sua forma mais simples, um fluoroscópio consiste em uma fonte de raios-X e em uma tela fluorescente, entre as quais um indivíduo é colocado. A variável "CP" ("Tipo de dor no peito") refere-se à intensidade de dor no peito, denominada "angina", que é devida ao baixo abastecimento de oxigênio do músculo cardíaco, geralmente devido à obstrução ou espasmos das artérias coronárias.

O resultado obtido na validação da metodologia proposta neste estudo foi considerado bastante satisfatório, visto que, com apenas 3 variáveis (THAL, CA e CP), foi possível estimar o diagnóstico de doença cardíaca com um percentual de acerto muito significativo quando comparado àqueles obtidos por outros autores e encontrados na literatura. Os resultados obtidos por Ho & Chou (2001) e por Hu, Li, Cai & Xu (2004), são da ordem de 83,0 % e 83,5 % de acerto, respectivamente, enquanto que o resultado obtido nos experimentos realizados neste estudo com a metodologia proposta, apresentou 91,0 % de acerto.

Outras técnicas de classificação de padrões conhecidas na literatura foram utilizadas, de forma a comparar seus resultados com os resultados obtidos neste estudo com a Rede Neural. Para tal análise comparativa, foram escolhidos o algoritmo C4.5 e a Análise Discriminante. O resultado apresentado pelo modelo de Rede Neural foi superior em relação aos resultados apresentados por estes modelos.

Acredita-se que os resultados obtidos sejam relevantes e que possam vir a auxiliar às condutas médicas em relação ao diagnóstico de cardiopatia isquêmica, podendo vir a ser úteis como ponto de partida na prevenção e/ou tratamento de doenças cardíacas. Espera-se que os resultados obtidos sirvam também de incentivo para outras iniciativas nesta direção, sendo, por exemplo, estudados também do ponto de vista clínico, de forma a contribuir não só para o diagnóstico, mas também para o prognóstico de doença cardíaca. Como proposta para trabalhos futuros, fica a sugestão de se utilizar a nova gama de informações advindas dos estudos recentes com respeito às células-tronco, com o objetivo de se identificar variáveis que possam propiciar um diagnóstico (e/ou um prognóstico) cada vez mais preciso no que diz respeito às doenças cardíacas.

Referências Bibliográficas

(1) Aha, D.W. (2001). Heart Disease Databases. <http://www.ics.uci.edu/pub/machine-learning-databases/heart-disease/heart-disease.names>. Current: Oct 2001.

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

(3) Bishop, C.M. (1995). *Neural Networks for Pattern Recognittion*. Clarendon Press, Oxford.

(4) Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*. Wiley, New York.

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

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

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

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

(9) Ho, C.S. & Chou, J.S. (2001). Fuzzy ARTRON: A General-purpose Classifier Empowered by Fuzzy ART and Error Back-propagation Learning. *Journal of Information Science and Engineering*, **17**, 683-695.

(10) Hu, Z.H.; Li, Y.G.; Cai, Y.Z. & Xu, X.M. (2004). An Empirical Comparison of Ensemble Classification Algorithms with Support Vector Machines. *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August 2004.

(11) Kwak, N. & Choi, C. (2002). Input Feature Selection for Classification Problems. *IEEE Trans. Neural Networks*, **13**(1), 143-159.

(12) Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, **2**, 164-168.

(13) MacKay, D.J.C. (1992). Bayesian interpolation. *Neural Computation*, **4**(3), 415-447.

(14) Marquardt, D. (1963). An Algorithm for Least Squares Estimation of Nonlinear Parameters. *SIAM J. Appl. Math*., **11**, 431-441.

(15) Medeiros, M.C. & Pedreira, C.E. (2001). What are the effects of forecasting linear time series with neural networks? *Engineering Intelligent Systems*, **9**, 237-242.

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

(17) Principe, J.C.; Euliano, N.R. & Lefebvre, W.C. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley.

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

(19) Setiono, R. & Liu, H. (1997). Neural Network Feature Selector. *IEEE Trans. Neural Networks*, **8**, 654-661.