2
Uma Metodologia para Diagnóstico de Doença Cardíaca

Este capítulo se inicia com a descrição do banco de dados; em seguida é apresentada a proposta de estratégia utilizada para seleção do conjunto de variáveis; e finaliza-se apresentando o modelo de Rede Neural utilizado para diagnosticar doença cardíaca.

2.1
Descrição e Tratamento da Base de Dados

2.1.1
Descrição da Base de Dados

A base de dados de doenças do coração utilizada nesta dissertação ("Heart Disease Database") tem quatro subconjuntos diferentes com relação à origem (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 para cada indivíduo. Contudo, devido ao grande número de dados incompletos na maioria dos atributos, todos os experimentos publicados referem-se à utilização de somente 13 destes, os quais são listados a seguir:

→ Idade (AGE) – variando de 29 a 77 anos;
→ Sexo (SEX) – masculino ou feminino, sendo representados por 1 ou 0, respectivamente;
→ 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.

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

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

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

→ 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.

→ Máxima taxa de batimento cardíaco atingida (THALACH);

→ Angina induzida por exercício (EXANG):

- Valor 1: Sim;
- Valor 0: Não.

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

→ Inclinação da extremidade do segmento ST no exercício (SLOPE):

- Valor 1: Inclinado para cima;
- Valor 2: Plano;
- Valor 3: Inclinado para baixo.

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

→ 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: menos de 50 %, e 50 % ou mais de estreitamento do diâmetro do vaso sangüíneo, sendo representadas por “0” e “1”, respectivamente. Convencionou-se que os indivíduos com valor “0” na classe de saída seriam chamados de “indivíduos não-doentes”, e os indivíduos com valor “1”, de “indivíduos doentes”.

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

São comuns as situações onde nem todas as variáveis são igualmente importantes para a classificação de um determinado evento. Algumas das variáveis podem ser redundantes ou irrelevantes e, mais do que isso, é sabido que encontrar uma combinação apropriada de variáveis explicativas pode melhorar significativamente o desempenho do modelo (Fukunaga, 1972). Esta questão é ainda mais delicada quando as variáveis alimentam um modelo não-linear.

O tratamento das variáveis se dará em dois níveis, primeiramente serão explorados aspectos de redundância e capacidade de transmitir informação ao desfecho, em seguida, já explorando os resultados numéricos, diferentes conjuntos de variáveis serão testados frente ao modelo proposto.

2.1.2
Tratamento da Base de Dados

O tratamento da base de dados utilizada nesta dissertação foi baseado na seleção de variáveis através de *Informação Mútua*. 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) (ver mais detalhes no Anexo A). 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 ← F-{fi}) e entra no conjunto S (S ← {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 a seguir:

Tabela 2.1 – Resultado Preliminar de Seleção e Ordenação de 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

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.

2.2
Uma Proposta de Modelagem para Diagnóstico de Doença Cardíaca

O diagnóstico de doença cardíaca 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”) (ver mais detalhes no Anexo B) 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 nesta dissertação, 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 (variáveis 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 alta probabilidade do indivíduo ser doente e próximos de 0 indicam baixa probabilidade deste ser doente. O ponto de corte adotado para diferenciar alta de baixa probabilidade foi 0,5 e desta forma considerou-se apenas dois tipos de saída: doente, para valores iguais ou acima de 0,5, e não-doente, para valores abaixo de 0,5.

Nas implementações do próximo capítulo, as variáveis foram normalizadas, por razões numéricas, no intervalo [-1;1].

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):

$$\hat{\psi} = \arg\min_{\psi} Q_1(\psi) = \arg\min_{\psi} \sum_{t=1}^{N} (y_t - G(x,\psi))^2 \tag{2.1}$$

Neste ponto 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 e Pedreira, 2001).

Atualmente, diversas metodologias são utilizadas para solucionar o problema de *"overfitting"* (Haykin, 1998). Nesta dissertação, 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 e Pedreira (2001), o problema de estimação passa a ser definido como:

$$\hat{\psi} = \arg\min_{\psi} Q_T(\psi) = \arg\min_{\psi} \sum_{t=1}^{N} (\eta Q_1(\psi) - \phi Q_2(\psi))^2 \tag{2.2}$$

Onde a função de penalização é a soma do quadrado dos parâmetros (α, γ):

$$Q_2(\psi) = \sum_{h=0}^{H} \alpha_h^2 + \sum_{h=0}^{H} \sum_{i=0}^{I} \gamma_{hi}^2 \tag{2.3}$$

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 & 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 nesta dissertação 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).

Normalmente utiliza-se a forma extrema de validação cruzada múltipla, conhecida como o "método deixe um de fora" (*leave-one-out method*, Haykin, 1998 e Bishop, 1995), se o número de exemplos rotulados disponíveis, N, for severamente limitado. Neste método são utilizados N-1 exemplos para treinar o modelo, e este é validado testando-o sobre o exemplo deixado de fora. Deixando de fora cada vez um exemplo diferente para a validação, o experimento é repetido um total de N vezes. O erro quadrado na validação é então a média sobre as N tentativas do experimento (Haykin, 1998, Bishop, 1995). Dessa forma, no caso do banco de dados desta dissertação, o uso do método de *leave-one-out* seria o mais

indicado. Porém, foi feito diferente para o conjunto de dados desta dissertação: para que fosse possível estabelecer uma comparação com pesquisas previamente realizadas e encontradas na literatura, e também aplicar a metodologia proposta nesta dissertação, 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). Assim, 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.