APLICAÇÃO DE REDES NEURAIS PROBABILÍSTICAS À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Alfredo Ricardo de Faria Passos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientadores: Basílio de Bragança Pereira
Amália Faria dos Reis

Rio de Janeiro
Fevereiro de 2010

APLICAÇÃO DE REDES NEURAIS PROBABILÍSTICAS À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Alfredo Ricardo de Faria Passos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

_____

Prof. Basílio de Bragança Pereira, Ph. D.

_____

Drª. Amália Faria dos Reis, D.Sc.

_____

Profª. Laura Silvia Bahiense da Silva Leite, D.Sc.

_____

Dr. Nelson Albuquerque de Souza e Silva, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2010

Passos, Alfredo Ricardo de Faria

Aplicação de Redes Neurais Probabilísticas à classificação do risco de morte de pacientes com síndrome coronariana aguda/ Alfredo Ricardo de Faria Passos. - Rio de Janeiro: UFRJ/COPPE, 2010

XI, 52 p.: il.; 29,7 cm.

Orientadores: Basílio de Bragança Pereira

Amália Faria dos Reis

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2010.

Referências Bibliográficas: p. 48-52.

1. Redes Neurais Probabilísticas. 2. Síndrome Coronariana Aguda. I. Pereira, Basílio de Bragança et al. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

Ao meu pai, Edilmar Passos,
Ao meu irmão, Vitor,
à minha mãe, Icléa,

AGRADECIMENTOS

Ao meu pai, em pensamento, meu herói em vida e símbolo de sabedoria e honestidade.

Ao meu irmão Vitor, pelo exemplo de flexibilidade, e felicidade com a vida.

A minha mãe, Icléa, por me mostrar o caminho da vida e de se estar de bem com a vida.

Ao professor Basílio pelos esclarecimentos e pelo subsídio ao desenvolvimento deste trabalho, assim como pelos preciosos ensinamentos ao longo das disciplinas por mim cursadas.

À professora Amália, pela orientação e precioso subsídio na área clínica e análise crítica deste trabalho ao longo de sua elaboração.

À professora Laura Bahiense pela participação na banca e pelos importantíssimos ensinamentos na área de Otimização ao longo da minha estadia na COPPE.

A minha segunda mãe Cremilda, pelo símbolo de garra, caráter e dedicação às pessoas que ama.

Ao professor Luis Otávio Façanha, amigo e debatedor de temas ligados à econometria no IE UFRJ.

Aos meus saudosos professores do Colégio de Aplicação, pela contribuição acadêmica e me ajudarem na formação como cidadão, em especial Gustavo Bernardo, Cristina Ferreira e João Carlos Cataldo.

Aos professores do IE UFRJ, por me ajudarem a desenvolver o pensamento crítico sobre a sociedade e o meu papel nela.

Ao amigo Márcio, pelo apoio nos momentos difíceis e pelas conversas francas.

A Deus por me conceder inteligência e energia para realizar minha missão no universo.

Resumo da Dissertação apresentada à COPPE / UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APLICAÇÃO DE REDES NEURAIS PROBABILÍSTICAS À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Alfredo Ricardo de Faria Passos

Fevereiro/2010

Orientadores: Basílio de Bragança Pereira
Amália Faria dos Reis

Programa: Engenharia de Produção

A Síndrome Coronariana Aguda continua sendo uma das principais causas de morte no Brasil. Este fato faz com que seja um tema de extrema relevância em saúde pública.

Diversas modelagens têm sido propostas neste sentido, com o objetivo de se identificar variáveis capazes de explicar os possíveis desfechos desta doença.

Este trabalho objetivou aplicar modelos de redes neurais probabilísticas à classificação do risco de morte de pacientes internados com Síndrome Coronariana Aguda. O classificador assim construído foi capaz de predizer com altíssima acurácia os desfechos destes pacientes, em sobreviventes e não-sobreviventes durante o período de internação.

Para tanto utilizaram-se diversas combinações de variáveis explicativas, algumas das quais foram capazes de produzir classificadores com acurácia máxima. Em particular, as variáveis Creatinina, Freqüência Cardíaca na Admissão e Idade foram as que forneceram melhor desempenho entre os classificadores construídos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

APPLICATION OF "PROBABILISTIC NEURAL NETWORKS" TO CLASSIFICATION OF THE RISK OF DEATH OF HOSPITALIZED PACIENT WITH ACUTE CORONARY SYNDROME

Alfredo Ricardo de Faria Passos

February/2010

Advisors: Basílio de Bragança Pereira
Amália Faria dos Reis

Department: Program of Production Engineering

The Acute Coronary Syndrome remains one of the main causes of death in Brazil. This fact makes it an extreme relevant issue of public health.

In this sense several modelings of this problem have been proposed, with the main goal of better identifying variables capable of explaining the possible outcomes of this desease.

This work aimed to apply probabilistic neural networks models to the classification of the risk of death of hospitalized patients with Acute Coronary Syndrome. The classifier so constructed was able to predict with very high accuracy the possible outcomes for these patients, in survivors and non-survivors along the hospitalization period.

For this purpose, several variable combinations were issued, some of them were able to produce maximum accuracy. In particular, Creatinine, Cardiac Frequency in Admission end Age were the ones which provided the best performance for the constructed classifiers.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Redes Neurais Artificiais (RNA) ou simplesmente Redes Neurais são modelos que se espelharam no cérebro humano visando à criação de um modelo que reproduzisse algumas características do funcionamento cerebral, que é altamente complexo, não-linear e de processamento paralelo.

De acordo com HAYKIN (2003), o cérebro humano tem a capacidade de organizar seus elementos constituintes (neurônios) para implementar computações muito mais rápidas do que os computadores digitais de hoje.

Neste sentido, uma Rede Neural é um método que visa reproduzir a forma pela qual o cérebro humano põe em prática uma determinada tarefa ou função de interesse. (HAYKIN, 2003).

Diversas aplicações têm sido dadas a tais modelos, em diversas áreas do conhecimento como: Engenharia (como em CHEN,2008), Economia-Finanças (LO, 1994) e Medicina. Nesta última linha de aplicações, as RNA's têm sido utilizadas como método de apoio ao diagnóstico e prognóstico de várias doenças como em REIS (2007).

As RNA's têm sido abordadas por diversas áreas de conhecimento humano, notadamente Estatística, *Data Mining*, Inteligência Artificial, Ciência da Computação, Neurociências e Máquinas de Aprendizado (*Machine Learning*).

Com o advento de maior capacidade de processamento dos computadores e o aumento progressivo do tamanho dos bancos de dados nas aplicações reais, a Estatística, como campo científico do saber também está mudando para abrigar estas novas realidades (IZENMANM, 2008).

Assim, maior atenção está sendo dada, por parte desta área do saber, a técnicas de descobrimento do conhecimento e informação (a partir de dados), ao invés da permanência do foco estatístico nas abordagens tradicionais. Segundo INZEMAN (2008) a origem de muitas dessas técnicas são puramente algorítmicas, enquanto as técnicas estatísticas mais tradicionais foram derivadas de otimização e do raciocínio probabilístico.

Corroborando com esta realidade, WASSERMAN (2004) constata o fato de que as técnicas e culturas de modelagem originalmente nascidas nas áreas de Ciência da Computação estão cada vez mais sendo adotadas pela comunidade estatística. Em contrapartida, os métodos de inferência tradicionais da estatística (calcados

fundamentalmente na estatística matemática), estão sendo cada vez mais utilizados como ferramental teórico pelos cientistas de Ciências da Computação. Desta forma, fica cada vez mais difícil se atribuir uma técnica de modelagem a uma área do conhecimento específica, como é o caso de Redes Neurais Artificiais (RNA).

Em todos esses casos se deseja averiguar a capacidade preditiva de tais procedimentos e modelos aplicados a bancos de dados específicos de uma dada situação (IZENMAN, 2008). O objetivo, neste campo do saber é extrair as propriedades descritivas e preditivas de grandes bancos de dados.

Este trabalho teve como objetivo criar modelos de Redes Neurais Artificiais (RNA), especificamente na construção de classificadores voltados ao problema do risco de morte por Síndrome Coronariana Aguda (SCA). O modelo de RNA apresentado para tal propósito neste estudo é conhecido como Rede Neural Probabilística, ou *Probabilstic Neural Network* (*PNN*).

Até o momento em que este trabalho estava sendo concluído muitas aplicações de *PNN*'s em problemas relacionados ao prognóstico e diagnóstico de doenças já haviam sido efetuadas, em particular no diagnóstico ou prognóstico de diversos tipos de câncer, como por exemplo em BERRAR e DUBITZKY (2003), GORUNESCU (2005), HUANG e LIAO (2003) e SHAN et all (2002). Muitas dessas aplicações se concentram na área do câncer de mama, tais como em SECRETAN et all (2007) e TIMEMY et all (2009).

Apesar desta vasta aplicação de *PNN*'s nas áreas médicas, não foram encontradas nos bancos de dados LILACS e MEDLINE publicações científicas que tenham aplicado *PNN's* em estudos com pacientes com Síndrome Coronariana Aguda (SCA). Na busca realizada durante o mês de janeiro de 2010 nestes bancos de dados foram utilizadas as palavras-chave "mortalidade", "infarto agudo do miocárdio", "angina instável", "sindrome coronariana aguda", "doença coronariana", "doença das coronárias", "acute myocardial infarction", "unstable angina", "acute coronary syndrome", "coronary disease" versus "rede", "neural", "rede neural artificial", "artificial neural network", "rede neural probabilística" e "probabilistic neural network".

O único trabalho encontrado que mais se aproximava do tema abordado foi o de VOSS (2002), que aplicou *PNN's* na predição da ocorrência de infarto agudo do miocárdio ou morte coronariana durante o seguimento de 10 anos de uma população de

5.159 de homens entre 35 e 64 anos, que participaram do estudo epidemiológico prospectivo PROCAM na Europa.

1.1 CONTEXTUALIZAÇÃO DA SITUAÇÃO-PROBLEMA, RELEVÂNCIA E JUSTIFICATIVA DO ESTUDO

As Doenças Isquêmicas do Coração (DIC), conjuntamente com as Doenças Cerebrovasculares (DCBV) são um tema de grande importância para a saúde brasileira, em função da magnitude dos óbitos por elas causados e registrados no país. Em 2004, as DIC foram responsáveis por 86.791 óbitos (REIS, 2007).

Segundo dados do DATASUS, as doenças do aparelho circulatório foram a principal causa de morte no Brasil, correspondendo a 27,88% dos óbitos no país em 2004 (REIS, 2007).

Ainda, no âmbito destas doenças, as DCBV corresponderam a 31,6% desses óbitos enquanto que as DIC (das quais a Síndrome Coronariana Aguda é uma forma de manifestação) corresponderam a 31,3%. A região sudeste foi a que apresentou, no mesmo ano, a maior concentração dos óbitos tanto devidos a doenças do aparelho circulatório (51,12% dos óbitos devidos a esta causa, no país) como por DIC's (53,4% dos casos). No Estado do Rio de Janeiro, Niterói foi o 3º município com maior mortalidade por DIC em 2004. O presente trabalho utilizará um banco de dados coletado por REIS (2007) e colaboradores composto de pacientes internados com Síndrome Coronariana Aguda no município de Niterói de julho de 2004 a junho de 2005.

Como exposto em REIS (2007), apesar de ter sido verificada no mundo inteiro, e inclusive no Brasil, uma redução das taxas de mortalidade por doenças do aparelho circulatório, esta redução não pôde ser explicada apenas pelo controle dos fatores de risco cardiovascular clássicos ou pela introdução de novas tecnologias como a revascularização miocárdica (seja cirúrgica ou por angioplastia). Portanto, existe a necessidade de se avançar no conhecimento, buscando entender a influência de outros fatores ambientais ou genéticos nesta mortalidade. Face a este contexto, no presente trabalho foram utilizados modelos de Redes Neurais Probabilísticas, como ferramenta na avaliação da influência das variáveis genéticas e clínicas na ocorrência de óbitos por síndrome coronariana aguda (SCA), já que esta metodologia permite a avaliação de

relações complexas (lineares ou não) entre as variáveis e o desfecho (óbito ou não óbito).

Neste sentido a principal justificativa para se adotar modelos de Redes Neurais Artificiais (RNA) na construção de um classificador do risco de morte derivado da Síndrome Coronariana Aguda (uma forma de manifestação das DIC) deve-se ao fato desse tipo de modelo (RNA) mapear um vetor de variáveis de entrada no espaço de categorias possíveis da variável resposta em um dado problema de classificação (que neste trabalho se constituem nas categorias óbito e não-óbito) de maneira não-linear; o que confere a tais modelos a possibilidade de melhor desempenho na classificação dos vetores de entrada quando comparados com casos onde este mapeamento estivesse por exemplo, restrito a formas lineares.

A construção de classificadores, a partir de máquinas de aprendizado como a RNA, que sejam capazes de prever com grande margem de acerto os possíveis desfechos, representa algo de grande utilidade para a sociedade brasileira e mundial, na medida em que tais dispositivos podem ser utilizados como instrumento de apoio na definição do diagnóstico e/ou prognóstico, auxiliando a tomada de decisão médica.

1.2 OBJETIVOS DO ESTUDO

Este trabalho pretende, com a construção do classificador do risco de morte por SCA a partir de uma modelo de Rede Neural Probabilística (RNP), cumprir os seguintes objetivos:

1- criar modelos de Redes Neurais Probabilísticas para a predição do risco de morte por SCA a partir dos classificadores construídos;

2- verificar os subconjuntos de variáveis que, por meio do classificador implementado por RNP, possuem maior poder de explicação da variável desfecho;

3- verificar se, a partir do conjunto de variáveis selecionadas, as variáveis genéticas desempenham um papel importante na classificação do risco de morte;

4- comparar os resultados obtidos, com os obtidos por REIS (2007) e COLLAZO (2009). Por conseguinte se deseja também comparar as metodologias de Redes Neurais Artificiais treinadas por *Backpropagation*, *Support Verctor Machines* e Redes Neurais Probabilísticas no que tange a performance desses métodos no mesmo banco de dados considerado messes estudos.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Este trabalho está organizado em seis capítulos. O primeiro capítulo é constituído da introdução, na qual se contextualiza e se apresenta a relevância do estudo e se explica o motivo da escolha da RNA como método de classificação.

No segundo capítulo é feito um apanhado dos estimadores não-paramétricos de densidades de probabilidade do tipo Núcleo e de classificadores de Parzen.

O capítulo três disserta sobre modelos Rede Neurais Probabilísticas, detalhando o esquema de funcionamento de um classificador implementado por meio deste tipo de RNA quando o estimador do Núcleo é uma função Gaussiana.

O capítulo quatro detalha a metodologia do estudo e da implementação das redes neurais probabilísticas no problema de classificação do risco de morte por SCA.

O capítulo cinco relata os resultados encontrados.

Por fim, o capítulo seis tece as conclusões acerca do estudo e considerações sobre trabalhos futuros.

2 ESTIMADORES NÂO-PARAMÉTRICOS DO TIPO NÚCLEO E CLASSIFICADORES DE PARZEN

Este capítulo se propõe a fazer um breve apanhado dos estimadores não-paramétricos conhecidos como estimadores de núcleo. Na primeira parte é feita uma revisão de aplicações destes estimadores na estimação de densidades, se verificando algumas de suas principais propriedades. Na segunda parte mostra-se como estes estimadores podem ser utilizados na construção de modelos de regressão não paramétrica. Na terceira parte se verá como estimadores de núcleo podem ser usados na construção de classificadores.

2.1 ESTIMADORES NÚCLEO NA ESTIMAÇÃO DE DENSIDADES DE PROBABILIDADE

Os estimadores não-paramétricos de densidade que vieram a ser chamados estimadores do tipo núcleo (ou *kernel estimators*) foram propostos inicialmente por PARZEN (1962), sendo também chamados de estimadores do tipo Janela de Parzen.

Basicamente, o método não–paramétrico de estimação de densidades de probabilidade consiste em se calcular tal estimativa, pela superposição de janelas, réplicas de mesma função (a ser chamada também de função núcleo ou *kernel*).

Formalmente, um estimador *kernel* de densidade é definido por:

$$(1)\ \hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

Onde:

K: é a função núcleo utilizada;

x_i: é cada dado (univariado) observado na amostra;

h: largura da banda, ou parâmetro de suavização.

Qualquer função suave K, tal que $K(x) \geq 0$, $\int K(x)dx = 1$ e $o_K^2 = \int x^2 K(x)dx > 0$ pode ser utilizada como função núcleo. Diversos tipos de funções respeitam essas propriedades, entre elas a função *kernel* Gaussiana $K(x) = 2\pi^{-1/2}e^{-x^2/2}$.

Outra função *kernel* muito utilizada é conhecida como *kernel* de Epanechnikov, dada pela expressão $K(x) = \left\{ \begin{array}{ll} \dfrac{\frac{3}{4}\left(1 - \frac{x^2}{5}\right)}{\sqrt{5}} & ; \text{se } |x| < 5 \\ 0 & ; \text{caso contrário} \end{array} \right\}$.

Segundo SPECHT (1990) e PARZEN (1962) Os estimadores *kernel* têm, como importante propriedade o fato de que são consistentes, convergindo para a verdadeira densidade sendo estimada, conforme o tamanho da amostra cresce.

Esta convergência implica que a verdadeira densidade será aproximada de forma suave. Além disso, esses estimadores convergem mais rápido para as verdadeiras densidades do que, por exemplo, os estimadores do tipo histograma.

Em geral, as larguras de banda são escolhidas por tentativa e erro. Assim como no caso univariado, a determinação desses parâmetros é de fundamental importância para a estimação de densidades multivariadas, por controlarem o balanceamento entre viés e variância deste estimador (e por conseguinte, seu risco), sendo assim parâmetros de suavização.

Os resultados originalmente encontrados por PARSEN (1962) foram posteriormente estendidos para o caso multivariado onde a densidade conjunta pode ser expressa como o produtório de funções *kernel* univariadas (assumindo-se, para tanto, que as variáveis aleatórias sejam independentes e identicamente distribuídas).

De forma genérica, dada uma função *kernel* K, vetores p-dimensionais $\mathbf{x_i} = (x_{i1},...,x_{ip})$ e um vetor de amplitudes de banda $\mathbf{h} = (h_1,...,h_p)$, o estimador *kernel* de densidades multivariadas é dado por:

$$(2)\ \hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

Onde: $K_h(\mathbf{x} - \mathbf{x}_i) = \dfrac{1}{n h_1 h_2 ....h_p}\left\{\prod_{j=1}^{p} K\left(\dfrac{x_j - x_{ij}}{h_j}\right)\right\}$

Para o caso da função *kernel* ser gaussiana, o estimador toma a forma:

$$(3)\ \hat{f}_n(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{N_K}\sum_{i=1}^{N_K} exp\left[-\frac{(\mathbf{x} - \mathbf{x}_{k_i})(\mathbf{x} - \mathbf{x}_{k_i})}{2\sigma^2}\right]$$

Costuma-se, por simplicidade, adotar como largura de banda para cada variável $h_j = hs_j$, onde s_j é o desvio padrão da j-ésima variável aleatória. Com esta simplificação, só há uma única largura de banda a ser escolhida.

2.2 CLASSIFICADORES DE PARZEN

Um classificador (ótimo) de Bayes opta por alocar um vetor **x** de variáveis em uma categoria k, em detrimento das demais se:

(4) $\pi_k L_k f_k(\mathbf{x}) > \pi_j L_j f_j(\mathbf{x})$, para todo $j \neq k$

Onde: π_k é uma probabilidade a priori de ocorrência de padrões da categoria k;

π_j é uma probabilidade a priori de ocorrência de padrões da categoria j;

L_k é o custo associado à decisão de que **x** pertence à clase k, quando pertence realmente à classe j, segundo a função de perda adotada L_k;

L_j é o custo associado à decisão de que **x** pertence à classe j, quando pertence realmente à classe k, segundo a função de perda adotada L_j;

Quando há uma igualdade entre esses dois termos, o classificador pode decidir arbitrariamente onde alocar **x**.

Pode-se demonstrar que um classificador assim construído minimiza o Risco de Bayes -como em RIPLEY (2004) -.

Da regra apresentada acima, por BISHOP (1995) e SPECHT (1990) se conclui que a fronteira de decisão de Bayes é definida para todos os vetores **x** que satisfazem:

(5) $f_A(\mathbf{x}) = k\, f_B(\mathbf{x})$

Com $k = \dfrac{\pi_B L_B}{\pi_A L_A}$

SPECHT (1990) apresenta como principal característica dessa fronteira de decisão de Bayes o fato de ser ótima de Bayes. Como nota o autor, a superfície de decisão de Bayes pode ser arbitrariamente complexa uma vez que não se faz nenhuma restrição sobre as funções densidades de probabilidade $f_A(\mathbf{x})$ e $f_B(\mathbf{x})$ a não ser o fato de serem não-negativas e integráveis à unidade.

O ponto central na igualdade apresentada no item (5) é que as probabilidades a priori, assim como as densidades das populações nas classes ou categorias, necessitam

ser determinadas. Com relação às probabilidades a priori, as mesmas ou já são conhecidas (por representarem, por exemplo, um grau de crença ou conhecimento prévio a cerca do problema) ou podem ser estimadas a partir dos dados (por exemplo considerando- se cada uma dessas probabilidades como a freqüência relativa observada dessas categorias em um grande banco de dados).

Portanto é a estimação das densidades de probabilidade das populações em cada classe que é o mais crucial para o desempenho do classificador, pois a precisão das fronteiras de decisão é muito sensível à maneira pela qual essas densidades são estimadas (SPECHT (1990) e PATTERSON (1996)).

Contudo, na maior parte dos problemas de classificação reais as densidades de probabilidade são desconhecidas, e a única fonte de informação disponível geralmente são os padrões observados, os quais se deseja classificar nas diversas categorias envolvidas no problema. A esse respeito, SPECHT, (1990) sugere a estimação de tais densidades a partir dos dados disponíveis. Nas palavras do autor: *"However, if the probability densities of the patterns in the categories to be separated are unknown, and all that is given is a set of training patterns (training samples), the it is these samples which provide the only clue to the unknown underlying probability densities".* (1990, p.110).

Um classificador, operacionalizado segundo o item (4) e que estima as densidades da população em cada classe segundo o método da Janela de Parzen é um Classificador de Parzen e toma decisões de classificação quanto a se um vetor de entrada **x** pertence a um padrão (categoria) k ou um padrão j após se calcular (estimar) as densidades de probabilidade $f_K(\mathbf{x})$ e $f_j(\mathbf{x})$ pelos métodos não-paramétricos de autoria do citado autor (PARZEN,1962), da seguinte forma :

$$(6)\ d(\mathbf{x}) = \theta_K, \text{ se } \pi_K L_K \hat{f}_K(\mathbf{x}) > \pi_j L_j \hat{f}_j(\mathbf{x})$$

$$(7)\ d(\mathbf{x}) = \theta_j, \text{ se } \pi_K L_K \hat{f}_K(\mathbf{x}) < \pi_j L_j \hat{f}_j(\mathbf{x})$$

Onde: $\hat{f}_K(\mathbf{x})$ e $\hat{f}_j(\mathbf{x})$ são as densidades dos padrões K e j estimadas pelo método do Núcleo.

Alternativamente, pode-se adotar uma regra de decisão mais simplificada, quando se atribui, a todo tipo de classificação equivocada um custo idêntico e unitário. Isso leva ao critério:

(8) $d(\mathbf{x}) = \theta_K$, se $\pi_K \hat{f}_K(\mathbf{x}) > \pi_j \hat{f}_j(\mathbf{x})$

(9) $d(\mathbf{x}) = \theta_j$, se $\pi_K \hat{f}_K(\mathbf{x}) < \pi_j \hat{f}_j(\mathbf{x})$

Portanto, trata-se de um classificador de Bayes, só que implementado de forma não-paramétrica quanto às referidas densidades. Quando estas densidades são estimadas por funções núcleo Gaussianas, a regra de decisão deste classificador passa a ser, para o caso de vetores **x** p-dimensionais:

(10) $d(\mathbf{x}) = \theta_K$, se $\pi_K \frac{1}{(2\pi)^{p/2}\sigma^p} \frac{1}{N_K} \sum_{i=1}^{N_K} exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{k_i})'(\mathbf{x}-\mathbf{x}_{k_i})}{2\sigma^2}\right] > \pi_j \frac{1}{(2\pi)^{p/2}\sigma^p} \frac{1}{N_j} \sum_{i=1}^{N_j} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{j_i})'(\mathbf{x}-\mathbf{x}_{j_i})}{2\sigma^2}\right]$

(11) $d(\mathbf{x}) = \theta_j$, se $\pi_K \frac{1}{(2\pi)^{p/2}\sigma^p} \frac{1}{N_K} \sum_{i=1}^{N_K} exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{k_i})'(\mathbf{x}-\mathbf{x}_{k_i})}{2\sigma^2}\right] < \pi_j \frac{1}{(2\pi)^{p/2}\sigma^p} \frac{1}{N_j} \sum_{i=1}^{N_j} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{j_i})'(\mathbf{x}-\mathbf{x}_{j_i})}{2\sigma^2}\right]$

Onde: i: representa o número de padrões

N_K : o número de padrões na categoria K;

N_j : o número de padrões na categoria j;

$\mathbf{x}_{ki}$: é todo vetor de entrada da amostra que pertença à categoria K;

$\mathbf{x}_{kj}$: é todo vetor de entrada da amostra que pertença à categoria j;

σ : parâmetro de suavidade da função-núcleo gaussiana;

p: dimensão do vetor **x**.

Quando se admite se tomar as probabilidades priori como sendo a freqüência relativa de cada categoria na amostra, isto é, $\pi_K = \frac{N_K}{N}$ e $\pi_j = \frac{N_j}{N}$, a regra se reduz a:

(12) $d(\mathbf{x}) = \theta_K$, se $\sum_{i=1}^{N_K} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{k_i})'(\mathbf{x}-\mathbf{x}_{k_i})}{2\sigma^2}\right] > \sum_{i=1}^{N_j} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{j_i})'(\mathbf{x}-\mathbf{x}_{j_i})}{2\sigma^2}\right]$, para $k \neq j$

(13) $d(\mathbf{X}) = \theta_j$, se $\sum_{i=1}^{N_K} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{k_i})'(\mathbf{x}-\mathbf{x}_{k_i})}{2\sigma^2}\right] < \sum_{i=1}^{N_j} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{j_i})'(\mathbf{x}-\mathbf{x}_{j_i})}{2\sigma^2}\right]$, para $k \neq j$

SPECHT (1990) apresenta o Classificador de Parzen assumindo esta hipótese sobre as prioris, porém deixa claro no mesmo artigo que outras formas de determinação das prioris também são possíveis de ser adotadas. Mais ainda, o autor, no mesmo trabalho, sugere também outras aproximações das densidades (pelo estimador do Núcleo) que não a Gaussiana.

É importante notar que a equação (13) implica que o conjunto inteiro de treinamento deve ser armazenado na memória do computador. A mesma também

implica que a quantidade de computação necessária para se classificar um ponto é proporcional ao conjunto de treinamento. Assim, aplicações de classificação com *PNN*'s a partir de grandes bases de dados são intensivas no uso de memória principal do computador.

SPECHT (1990) apresenta como principal característica de toda fronteira de decisão formada a partir de classificadores de Parzen, o fato de ser assintoticamente ótima de Bayes. Assim, a superfície (ou hiper superfície) de decisão assim formada assintoticamente se aproxima de uma superfície (ou hiper superfície) ótima de Bayes, no sentido de minimização do risco de Bayes em problemas de classificação de padrões.

SPECHT (1967) mostra que, no caso da função *kernel* ser uma função gaussiana multivariada, a fronteira de decisão de um classificador implementado por meio de uma rede PNN tende a um hiperplano conforme $\sigma \to \infty$ e para uma superfície altamente não-linear conforme $\sigma \to 0$.

2.3 ESTIMADORES NÚCLEO NA ESTIMAÇÃO DE REGRESSÕES NÃO-PARAMÉTRICAS

2.3.1 Regressão Nadayara-Watson

Estimadores do tipo Núcleo também podem ser usados na estimação de modelos de regressão não-paramétricos, procedendo a uma média ponderada dos valores amostrais da variável a ser explicada, Y. Este procedimento é conhecido como regressão Nadayara-Watson.

Assim, pode-se definir a citada regressão como:

$$(14)\ \hat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

Onde os pesos w_i atribuídos a cada observação Y_i são determinados por:

$$(15)\ w_i(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}$$

3 REDES NEURAIS PROBABILÍSTICAS (*PNN*) E REGRESSÃO GERAL POR REDES NEURAIS (*GRNN*)

3.1 REDES NEURAIS PROBABILÍSTICAS

As Redes Neurais Probabilísticas (ou *Probabilistic Neural Networks*, *PNN*) surgiram com os trabalhos pioneiros de Donald F. Specht (1988, 1990). Basicamente trata-se de uma implementação, via uma Rede Neural Artificial (RNA) de um mecanismo Classificador de Parzen (classificador assintoticamanete ótimo de Bayes, com as densidades de probabilidade da população em cada classe, estimadas pelo método da Janela de Parzen).

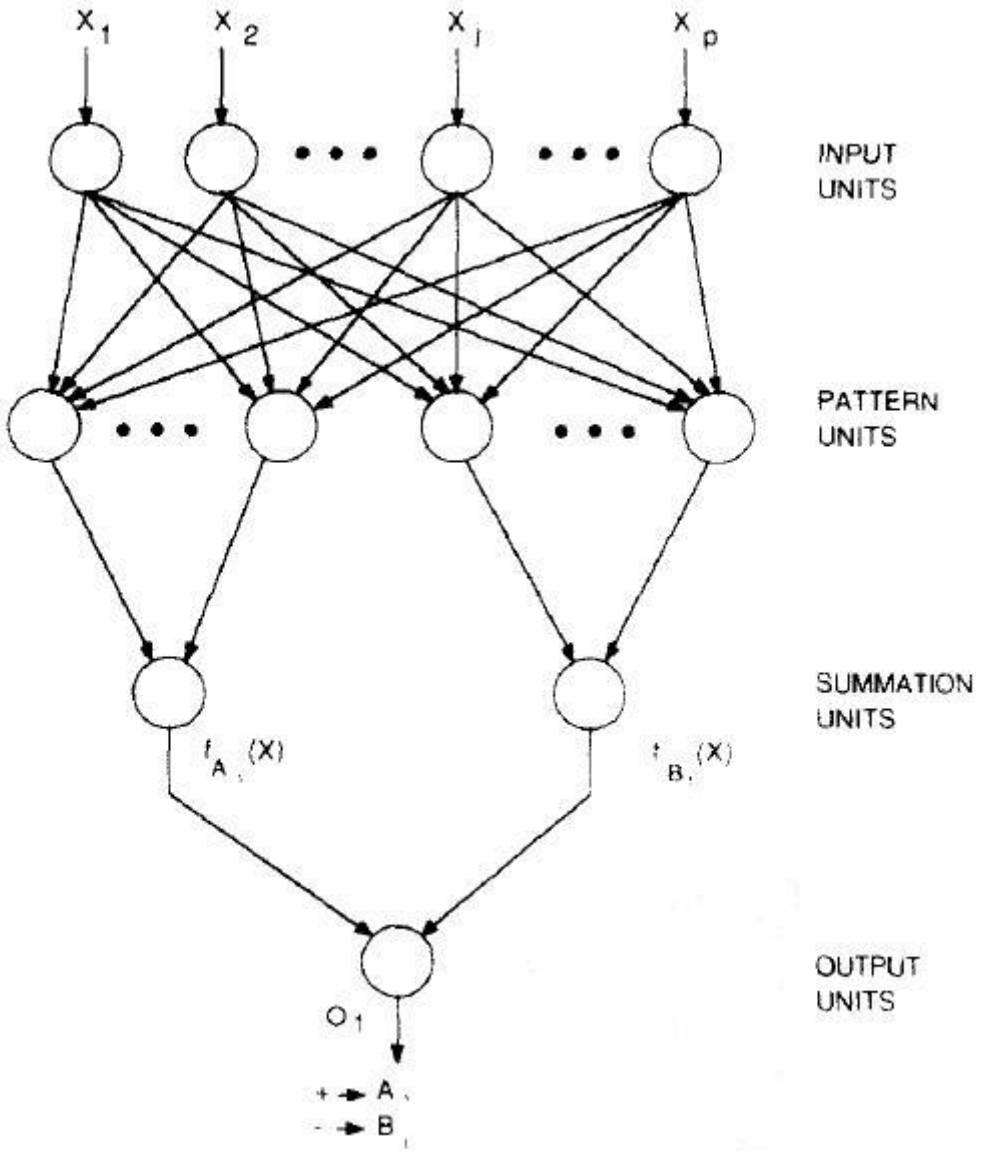
Como tal, a *PNN* decide quando um vetor de realizações de variáveis explicativas **x** pertence à classe k, ou a uma outra classe j qualquer segundo as relações apresentadas nos itens (6) e (7), do capítulo anterior.

As descrições das operações efetuadas em cada camada da *PNN*, a seguir apresentadas, referem-se ao caso da função núcleo ser gaussiana. Pode-se contudo se utilizar qualquer outra função núcleo na *PNN*, mas a mudança desta função implica em outras funções de ativação na rede. Como já adiantado anteriormente, Specht (1990) apresenta diversos outras funções *kernel* que podem ser utilizadas no método de estimação de densidades da Janela de Parzen, conjuntamente com a função de ativação correspondente a cada uma delas, para uma implementação via *PNN*.

Antes de se apresentar os dados à rede, deve-se proceder a uma etapa de normalização dos dados. Seja **x** um vetor de variáveis de entrada a ser apresentado à rede para seu treinamento, SPECHT (1990) sugere normalizá-lo de forma a que fique com comprimento unitário.

A arquitetura (ou topologia) da rede neural proposta originalmente por SPECHT (1990) para tal propósito é ilustrada nas figuras a seguir:

Figura 1: Arquitetura geral de uma *Probabilistic Neural Networks* (*PNN*)



Fonte: SPECHT (1990)

Na Figura 1 pode-se constatar que a arquitetura originalmente proposta por SPECHT (1990) de uma *PNN* é composta de quatro camadas. A primeira camada contém os nós de entrada, que correspondem às variáveis explicativas dos possíveis desfechos do problema de classificação (*input units* na Figura 1). Cada nó apresenta por vez um componente do vetor p-dimensional **x**, observado na amostra e utilizado como padrão no treinamento da rede. Esta camada tem como papel simplesmente prover os dados de entrada (componentes do vetor **x**) aos neurônios da camada seguinte.

Esses últimos são também chamados de unidades de padrão (*pattern units* na Figura 1 acima). A *PNN* possui uma arquitetura específica pelo fato de existirem tantas

unidades de padrão na segunda camada como de vetores de entrada no conjunto de treinamento. Deste modo, cada unidade de padrão armazena um vetor do conjunto de treinamento, assim como a sua verdadeira classe.

Essas unidades representam portanto, os vetores do conjunto de treinamento onde serão centradas funções *kernel* utilizadas no método de estimação de densidades de probabilidade conhecido como método da Janela de Parzen. Assim, cada função *kernel* utilizada na estimação não –paramétrica das densidades de probabilidade de **X** em cada classe terá como média um vetor observado no conjunto de treinamento.

As conexões que comunicam os nós na camada de entrada com os neurônios de padrões na camada imediatamente seguinte são ponderadas por pesos sinápticos, de tal forma que, a cada neurônio i da segunda camada (camada de unidades de padrão) existe um vetor de pesos $\mathbf{W_i}$ a ele associado, ligando este neurônio i a todos os nós da camada de entrada.

Deve-se ressaltar aqui que em muitas arquiteturas alternativas de *PNN* se considera que os vetores de pesos $\mathbf{W_i}$'s possuem todos os componentes unitários (como é o caso, por exemplo das arquiteturas alternativas propostas por ANÔNIMO (2009), DTREG *User's Manual* (2009) e WASSERMAN (1993), de tal forma que os sinais que chegam a cada unidade de padrão são simplesmente as componentes do vetor **x** do conjunto de treinamento (normalizado). Caso se assuma vetores de pesos $\mathbf{W_i}$'s diferentes do mencionado, deve-se proceder também a sua normalização, tal como efetuado com os vetores **x** de variáveis de entrada.

Seja qual for à forma como se especifica os vetores de pesos citados, os mesmos não sofrem alteração ao longo do treinamento da *PNN* (ao menos na bibliografia consultada), permanecendo fixos (PATTERSON, 1996, é o único autor que menciona a possibilidade dos pesos $\mathbf{W_i}$ serem ajustáveis, mas não incorpora essa possibilidade na rotina de treinamento de uma *PNN*).

Cada unidade de padrão procede então à operação $\left(\mathbf{x}-\mathbf{x}_{\mathbf{k}_i}\right)^T\left(\mathbf{x}-\mathbf{x}_{\mathbf{k}_i}\right)$, onde novamente $\mathbf{x}$ é um vetor apresentado à rede no treinamento e $\mathbf{x}_{\mathbf{k}_i}$ é o vetor da amostra que é armazenado (conjuntamente com a sua verdadeira classe) na unidade de padrão i. Após essa operação cada unidade de padrão aplica o resultado da operação anterior em uma função de ativação exponencial, obtendo assim, $\exp -\left[\frac{\left(\mathbf{x}-\mathbf{x}_{\mathbf{k}_i}\right)^T\left(\mathbf{x}-\mathbf{x}_{\mathbf{k}_i}\right)}{2\sigma^2}\right]$.

A seguir, as unidades de padrão mandam esses resultados para as unidades de soma (*summation units*, na Figura 1). Existem em uma *PNN* tantas unidades de soma quanto classes no problema abordado. A Figura 1 apresenta o esquema proposto por SPECHT (1990) para o caso de apenas duas classes (A e B), tendo assim somente dois nós de soma, uma para a classe A e outro para a classe B.

Deste modo, as unidades de padrão só se comunicam com apenas um nó de soma, correspondente à verdadeira classe dos vetores armazenados nas primeiras. Os nós da camada de soma então procedem à soma desses resultados. Assumindo-se que existam N padrões a serem apresentados à rede no conjunto de treinamento, e que desses N padrões N_k correspondam à classe k e N_j correspondam à classe j(assumindo-se por simplicidade somente duas classes no problema), então N_k unidades de padrão irão enviar seus sinais a uma das unidades de soma. As outras N_j unidades de padrão irão enviar seus sinais à outra unidade de soma.

No caso apresentado a título de exemplo por Specht (1990) na Figura 1, os resultados das operações efetuadas nas duas unidades de soma são $\sum_{i=1}^{N_A} \exp - \left[\frac{\left(\mathbf{x} - \mathbf{x}_{A_i}\right)^T \left(\mathbf{x} - \mathbf{x}_{A_i}\right)}{2\sigma^2} \right]$ e $\sum_{i=1}^{N_B} \exp - \left[\frac{\left(\mathbf{x} - \mathbf{x}_{B_i}\right)^T \left(\mathbf{x} - \mathbf{x}_{B_i}\right)}{2\sigma^2} \right]$.

Finalmente, os resultados obtidos nas unidades de soma são enviados às unidades de saída, que irá comparar os resultados recebidos e decidir por alocar **x** em uma das classes possíveis.

SPECHT (1990) apresenta uma arquitetura possível para tanto, no caso de apenas duas categorias (A e B), como ilustrado na Figura 2, abaixo.

Nesta Figura, há apenas um único nó de saída, que somará as quantidades calculadas $\sum_{i=1}^{N_A} \exp - \left[\frac{\left(\mathbf{x} - \mathbf{x}_{A_i}\right)^T \left(\mathbf{x} - \mathbf{x}_{A_i}\right)}{2\sigma^2} \right]$ e $\sum_{i=1}^{N_B} \exp - \left[\frac{\left(\mathbf{x} - \mathbf{x}_{B_i}\right)^T \left(\mathbf{x} - \mathbf{x}_{B_i}\right)}{2\sigma^2} \right]$. Porém, antes de efetuar propriamente esta operação de soma, submete a ultima quantidade a um produto da mesma com um termo $C = -\frac{\pi_B L_B}{\pi_A L_A} . \frac{N_A}{N_B}$. Como se verifica, o termo C é a razão das probabilidades priori, dividida pela razão das amostras em cada classe, multiplicada pela taxa de perdas.

Cada unidade de saída então irá somar $\hat{f}_A$ com $C\hat{f}_B$. Esse resultado será aplicado a uma função de ativação do tipo degrau (ou *Step*), com um valor de corte (*threshold*) igual a zero. Caso o valor da função de ativação seja +1 o classificador irá classificar **x** na classe A do exemplo, caso seja -1, o classificador de Parzen assim construído classificará o vetor na classe B.

Analisando-se mais detalhadamente estas últimas operações, pode-se compreender porque a *PNN* é um classificador de Parzen: assumindo-se que as estimativas das densidades de probabilidade da população em cada classe sejam dadas respectivamente por $\hat{f}_A$ e $\hat{f}_B$, e que as funções *kernel* utilizadas nestas estimações sejam gaussianas, então, sabe-se que os estimadores dessas densidades são dados por

$$\hat{f}_A(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{N_A}\sum_{i=1}^{N_A}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{A_i})^T(\mathbf{x}-\mathbf{x}_{A_i})}{2\sigma^2}\right] \text{ e}$$

$$\hat{f}_B(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{N_B}\sum_{i=1}^{N_B}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{B_i})^T(\mathbf{x}-\mathbf{x}_{B_i})}{2\sigma^2}\right] \text{ respectivamente,}$$

então o classificador implementado pela *PNN* apresentada irá decidir por alocar um vetor **x** na classe A se $\pi_A L_A \hat{f}_A(\mathbf{x}) > \pi_B L_B \hat{f}_B(\mathbf{x})$. Da mesma forma irá optar por alocar **x** na classe B se o sinal da desigualdade for invertido.

Com algumas manipulações algébricas é fácil verificar que a *PNN* alocará **x** à classe A se:

$$(16)\ \sum_{i=1}^{N_{A_i}}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{A_i})^T(\mathbf{x}-\mathbf{x}_{A_i})}{2\sigma^2}\right] - \frac{\pi_B}{\pi_A}\frac{L_B}{L_A}\frac{N_A}{N_B}\sum_{i=1}^{N_B}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{B_i})^T(\mathbf{x}-\mathbf{x}_{B_i})}{2\sigma^2}\right] > 0$$

De modo análogo, alocará **x** na classe B se o resultado da operação for negativo, ou seja:

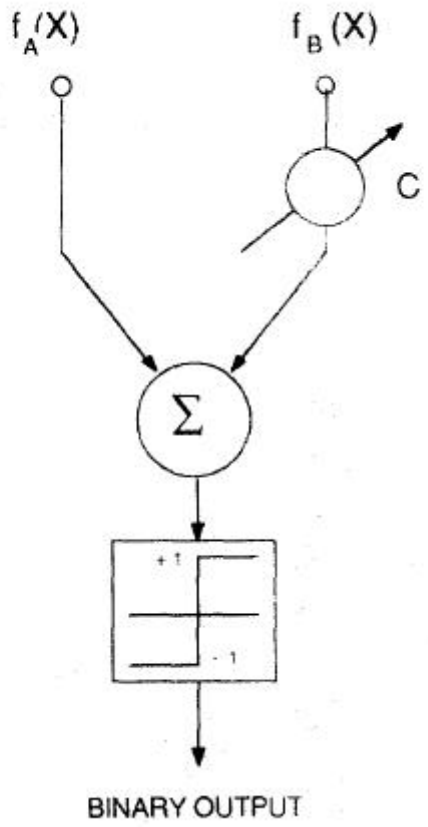
6.12)

$$(17)\ \sum_{i=1}^{N_{A_i}}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{A_i})^T(\mathbf{x}-\mathbf{x}_{A_i})}{2\sigma^2}\right] - \frac{\pi_B}{\pi_A}\frac{L_B}{L_A}\frac{N_A}{N_B}\sum_{i=1}^{N_B}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_{B_i})^T(\mathbf{x}-\mathbf{x}_{B_i})}{2\sigma^2}\right] < 0$$

Logo a arquitetura proposta por Specht implementa de fato um classificador de Parzen. Note-se que nesta estrutura proposta, que o próprio valor de corte (*threshold*) da

função de ativação (do tipo degrau) forçará o classificador assim construído a optar por alocar **x** em uma das classes existentes, não existindo aqui, portanto, a possibilidade do classificador ficar em dúvida.

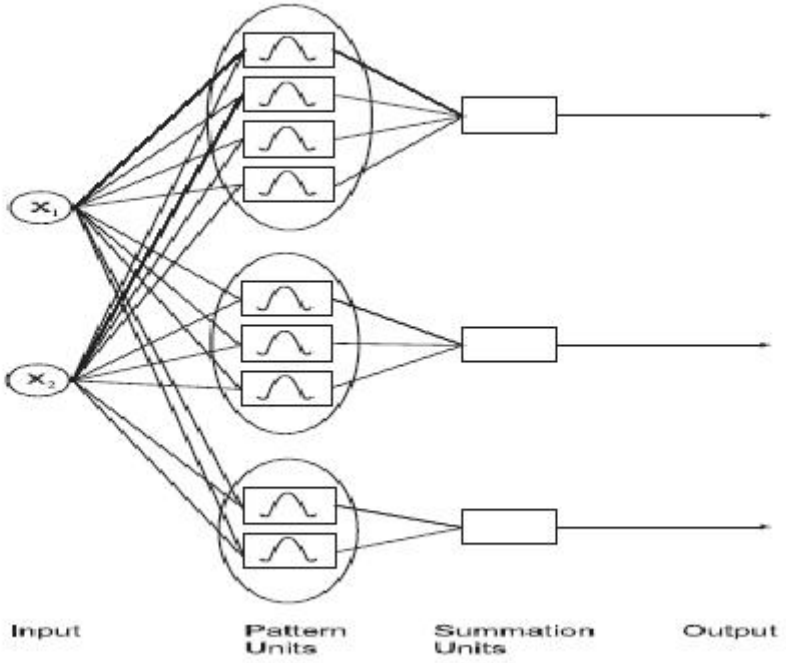
Figura 2: A unidade de saída da *PNN* segundo SPECHT (1990):



Fonte: SPECHT (1990)

Outras arquiteturas também são capazes de implementar uma *PNN*. PEREIRA e RAO (2006), apresentam a seguinte topologia para uma *PNN*, como na figura abaixo:

Figura 3: Topologia alternativa de uma *PNN*:



Fonte: PEREIRA E RAO (2006).

Na topologia apresentada por estes autores, o resultado das operações nas unidades de soma (após as considerações dos custos de classificação equivocada e das probabilidades prioris) são diretamente comparados, e o maior valor corresponde à classe que a *PNN* irá alocar um dado vetor **x**.

Como se pode notar, uma rede neural do tipo *PNN* funciona de forma semelhante a um classificador baseado no critério de vizinhança "dos k mais próximos" (ou *K Nearest Neighborhood*), sendo as diferenças mais importantes, os seguintes fatos: a rede *PNN* generaliza o critério de classificação baseado no conceito "dos K mais próximos" por considerar todos os pontos do conjunto de treinamento, computando-se a distância radial de um ponto (vetor do conjunto de treinamento) considerado em relação a todos os outros pontos deste conjunto; cada distância radial computada é ponderada por um peso, sendo cada peso representado por funções *kernel*.

Portanto, uma rede *PNN* operacionaliza um classificador baseado no conceito de vizinhança ponderada (*weighted neighborhood*), sendo os pesos oferecidos por *kernels*.

3.1.1) Escolha dos desvios-padrões da densidade multivariada estimada por funções *kernel*.

Como apresentado no capítulo 3, a escolha apropriada dos termos relacionados com o grau de suavização das funções *kernel* é fundamental para o bom desempenho do classificador de Parzen, na medida em que são esses termos que controlam o *trade-off* entre viés e variância destes estimadores. Mais especificamente o formato da superfície de decisão depende da escolha destes termos.

No caso de se adotar como função *kernel* densidades normais multivariadas, esses termos correspondem ao desvio-padrão de cada variável na função *kernel.* Esses desvios-padrão devem ser estimados empiricamente, visando-se minimizar uma estimativa do risco do classificador assim considerado.

3.1.2) Características da *PNN*

De acordo com SPECHT (1990) algumas das principais vantagens do *PNN* sobre redes *Multi Layer Perceptron* treinadas por *Backpropagation* são:

-O treinamento da rede *PNN* é fácil e praticamente instantâneo, sendo muito mais rápido do que o de redes treinadas por *Backpropagation*. Specht reporta experimentos nos quais o treinamento do *PNN* foi duzentas mil vezes mais rápido do que a de redes treinadas pelo segundo algoritmo. Na visão de SPECHT (1990) essa é uma das vantagens mais importantes de uma rede *PNN* sobre redes treinadas por *Backpropagation*.

-Experimentos efetuados por SPECHT (1990) e WASSERMAN (1993) revelaram sensibilidade, especificidade e acurácia, dos classificadores implementados por uma rede *PNN*, bem superiores às obtidas por redes neurais treinadas por *Backpropagation*.

-A rede *PNN* opera completamente em paralelo, sem necessidade de *feedback* de neurônios individuais para neurônios em camadas anteriores.

Além disso, outras características importantes do *PNN* são:

- O formato das superfícies de decisão de classificadores estruturados a partir de tais redes neurais pode ser tornado tão simples ou tão complexo quanto se queira, bastando para isso se alterar o vetor de parâmetros correspondente aos desvios-padrões da distribuição.

- A superfície de decisão implementada por *PNN* é assintoticamente ótima de Bayes - o que significa, de acordo com WASSERMAN (1993), convergência garantida para um classificador de Bayes se um número suficiente de padrões de treinamento for oferecido à rede -.

- A *PNN* permite treinamento incremental de forma rápida, pois pode-se incorporar à rede novos exemplos de treinamento sem dificuldades;

- De acordo com WASSERMAN (1993) e SPECHT (1990) classificadores implementados por meio de uma *PNN* são robustos a exemplos com muito ruído, onde amostras errôneas são toleradas. Além disso, σ pode ser tornado menor conforme o tamanho da amostra aumenta, sem necessidade de re-treinamento.

3.2) Redes Neurais de Regressão Generalizada

Esta seção apresenta uma rede neural similar a *PNN*, porém com a finalidade de prover estimativas de variáveis contínuas, implementando assim uma regressão não-paramétrica com estimadores de densidades de probabilidade do tipo Janela de Parzen.

Tal procedimento é conhecido por Redes Neurais de Regressao Generalizada, ou *Generalized Regression Neural Network* (ou abreviadamente *GRNN*) também de autoria de SPECHT (1991).

A citada rede neural implementa uma regressão não-paramétrica de uma variável resposta (ou vetor de variáveis-resposta) $\mathbf{y}$ dado um conjunto de observações de um vetor de variáveis de entrada $\mathbf{x}$. A apresentação a seguir será feita considerando-se y como uma única variável, embora os mesmos resultados sejam extensíveis ao caso de y possuir dimensão maior que um.

A técnica de regressão geral consiste em se determinar o valor esperado da variável resposta (y) condicionado a uma realização $\mathbf{x}$ do vetor de variáveis aleatórias explicativas $\mathbf{X}$, ou seja, $E[y/\mathbf{x}]$. Essa expectância condicionada, por sua vez, pode ser expressa através da densidade conjunta $f(\mathbf{x}, y)$ como:

$$(18)\ E[y/\mathbf{x}] = \frac{\int_{-\infty}^{+\infty} y f(\mathbf{x}, y) dy}{\int_{-\infty}^{+\infty} f(\mathbf{x}, y) dy}$$

Onde $f(\mathbf{x}, y)$ é a densidade de probabilidade conjunta de $\mathbf{x}$ e y. Nas aplicações onde esta densidade conjunta não é conhecida, deve-se estimá-la a partir de dados

observados, por algum método. No caso da *GRNN*, o estimador adotado, tal como na *PNN* é o estimador da janela de Parzen.

O estimador de densidade $\hat{f}(\mathbf{x}, y)$ por este método, como visto, é baseado nos valores amostrais $\mathbf{x}_i$ e y_i. Adotando-se funções *kernel* gaussianas, o estimador é determinado por:

$$19)\ \hat{f}(\mathbf{x}, y) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^{(p+1)}} \cdot \frac{1}{n}\sum_{i=1}^{n} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{2\sigma^2}\right].\exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right]$$

Substituindo-se a expressão do estimador de densidade na expressão de valor esperado condicionado, obtêm-se:

$$(20)\ \hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{2\sigma^2}\right]\int_{-\infty}^{+\infty} y\exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right]dy}{\sum_{i=1}^{n} \exp\left[-\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{2\sigma^2}\right]\int_{-\infty}^{+\infty} \exp\left[-\frac{(y-y_i)^2}{2\sigma^2}\right]dy}$$

Definindo-se a função $D_i^2 = (\mathbf{x}-\mathbf{x}^i)^T(\mathbf{x}-\mathbf{x}^i)$ e resolvendo-se as integrações da expressão acima, pode-se representar a estimativa de y, pela regressão em questão como:

$$(21)\ \hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}{\sum_{i=1}^{n} \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}$$

Da expressão acima pode-se constatar que a estimativa de y fornecida pela regressão ora apresentada consiste numa média ponderada de todas as observações de y.

Como acontecia com a rede *PNN*, σ é o termo de suavização dos estimadores de densidade e controlam o *tradeoff* entre viés e variância da regressão não-paramétrica.

Com σ grande, $\hat{y}(\mathbf{x})$ é a média amostral dos valores y_i observados; para valores pequenos de σ, $\hat{y}(\mathbf{x})$ assume o valor de y_i associado com a observação mais próxima de $\mathbf{x}$. Para valores intermediários de σ, todos os valores y_i são levados em consideração no cômputo da estimativa $\hat{y}(\mathbf{x})$.

Um valor apropriado de σ pode ser determinado empiricamente, tal como ocorre com a rede *PNN*. SPECHT (1991) sugere o uso validação cruzada *Leave One Out* para determinação de um vetor apropriado de sigmas.

A normalização das variáveis de entrada é necessária, sendo portanto uma etapa de pré-processamento adotada antes do treinamento da rede. A normalização visa garantir que as variáveis envolvidas tenham aproximadamente as mesmas faixas de valores.

Tal como ocorre com a *PNN*, pode-se adotar qualquer janela de Parzen, na estimativa da densidade em questão, não havendo, razão para se restringir qualquer aplicação ao *kernel* gaussiano.

3.2.1) *GRNN* e *Clustering*

Em aplicações com grandes massas de dados, a *clusterização* dos mesmos é necessária. De acordo com SPECHT (1991), nessas situações, a regressão proposta pode ser determinada pela seguinte relação:

$$(22)\ \hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{m} A_i \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}{\sum_{i=1}^{m} B_i \exp\left[-\frac{D_i^2}{2\sigma^2}\right]}$$

Onde:

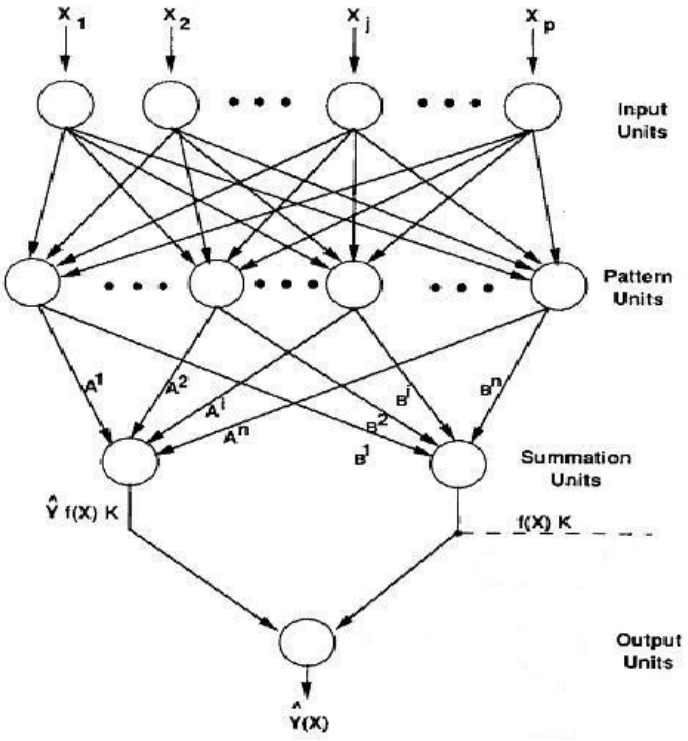
A_i é a soma dos valores observados y_i da variável resposta, em cada *cluster* i;

B_i é o número de padrões (vetores $\mathbf{x}$) em cada *cluster* i.

3.2.2) Implementação via Rede Neural

A figura 4 abaixo apresenta um esquema de como a regressão não-paramétrica em questão pode ser implementada em uma estrutura de rede neural, de acordo com SPECHT (1991).

Figura 4: Implementação de Regressão Geral por Redes Neurais *Feed Foward*:



Fonte: SPECHT (1991).

A camada de unidades de *input*, tal como na rede *PNN* é meramente uma camada de unidades de distribuição, que provê os valores normalizados das variáveis para as unidades da camada seguinte, a camada de padrões.

Nesta segunda camada, tem-se que cada unidade armazena um vetor da amostra (padrão) ou um centro de *cluster*, caso se tenha procedido a clusterização dos dados. Quando um novo vetor **x** entra na rede, ele é subtraído de cada vetor representando os centros de *cluster*, armazenado na memória principal do computador. Essas diferenças são elevadas ao quadrado e somadas. Estes somas são então aplicadas a uma função de ativação exponencial.

Esses resultados são enviados à próxima camada, que contém as unidades de soma. As conexões sinápticas entre as unidades de padrão e as unidades de soma possuem pesos específicos da arquitetura *GRNN*: as conexões ligando as unidades de padrão a uma das unidades de soma terão pesos A_i (isto é, a soma dos valores das variáveis-resposta y relacionadas aos vetores **x** em cada centro de *cluster* i), as outras conexões relacionadas à outra unidade de soma terão pesos B_i (que representam o número de padrões em cada *cluster*).

As unidades de soma efetuam então um produto interno do vetor de pesos com o vetor composto dos sinais enviados pelas unidades de padrão. O nó de saída simplesmente dividirá um resultado pelo outro, obtendo assim a estimativa $\hat{y}(\mathbf{x})$.

4 METODOLOGIA

Como inicialmente apontado na introdução deste trabalho, o problema de classificação a ser abordado com *PNN*'s consiste na classificação do risco de morte por Síndrome Coronariana Aguda (SCA) utilizando-se o mesmo conjunto inicial de variáveis de REIS (2007). Este banco de dados foi obtido a partir de um estudo de coorte prospectiva, composta por pacientes internados com o diagnóstico de Síndrome Coronariana Aguda (SCA) com ou sem supra desnível do ST e acompanhados durante a internação hospitalar até a alta ou óbito.

Esses pacientes foram sendo introduzidos nessa base, durante o período de julho de 2004 a junho de 2005, em quatro hospitais particulares e três hospitais públicos no município de Niterói, RJ.

Em REIS (2007) foram analisadas 25 variáveis de 411 pacientes incluídos no estudo ao longo de 12 meses consecutivos.

É necessário, contudo, destacar algumas particularidades desse conjunto de dados:

-A baixa ocorrência de óbitos (somente 37 dos 411 casos estudados);

-A grande parcela de pacientes com informações ausentes com relação às variáveis consideradas: somente 231 pacientes tinham informações completas referentes às 25 variáveis selecionadas; todos os demais apresentavam dados ausentes com respeito a pelo menos uma dessas variáveis.

Essas particulares conferem ao problema assim definido uma grande complexidade caso se desejasse esgotar cada possível combinação de variáveis (cada combinação implicando em um determinado número de observações completas). Além disso, cada possível combinação de variáveis traria implicitamente um número diferenciado de (já escassos) casos de desfecho com óbito.

Perante este quadro, com o objetivo de reduzir o número de variáveis e aumentar o número de informações disponíveis, foram utilizados no presente trabalho diferentes critérios de seleção de variáveis (cinco critérios estatísticos), como detalhados na próxima seção.

4.1 ESQUEMA DE VALIDAÇÂO CRUZADA ADOTADO

Dada a escassa amostra disponível para a implementação das *PNN*'s, se optou por um esquema de validação cruzada do tipo *Leave One Out*, que corresponde a se retirar um padrão (vetor de entrada) por vez do banco de dados para colocá-lo como (único) elemento no conjunto de validação. Outra justificativa para esta opção foi permitir uma melhor comparação dos resultados aqui obtidos com os de REIS (2007) e COLLAZO (2009).

Assim, o treinamento da rede neural se repetirá N vezes, sendo que, ao longo de cada treinamento, um padrão diferente irá para o conjunto de validação e N-1 padrões são utilizados para treinamento da rede, sendo o treinamento interrompido toda vez que N-1 padrões no conjunto de validação são apresentados a rede, quando então o modelo é validado com o elemento deixado de fora deste conjunto.

4.2. MEDIDA DE ERRO UTILIZADA NO TREINAMENTO, VALIDAÇÃO E TESTE DAS REDES.

Foi adotado como medida de erro, nos processos de treinamento, validação e teste das redes o Erro Médio Quadrático (*Mean Squared Error*, ou *MSE*).

4.3 *SOFTWARES* E *HARDWARES* UTILIZADOS

As *PNN*'s foram treinadas e testadas no *software* DTERG versão 9.2 corporativa, sob cortesia de seu criador, Phillip H. Sherrod. Os dados foram formatados e organizados em planilhas eletrônicas do Excel do Microsoft Office 2007 Student Version. Os critérios de seleção de variáveis foram implementados no *software* estatístico R versão 2.7.0. Os bancos de dados de entrada para as redes e os relatórios dos softwares DTERG e R foram organizados em planilhas Excel, também pertencentes ao Microsoft Office 2007 Student Version. Por fim, os textos, fórmulas e tabelas foram criados no software Word 2007.

Todas as implementações foram feitas em duas arquiteturas PC diferentes: a primeira correspondendo a um PC com CPU Pentium 4 HT 3.06 GHz, 2 GB de RAM, Barramento 533 MHz, com sistema operacional Windows XP Home Edition. A segunda

arquitetura corresponde a um notebook Intel Core Duo 2 de 2.2 GHz, com 2.5GB de RAM, barramento 1Ghz e sistema operacional Windows Vista Home Edition.

4.4 PADRONIZAÇÃO DOS DADOS

Os valores das variáveis preditoras são padronizados no *software* DTREG, subtraindo-se cada um pela sua mediana e dividindo-se este resultado pela amplitude do interquartil correspondente a este valor.

4.5 CONFIGURAÇÃO DAS *PNN's*

As *PNN*'s utilizadas neste trabalho foram ajustadas, treinadas e testadas no *software* DTREG, que permite muitas opções de configuração das mesmas. A seguir são apresentadas todas as escolhas de configuração e ajustes dos modelos utilizados neste *software*.

4.5.1 Arquitetura geral da *PNN* no *software* DTREG.

As Arquiteturas das *PNN*'s implementadas pelo DTREG seguem o modelo de arquitetura apresentada por SPECHT (1990), como exemplificada na Figura 1 do capítulo 3 deste trabalho. A única diferença é que o nó de decisão, na ultima camada da rede, não aplica os resultados obtidos em uma função de ativação do tipo degrau, ele simplesmente compara os *inputs* recebidos e decide por alocar **x** na categoria correspondente ao *input* de maior valor.

A seguir, detalha-se as escolhas efetuadas durante o treinamento, validação e teste, de cada um dos elementos componentes desta arquitetura.

4.5.1.1. Número de unidades na camada de *input*

Há um neurônio nessa camada para cada variável preditiva contínua utilizada. Para cada variável categórica de k classes haverá k-1 variáveis *dummy* 0-1, representadas por k-1 neurônios nesta camada.

4.5.1.2. Número de neurônios na camada de padrões.

Embora a arquitetura da *PNN* propriamente dita tenha um número de unidades na camada de padrões exatamente igual ao número de vetores (padrões) apresentados a

mesma no treinamento, o *software* DTREG permite algumas opções de remoção de neurônios dessa camada, visando à otimização e simplificação da arquitetura da rede, assim como algum ganho de performance preditiva da mesma.

Neste *software* o processo de remoção de neurônios na camada mencionada é iterativo (e geralmente lento), pois cada modelo, com um determinado número fixo de unidades, é avaliado e o modelo com menor erro é escolhido. Em outras palavras, para cada número escolhido de neurônios nesta camada, o *software* remove um neurônio por vez da mesma, treinando e avaliando a rede com essa escolha. A seguir, devolve esse neurônio e retira outro da mesma, reavaliando o modelo com essa escolha. Por fim decide pela arquitetura que produzir o menor erro. Esse processo é repetido variando-se o número total de neurônios na camada considerada.

O processo de remoção de neurônios na camada de padrões pode ser efetuado de diversas formas:

a) Para minimização de erro – neurônios são removidos nesta camada enquanto o erro no conjunto de validação (o software sempre utiliza, para esses propósito, validação cruzada do tipo Leave One Out) permanece constante ou diminui.

b) Para minimizar neurônios - neurônios são removidos até que o erro *leave one out* exceda o erro correspondente ao modelo completo (com todas as unidades de padrão).

c) Número pré-determinado de neurônios nesta camada – Os neurônios são removidos até que o número especificado de unidades nesta camada seja alcançado.

Para cada uma dessas opções há a opção de re-treinamento após a remoção de cada neurônio, o que torna o processo de remoção ainda mais lento (algumas sessões de treinamento, das PNN's com possibilidade de remoção de neurônios e re-treinamento duraram mais de 12 horas com a arquitetura de hardware Core 2 Duo).

Não se verificou nenhum ganho de acurácia no conjunto de teste utilizado, com nenhum dos esquemas mencionados de remoção de neurônios na camada de padrões, se optando, portanto, pela arquitetura completa, típica da *PNN*. Como mencionado anteriormente, se a amostra de dados tem N padrões, esta camada terá N-1 unidades, haja visto o esquema de validação cruzada *Leave One Out* adotado.

4.5.1.3. Escolha da função *kernel*.

A função *kernel* escolhida para a estimação das densidades de probabilidade da população em cada classe nas *PNN*'s consideradas, foi a gaussiana, por ter sido esta função que apresentou os melhores resultados (em termos de acurácia) nos conjuntos de treinamento, validação e teste.

4.5.1.4 Escolha dos parâmetros sigma (desvios-padrão).

A versão utilizada do DTREG permite muitas possibilidades de configuração de como será determinado o vetor de sigmas para a estimativa de densidade multivariada na *PNN*. O software permite se determinar um único valor sigma para todas as variáveis envolvidas, um sigma específico para cada variável preditora, ou ainda um valor sigma específico para cada variável preditora.

Para qualquer uma dessas opções, os valores sigma ótimos são determinados utilizando-se o algoritmo de otimização numérica não-linear conhecido como Gradiente Conjugado (GC) na função de erro médio quadrático (*MSE*). Ao longo deste processo de otimização, o *software* usa o método de validação *Leave One Out* (isto é, deixa um padrão de fora) para avaliar a performance da rede treinada com um conjunto de valores sigma considerados.

Deve-se destacar que existem várias versões do Gradiente Conjugado. O manual do DTREG não especifica que versão é implementada na minimização do *MSE* no conjunto de validação. Maiores referências sobre os métodos Gradiente Conjugado se encontram em BERTSEKAS (2003).

Com cada escolha de vetor sigma, o *software* procede a N treinamentos, cada vez utilizando N-1 padrões para treinamento e validando o modelo com o padrão deixado de fora. Após esses N treinamentos, se faz a média dos erros quadráticos (*MSE*), se calcula uma nova direção de descida e o algoritmo GC computa então um novo vetor sigma. Todo o procedimento é repetido até que um dos critérios de parada deste algoritmo seja alcançado.

Para se tentar reduzir a possibilidade do GC convergir para mínimos locais não-globais, o algoritmo é repetido para diversos vetores iniciais de sigmas. Esses valores iniciais para os sigmas são determinados (por um processo de aleatorização) dentro de uma faixa de valores possíveis pré-especificada pelo usuário. Neste trabalho, para cada rede treinada, foram feitas vinte inicializações de treinamentos com diferentes vetores

iniciais de sigmas, cujas componentes foram determinadas desta forma (foram tentados também treinamentos com mais de vinte inicializações, sem ganhos de performance e com custos crescentes de tempo de computação).

Foi verificado ao longo dos treinamentos de *PNN*'s de diversas configurações, que a faixa de [0,0001; 0,7] para os valores iniciais sigma a serem utilizados pelo CG produzia consistentemente sensibilidades e especificidades mais altas do que outras, para a grande maioria dos conjuntos de variáveis considerados, sendo portanto esta a faixa adotada para a escolha dos componentes do vetor de sigmas a ser utilizado como ponto inicial para o algoritmo do GC em cada inicialização aleatória. Assim, cada implementação do algoritmo CG selecionava aleatoriamente valores sigma (não necessariamente iguais para cada variável) neste intervalo.

4.5.1.4.1. Critérios de parada para o algoritmo do GC

O algoritmo GC implementado pelo DTREG na escolha dos valores sigma para as *PNN*'s durante o treinamento, combina uma série de critérios de parada, interrompendo o processo de otimização assim que um deles for atingido. Os critérios de parada utilizados pelo *software* são:

a) Número máximo de iterações: por este critério de parada, uma vez que o algoritmo GC tenha alcançado este número máximo de iterações permitidas, ele pára. Foi escolhida a opção de cinco mil (5.000) iterações para este critério.

b) Tolerância de convergência absoluta: se o erro medido em uma iteração do GC é menor do que o valor escolhido para este parâmetro, então o software assume que o algoritmo convergiu para a solução ótima e interrompe o treinamento da rede. Foi escolhido para este critério o valor *defaut* do *software* de $1x10^{-8}$.

c) Tolerância de convergência relativa: se o erro cai por menos do que o valor estabelecido para este critério, então o treinamento também é encerrado. Foi mantido para este critério o valor *defaut* de $1x10^{-4}$.

d) Tempo máximo de execução do algoritmo do GC. Este critério de parada interrompe o treinamento da rede assim que o tempo de treinamento atinja o valor estabelecido para este critério. Esta opção foi desativada nos treinamentos das *PNN*'s consideradas neste trabalho.

A escolha dos parâmetros relativos aos critérios de parada do algoritmo GC foi feita empiricamente visando-se buscar um *tradeoff* adequado entre o tempo de execução

do treinamento e performance preditiva (em termos de sensibilidade e especificidade dos modelos finais, nos conjuntos de teste).

4.5.1.5 Escolha da função de perda utilizada no classificador

A função de perda utilizada para os classificadores implementados pelas redes neurais consideradas foi a que atribui custos unitários para todo tipo de classificação equivocada e custo zero para as classificações corretas.

4.5.1.6. Escolha das probabilidades *prioris* para cada categoria

As probabilidades *prioris* escolhidas foram de 0,5 para a categoria 0 (a categoria dos pacientes que sobreviveram durante o período de estadia hospitalar) e 0,5 para a categoria 1 (dos pacientes que faleceram ao longo desse período). Essas probabilidades foram escolhidas de acordo com a performance do classificador e não com o propósito de refletirem um conhecimento prévio sobre o problema analisado. Assim diversas combinações de probabilidades prioris foram testadas, e a escolha mencionada foi a que produziu as melhores sensibilidades e acurácias.

4.6. BANCO DE DADOS UTILIZADO PARA TESTE DOS MODELOS FINAIS.

Embora o ideal seja utilizar, como conjunto de teste dos modelos finais, um banco com padrões ainda não apresentados à rede durante o treinamento e validação, pela escassez de dados, o mesmo banco de dados utilizado nas duas fases anteriores foi utilizado como conjunto de teste.

4.7 VARIÁVEIS E CRITÉRIOS DE SELEÇÂO

4.7.1 Variáveis Consideradas

As 25 variáveis constantes do estudo de REIS (2007) representavam informações referentes a diferentes aspectos dos pacientes. Assim o trabalho envolveu variáveis relacionadas a aspectos físicos e sociais (Idade, Índice de Massa Corporal, Sexo, Escolaridade), variáveis clínicas anteriores à internação atual (Infarto Agudo do Miocárdio Prévio, Qualquer Revascularização Prévia, Tabagismo, História Familiar de Doença Arterial Coronariana), variáveis clínicas e laboratoriais obtidas na admissão hospitalar dos pacientes (Tipo de SCA, Tempo para o Primeiro Atendimento Médico,

Freqüência Cardíaca, Classe Killip e Creatinina). Este trabalho também utilizou uma variável relacionada ao nível de atividade física dos pacientes (Grupo Atividade Física).

Variáveis relativas à presença de fatores de risco também foram utilizadas, como as variáveis de diagnóstico de: Hipertensão Arterial Sistêmica, Colesterol Total Elevado, Triglicerídeos Elevados e Colesterol-HDL Baixo.

Por fim, foram acrescentadas as variáveis genéticas utilizadas por REIS (2007). Estas variáveis foram: alelo D e I do polimorfismo do gene da enzima conversora da angiotensina I (ECA); alelo M e T do polimorfismo M235T do gene do angiotensinogênio (AGT) e os alelos E2, E3 e E4 do polimorfismo do gene da apolipoproteína E (APO E).

Além destas variáveis, utilizou-se também a variável desfecho (representando o desdobramento final de cada caso de internação, com o óbito do paciente ou o não-óbito).

A variável de desfecho possui duas classes e identifica se ocorreu o óbito ou não do paciente, representando o desdobramento final de cada caso de internação .

DEFINIÇÃO DAS VARIÁVEIS

Variáveis Antropométricas e Sociais

- Idade

A idade foi medida em anos e corresponde a uma variável inteira e não categórica.

-Índice de Massa Corporal (IMC)

O IMC (Kg/m^2) expressa à relação entre o peso e a altura de um indivíduo, sendo representado por uma variável contínua. Ele é o método mais prático e rápido para avaliar o grau de risco associado à obesidade.

-Sexo

O sexo é uma variável categórica de dois níveis: masculino e feminino.

-Escolaridade

A escolaridade é representada por uma variável categórica com cinco classes: analfabeto, primeiro grau (completo ou não), segundo grau (completo ou não), nível universitário (completo ou não), e pós-graduado.

É importante salientarmos que ao constatarmos na sociedade brasileira um forte vínculo entre escolaridade e nível socio-econômico, esta variável traz consigo tacitamente algumas clivagens imanentes à qualidade de vida das diversas classes sociais, como, por exemplo, tipo de alimentação, acesso a remédios e a atendimentos médicos, nível de estresse e outras.

-Grupo Atividade Física (AF)

Os indivíduos foram classificados quanto à variável Atividade Física (AF) em duas categorias ou grupos: os que praticam atividades físicas durante 30 minutos consecutivos três ou mais vezes durante a semana , os quais foram considerados como ativos ou os sedentários, em caso contrário.

Variáveis Clínicas anteriores à Internação

-Infarto do Miocárdio Prévio

O infarto do miocárdio prévio é representado por uma variável bi-classe, sendo considerado para sua caracterização o relato do paciente de infarto prévio ou evidências do mesmo presente em exames complementares, tais como eletrocardiograma, ecocardiograma, cintilografia miocárdia ou angiografia coronariana.

-Qualquer Revascularização Prévia (QRP)

A variável QRP é bi-classe e engloba a realização prévia à internação de angioplastia coronariana e/ou de cirurgia de revascularização miocárdia (CRM). Para sua caracterização, foram considerados laudos médicos ou relatos detalhados dos pacientes, que deveriam possuir evidência de cicatriz cirúrgica compatível, no caso de CRM.

-Tabagismo

Os indivíduos foram categorizados em três classes:

* não tabagistas: os fumantes passivos e as pessoas que nunca fumaram ou fumaram menos de cinco cigarros por menos de cinco anos;

* ex-tabagistas: pessoas que pararam de fazer uso do tabaco há mais de seis meses; e

* tabagistas atuais e eventuais: pessoas que fizeram uso regular do tabaco durante os seis meses que antecederam a coleta dos seus dados.

-História Familiar de Doença Arterial Coronariana (DAC)

A história familiar de DAC é uma variável bi-classe, que foi coletada apenas em primeiro grau, isto é, entre pais e filhos. Foi considerada positiva a história familiar de DAC se pais ou filhos dos pacientes tivessem tido doença coronariana (IAM ou angina pectoris) ou tivessem sido submetidos à angioplastia coronariana ou CRM com idade < 55 anos para homens e < 65 anos para mulheres.

Variáveis Clínicas e Laboratoriais na Admissão Hospitalar

-Tipo de Síndrome Coronariana Aguda (SCA)

O tipo de SCA é uma variável categórica com três classes: angina instável, IAM sem supra desnível do ST e IAM com supradesnível do ST.

-Tempo para o primeiro Atendimento Médico (Delta t1)

Esta variável Delta t1 mede o intervalo de tempo, em horas, entre o início da dor até a ocorrência do primeiro atendimento médico, em qualquer posto de saúde ou hospital. É representado por uma variável contínua.

-Freqüência Cardíaca (FC)

A freqüência cardíaca é uma variável inteira, medida “no primeiro atendimento médico após o início do quadro coronariano agudo.” (REIS, 2007)

-Classe Killip

Os indivíduos foram categorizados em quatro classes segundo REIS (2007):

* KILLIP I: ausência de sinais de insuficiência cardíaca;

* KILLIP II: presença de estertores em bases pulmonares ou de terceira bulha na ausculta cardíaca;

* KILLIP III: presença de edema agudo de pulmão e;

* KILLIP IV: choque cardiogênico.

-Creatinina

A creatinina sérica é uma variável contínua que foi coleta na sua primeira medida após dar entrada no hospital e que avalia a função renal do paciente.

Variáveis de Diagnóstico de Fatores de Risco

-Hipertensão Arterial Sistêmica (HAS)

A HAS é uma variável bi-classe que caracteriza indivíduos com pressão arterial sistólica superior ou igual a 140 mmHg ou com pressão diastólica superior ou igual a 90 mmHg ou ainda indivíduos que faziam uso de anti-hipertensivos.

-Colesterol Elevado

O Colesterol Total Elevado é representado por uma variável bi-classe, sendo caracterizado como portadores de colesterol elevado os indivíduos com colesterol total acima de 200 mg/dl na primeira medida após admissão hospitalar ou uso prévio de drogas hipolipemiantes (estatina e/ou fibrato).

-Triglecerídeos Elevados

Os Triglecerídeos Elevados constituem uma variável bi-classe, na qual os indivíduos são considerados portadores de triglicerídeos elevados se apresentassem mais de 150 mg/dl de triglecerídeos na primeira medida após admissão hospitalar ou uso prévio de drogas hipolipemiantes (estatina e/ou fibrato).

-Colesterol-HDL Baixo

O Colesterol-HDL Baixo é uma variável bi-classe, na qual são considerados portadores de Colesterol-HDL Baixo os indivíduos com colesterol-HDL inferior a 40

mg/dl na primeira medida após admissão hospitalar ou uso prévio de drogas hipolipemiantes (estatina e/ou fibrato).

Variáveis Genéticas

Para entendermos a estruturação das variáveis genéticas, é preciso que nos familiarizemos com os conceitos abaixo:

- Polimorfismo genético: são mutações presentes em mais de 1% da população, tornando possível a observação de diferentes formas alélicas de um mesmo lócus gênico;

-Gene: constitui a peça central da hereditariedade. Ele é um segmento do DNA, sendo responsável pela produção de uma proteína ou pela determinação de uma característica do indivíduo;

-Alelo: é cada uma das formas que um determinado gene pode possuir. Por exemplo, o gene que estabelece a cor dos olhos é o mesmo em todos os indivíduos, contudo ele apresenta pequenas nuances (ou diferentes alelos) se uma pessoa possui olhos azuis, verdes ou castanhos; e

-Genótipo: é a constituição gênica dos indivíduos. Por exemplo, se um gene possui os alelos R, S e T, uma determinada pessoa pode possuir um dos seguintes genótipos deste gene (e apenas um): RR, SS, TT, RS, RT ou ST.

Neste estudo, o nível de análise foi limitado à presença dos alelos nos indivíduos. Isto se justifica ao constatarmos que a presença de determinados genótipos é bastante reduzida na amostra. Foram consideradas sete variáveis genéticas:

-alelo D do polimorfismo da enzima conversora da angiotensina I (ECA);

-alelo I do polimorfismo da enzima conversora da angiotensina I (ECA);

-alelo M do polimorfismo M235T do gene do angiotensinogênio (AGT);

-alelo T do polimorfismo M235T do gene do angiotensinogênio (AGT);

-alelo E2 do polimorfismo do gene da apolipoproteína E (APO E);

-alelo E3 do polimorfismo do gene da apolipoproteína E (APO E); e

-alelo E4 do polimorfismo do gene da apolipoproteína E (APO E).

4.7.2. CRITÉRIOS DE SELEÇÃO DE VARIÁVEIS

A partir do conjunto inicial de variáveis acima relatado, foram utilizados diversos critérios estatísticos de seleção de variáveis.

Assim, foram empregados neste trabalho os mesmos métodos para seleção de variáveis que os adotados em COLLAZO (2009) e REIS (2007) para fins de facilitar a comparação de resultados com estes autores. Neste sentido, o método de seleção proposto inicialmente por CHEN, T. et al (2008), os três métodos criados por COLAZO (2009) e o MFISU utilizado por REIS (2007) são empregados na seleção de variáveis para treinamento validação e teste das *PNN*'s aqui utilizadas. Esses métodos serão chamados aqui de Critérios Chen, Adaptado, Dual, Adaptado Dual e MFISU.

Maiores referências sobre os mesmos se encontram em BATITI (1994), COLLAZO (2009), CHEN (2008) e KWAK (2002).

Para os critérios estatísticos de seleção de variáveis, se adotou no estudo, conjuntos formados pelas três, quatro, cinco e seis melhores variáveis ranqueadas segundo cada critério. As seleções de variáveis efetuadas por cada método se encontram na tabela abaixo.

Tabela 4: Variáveis selecionadas segundo cada critério

CHEN	**Adaptado**	**Dual**	**Adaptado Dual**	**MIFS-U**
QRP	QRP	Creatinina	Creatinina	Idade
Alelo E3	HAS	FC adm	FC adm	QRP
Delta t1	Delta t1	Idade	Killip	Creatinina
Alelo E2	Alelo I	HDL	Idade	IMC
Infarto Prévio	Infarto Prévio	Tabagismo	Alelo E3	Genótipo DD
HAS	Atividade Física	Escolaridade	Alelo D	Genótipo E4E4

Para todos os critérios de seleção de variáveis considerados, cada conjunto de variáveis foi utilizado com o maior número possível de observações completas. Assim, pacientes com informações ausentes segundo cada combinação de variáveis, não entraram nos conjuntos de treinamento, validação e teste.

5 RESULTADOS

Como adiantado no capítulo anterior, este estudo aplicou redes *PNN*'s ao problema de classificação do risco de morte por Síndrome Coronariana Aguda utilizando para isso o mesmo banco de dados utilizado por REIS (2007) e COLLAZO (2009).

Com este propósito, se treinou e testou *PNN's* com variáveis selecionadas segundo os mesmos critérios de seleção utilizados por estes autores.

As tabelas abaixo, resumem a performance das redes *PNN* ao longo dos diversos conjuntos de variáveis utilizados e dos diversos critérios de seleção adotados.

Tabela 5.1: Performance das *PNN*'s para as variáveis do critério CHEN

Variáveis	observações completas	acurácia (%)	especificidade (%)	sensibilidade (%)
QRP, alelo E3, Delta t1	372	81,18%	81,16%	81,48%
QRP, alelo E3, Delta t1, alelo E2	372	81,99%	82,03%	81,48%
QRP, alelo E3, Delta t1, alelo E2, Infarto Prévio	362	84,81%	84,27%	92,00%
QRP, alelo E3, Delta t1, alelo E2, Infarto Prévio, HAS	362	88,67%	88,43%	92,00%

Na tabela 5.1 observa-se que o alelo E3, na presença das demais variáveis selecionadas segundo o critério CHEN, não gerou classificadores com acurácias que alcançassem ao menos 90% na predição do desfecho. Todas as combinações de variáveis segundo este critério não levaram a um desempenho tão bom quanto os observados nas demais tabelas abaixo. O acréscimo da variável Infarto Prévio melhorou bastante a sensibilidade do modelo (um aumento de 10,52%) e aumentou a

especificidade em 2,24%. A posterior inclusão da variável HAS melhorou um pouco a especificidade da rede em 4,16%.

A inclusão da variável alelo E2 gerou uma ligeira melhoria na especificidade das redes.

Tabela 5.2: Performance das *PNN*'s para as variáveis do critério Adaptado

Variáveis	observações completas	acurácia (%)	especificidade (%)	sensibilidade (%)
QRP, HAS, Delta t1	386	73,83%	72,27%	93,10%
QRP, HAS, Delta t1, alelo I	384	79,69%	78,15%	100,00%
QRP, HAS, Delta t1, alelo I, Infarto Prévio	373	90,35%	90,23%	92,00%
QRP, HAS, Delta t1, alelo I, Infarto Prévio, Grupo Atividade Física	373	91,15%	91,09%	92,00%

Como se pode notar na Tabela 5.2, e tal como também ocorrido no estudo de COLLAZZO (2009) a incorporação do alelo I ao conjunto de variáveis QRP, HAS e Delta t1 propiciou uma melhora mais evidente na sensibilidade do classificador empregando estas variáveis. Provavelmente a principal razão da melhora neste item deve-se ao fato de que, a variável alelo I, ao ser incluída no conjunto de variáveis, elimina dois pacientes que evoluíram para óbito. Estes pacientes correspondiam a casos que a rede *PNN* tinha dificuldade de classificar corretamente –possivelmente, estes pacientes são *outliers* - de forma que a eliminação dos mesmos melhorou substancialmente a sensibilidade do modelo. Na mesma tabela, podemos observar também que o acréscimo da variável Infarto Prévio melhorou enormemente a especificidade do modelo em 12,08%, causando, ao mesmo tempo uma piora de 8% na

sua sensibilidade. A adição da variável Grupo Atividade Física melhorou ligeiramente (em menos de 1%) a especificidade.

Tabela 5.3: Performance das *PNN*'s para as variáveis do critério Dual

Variáveis	Observações completas	acurácia (%)	especificidade (%)	sensibilidade (%)
Creatinina, FC adm, Idade	364	100,00%	100,00%	100,00%
Creatinina, FC adm, Idade, Diag HDL Baixo	323	99,69%	99,66%	100,00%
Creatinina, FC adm, Idade, Diag HDL Baixo, Grupo Tabaco	323	100,00%	100,00%	100,00%
Creatinina, FC adm, Idade, Diag HDL Baixo, Grupo Tabaco, Escolaridade	317	100,00%	100,00%	100,00%

Na tabela 5.3 observa-se que com apenas três variáveis, Creatinina, FC adm e Idade já se obtêm 100% de acurácia, especificidade e sensibilidade. A adição da variável Diag HDL Baixo levou a discreta diminuição na acurácia e especificidade do classificador. A adição das variáveis Grupo Tabaco e Escolaridade não trouxe benefícios, pois já se tinha sido alcançado 100% de acurácia com as três primeiras variáveis consideradas

Tabela 5.4: Performance das *PNN*'s para as variáveis do critério Adaptado Dual

Variáveis	observações completas	acurácia (%)	especificidade (%)	sensibilidade (%)
Creatinina,FC, adm, Killip	347	89,91%	88,96%	100,00%
Creatinina, FC adm, Killip, Idade	347	100,00%	100,00%	100,00%
Creatinina, FC adm,	336	99,40%	99,35%	100,00%

Killip, Idade, alelo E3				
Creatinina, FC adm, Killip, Idade,alelo E3, alelo D	336	97,92%	97,72%	100,00%

Com relação ao critério Adaptado Dual (Tabela 5.4), o conjunto ótimo de variáveis é o formado por Creatinina, FC adm, Killip e Idade, na medida que as inclusões das variáveis genéticas E3 e D propiciaram pioras sucessivas na especificidade e acurácia dos modelos.

Tabela 5.5: Performance das *PNN*'s para as variáveis do critério MIFS-U

Variáveis	observações completas	Acurácia (%)	especificidade (%)	sensibilidade (%)
Idade, QRP, Creatinina	393	83,72%	82,27%	100,00%
Idade, QRP, Creatinina, IMC	378	99,74%	99,72%	100,00%
Idade, QRP, Creatinina, IMC, DD	377	91,51%	90,88%	100,00%
Idade, QRP, Creatinina, IMC, DD, E4E4	365	95,07%	94,71%	100,00%

Na tabela 5.5 as variáveis Idade, QRP, Creatinina e IMC obtiveram excelente desempenho e o acréscimo das variáveis genéticas DD e E4E4 levaram a piora dos resultados. Pode-se constatar que a inclusão da variável IMC foi capaz de elevar a especificidade em 17,45%, o que por sua vez causou um aumento de 16,02% na acurácia.

O papel e a relevância das variáveis genéticas na classificação do risco de morte não parece ter sido totalmente elucidado com os resultados obtidos. A melhora de desempenho com a inclusão do alelo I ao conjunto das três melhores variáveis segundo o critério Adaptado (tabela 5.2) possivelmente se deveu aos fatores anteriormente mencionados e não pelo poder explicativo da variável em si. A inclusão sucessiva dos alelos E3 e D ao conjunto das três melhores variáveis no critério Adaptado Dual (tabela

5.4) pioraram os classificadores construídos a partir delas. A adição da variável E2 ao conjunto das três melhores variáveis segundo o critério de Chen (tabela 5.1) gerou uma melhora muito pequena (menor que 1%) na acurácia e na especificidade).

Na tabela 5.4 que descreve os resultados selecionando as variáveis pelo critério Adaptado Dual, a inclusão da variável Idade propiciou uma melhora significativa no desempenho da rede neural, no que tange à especificidade e acurácia do modelo. Sua inclusão no conjunto de treinamento levou a especificidade de 88,96% para 100% (causando aumento da acurácia do modelo de 89,91% para 100%). Provavelmente, a principal razão desta melhoria se deve ao fato de que todos os pacientes constantes da amostra declararam suas idades, não havendo informações ausentes com respeito a ela no banco de dados utilizado. Assim, a inclusão desta variável não implica em perda de dados no conjunto de treinamento.

Outro aspecto importante que se pode concluir a partir das tabelas acima é que, comparando-se as três melhores variáveis segundo o critério Dual (tabela 5.3) com as três melhores do critério Dual Adaptado (tabela 5.4) , vê-se que elas diferem apenas por uma variável: no primeiro critério, as três melhores variáveis correspondem a Creatinina, FC adm e Idade e no segundo a variável Killip substitui a variável Idade.

No entanto, a performance das redes nesse dois conjuntos é significativamente diferente. No primeiro conjunto de variáveis (de acordo com o critério Dual) o classificador atingiu o máximo em todos os quesitos (acurácia, especificidade e sensibilidade). No entanto, o classificador construído com o segundo conjunto possui performance bem inferior (acurácia de 89,91%, especificidade de 88,96% e sensibilidade de 100%), indicando que a variável Idade é muito mais adequada na explicação do desfecho (sobretudo para os desfechos sem óbito) do que a variável Killip. O mesmo efeito aliás foi verificado também por Collazo na aplicação de *Support Vector Machines* (*SVM*'s) na classificação do risco aqui tratado com o mesmo conjunto de dados.

Assim, esse resultado revela que o poder explicativo da variável Killip, na presença de Creatinina e FC adm não é tão bom quanto o da variável Idade e que a perda de capacidade explicativa do modelo, com a troca da variável Idade por Killip, não é devida a um menor número de observações completas. Isso pode ser verificado na tabela 5.4, onde a posterior inclusão de Idade no modelo com a variável KIllip, embora não aumente essas observações, traz novamente a acurácia do modelo para 100%.

Com o apresentado até então, pode-se concluir que as variáveis que foram mais bem sucedidas na classificação do risco aqui estudado foram Creatinina, FC Adm e Idade. Contudo, diferentemente do observado por Collazo na aplicação de *SVM*'s neste problema de classificação, as variáveis QRP, HAS e Delta t1 não obtiveram resultados tão bons na predição do desfecho, quando o classificador passa a ser construído por redes neurais probabilísticas.

Tabela 5.6: Desempenho Médio das *PNN*'s em cada critério de seleção de variáveis

Critérios	No médio de observações completas	acurácia média (%)	especificidade média (%)	sensibilidade média (%)
Critério CHEN	367,00	84,16%	83,97%	86,74%
Critério Adaptado	373,00	90,75%	90,66%	92,00%
Critério Dual	321,00	99,90%	99,89%	100,00%
Critério Adaptado Dual	341,50	96,81%	96,51%	100,00%
Critério MFIS-U	373,33	95,44%	95,10%	100,00%

Como se pode constatar a partir da Tabela 5.7, tomando-se a acurácia média como medida resumo de desempenho, o critério de seleção de variáveis que gerou classificadores do tipo PNN com melhor desempenho médio foi o Critério Dual, com acurácia média de 99,90%, especificidade média de 99,89% e sensibilidade média de 100%.

Em ordem decrescente de desempenho, seguindo-se o critério Dual, tem-se o critério Adaptado Dual (com acurácia média de 96,81%), o critério MFIS-U (95,44%), o critério Adaptado (90,75%) e o critério de Chen (84,16%).

Analisando-se estes resultados, já se esperava um bom desempenho dos critérios Adaptados sobre o de CHEN et al. (2008), por ordenarem as variáveis diretamente pela quantidade principal. Assim, os critérios Adaptados geraram classificadores com desempenho médio superior ao inicialmente proposto por Chen et al, como também observado por COLAZZO (2009).

O critério MFIS-U, com sua estratégia gulosa de ir selecionando variáveis de acordo com o grau de contribuição ao índice de informação mútua conjunta, também propiciou classificadores extremamente atraentes, obtendo a terceira maior acurácia média dentre os critérios de seleção de variáveis utilizados (95,44%). Em termos de sensibilidade média, o mesmo se encontra empatado em primeiro lugar, com os critérios Dual e Adaptado Dual (todos com sensibilidade média máxima).

Dos dados apresentados na tabela acima, pode-se concluir que o desempenho médio das *PNN*'s foi semelhante quando utilizando variáveis selecionadas segundo os critérios Dual, Dual Adaptado e MFIS-U, com a performance segundo o critério Dual sendo um pouco melhor do que os outros dois critérios.

Analisando-se os resultados obtidos com as *PNN*'s construídas com as três, quatro, cinco e seis melhores variáveis segundo cada critério de seleção adotado constatou-se de imediato a performance extremamente positiva das mesmas no problema de classificação em questão: dos conjuntos de variáveis utilizados, sete geraram classificadores que obtiveram 100% de acurácia (o que implica em 100% de sensibilidade e especificidade simultaneamente), o que é de fato um resultado excelente.

Estes resultados superaram os obtidos pelos citados autores: COLLAZO (2009) obteve como melhor performance 97,7% de acurácia, 98,6% de especificidade e 87,5% de sensibilidade combinando variáveis dos critérios Adaptado e Adaptado Dual (mais especificamente as variáveis Creatinina, QRP, Idade e HAS). REIS (2007) obteve como melhor performance com RNA's *feedfoward* 70,7%, 69,8% e 78,3% de acurácia, especificidade e sensibilidade respectivamente.

Uma outra característica marcante dos resultados encontrados com as *PNN's* implementadas com os subconjntos de variáveis empregados, é que as sensibilidades observadas nas redes testadas (isto é, a parcela das vezes que os modelos foram capazes de prever corretamente os casos de desfecho com óbito) foram muito mais elevadas do que as verificadas nos resultados de COLLAZO (2009) e REIS (2007). Este fato merece destaque pois, nos estudos destes autores, para a maior parte de combinações de variáveis, a sensibilidade foi significativamente menor que a especificidade, em função da baixa participação de casos com desfecho de óbito na amostra utilizada.

A terceira menor sensibilidade encontrada neste estudo foi de 92,00% (para *PNN*'s treinados com as cinco e seis variáveis melhor ranqueadas no critério de Seleção de Chen et all –vide Tabela 5.1- e para as *PNN*'s treinados com as cinco e seis melhores

variáveis no critério Adaptado –Tabela 5.2-). Este valor é superior à sensibilidade máxima encontrada por Collazo (que atingiu 90,6% com *Suppot Vector Machines* treinados com variáveis provenientes da combinação dos critérios Adaptado e Adaptado Dual, mais especificamente: QRP, Idade e Creatinina).

À exceção das *PNN*'s treinadas com as variáveis selecionadas pelo critério proposto por CHEN (2008), em todos os outros critérios se verificaram pelo menos um caso com sensibilidade de 100% o que ressalta ainda mais as capacidades classificatórias de tais modelos. Mais especificamente, todas as *PNN*'s treinadas com as variáveis selecionadas segundo os critérios Dual, Adaptado Dual e MIFS-U geraram sensibilidades de 100%.

Acredita-se que as altas sensibilidades encontradas provêm da característica intrínseca das redes *PNN* de lidar bem com *outliers*, que provavelmente existem no banco de dados utilizado. Corroborando com esta posição constatou-se que os valores-sigma associados aos modelos finais (já treinados) eram baixos, o que torna as *PNN*'s mais capazes de lidar com *outliers*, na medida em que pontos extremos recebem maior importância (maior probabilidade segundo a densidade estimada pelo método da Janela de Parzen) ao longo do processo de classificação do que em *PNN*'s com valores sigmas maiores.

6 CONCLUSÕES

O primeiro e principal objetivo deste estudo foi criar modelos de Redes Neurais Probabilísticas. Esses modelos se mostraram capazes de classificar satisfatoriamente o risco de morte de pacientes com SCA, com vistas a futuramente tentar fornecer mais um subsidio, tipicamente de Pesquisa Operacional, ao trabalho clínico na predição do desfecho de pacientes internados com esta doença.

Com relação a este primeiro objetivo, considera-se o mesmo plenamente atingido, pois, como visto, os classificadores aqui construídos obtiveram desempenho excepcional nesta tarefa, atingindo, em vários casos, acurácia máxima, o que os credencia para tais finalidades.

O segundo objetivo deste trabalho foi verificar que subconjunto de variáveis melhor prediz o desfecho. Considera-se, à luz dos resultados obtidos que o mesmo foi alcançado, uma vez que as variáveis Creatinina, FC Adm e Idade foram as que propiciaram a melhor classificação do risco aqui analisado .

O terceiro objetivo do estudo foi verificar o papel das variáveis genéticas na classificação desse risco. Neste sentido verificou-se que apesar de alguns alelos ou genótipos terem sido selecionados inicialmente entre as variáveis com maior relação com o desfecho pelos critérios utilizados, os classificadores mais bem sucedidos obtiveram máxima acurácia apenas com variáveis clínicas. Assim, o presente trabalho não sugere que estes polimorfismos analisados acrescentem informação significativa às variáveis clínicas selecionadas, na predição do desfecho estudado.

O quarto objetivo deste estudo foi comparar o desempenho das redes neurais do tipo *PNN* na classificação do risco aqui analisado, com os desempenhos obtidos por modelos do tipo *SVM* - implementados por COLLAZO (2009) - e redes *Multi Layer Perceptron* treinadas pelo algoritmo de *Backpropagation* – implementadas por REIS (2007) - aplicados ao mesmo problema.

Nesse sentido este estudo treinou e testou redes *PNN* criadas a partir de conjuntos de variáveis selecionadas segundo os mesmos métodos de seleção adotados por estes

dois autores, constatando o melhor desempenho dessas redes na classificação do risco em questão.

Assim, conclui-se que este estudo atingiu os objetivos estabelecidos a priori e que este tipo de modelo tem potencial para vir a ser mais amplamente utilizada na prática clínica, contribuindo na tomada de decisões médicas e na conseqüente redução do número de óbitos devidos à Síndrome Coronariana Aguda no Brasil.

Levando-se em consideração os resultados extremamente satisfatórios aqui obtidos com *PNN*'s na classificação do risco de morte de pacientes com SCA, este trabalho tem como perspectiva futura, a criação de classificadores baseados em tais modelos a partir de bancos de dados de maior porte. Também visa a maior divulgação de tais técnicas como método de apoio à decisão na prática clínica.

REFERÊNCIAS BIBLIOGRÁFICAS

AHMED, N. K., ATIYA, A. F.,2002, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting", *IBM Center for Advanced Studies in Cairo*, Cairo, Disponível em: <: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.114.8923>. Acesso em: 07 set. 2009.

ALENCAR, G. A., CALÔBA, L. P., 2003, "Artificial Neural Networks as Rain Attenuation Predictors in Earth-Space Paths", *IEEE Ont. Symp.on Circuits and Systems*, Bungkok.

ANÔNIMO, 2009, *Probabilistic Neural Networks*. Disponível em: <: http://homepages.gold.ac.uk/nikolaev/311pnn.htm >. Acesso em: 07 set. 2009.

AL-TIMEMY, A. H., AL-NAIMA, F., M., QAEEB, N. H., 2009, "Probabilistic Neural Network for Breast Biopsy Classification", *MASAUM Journal of Computing*, v.1, issue 2, (Set), pp.199-205.

BATTITI, R., 1994, "Using Mutual Information for Selecting Features in Supervised Neural Net, Learning", *IEEE Transactions on Neural Networks*, v.5, n.4, (Jul), pp. 537-550.

BERRAR, D.P., DOWNES, C.S., DUBITZKY, W., 2003, "Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks", *Pacific Symposium on Biocomputing*, v. 8, pp.5-16.

BERTSEKAS, DIMITRI P., 2003, *Nonlinear Programming*. 2 ed. Belmont, Athena Scientific.

BISHOP, C. M., 1995, *Neural Networks for Pattern Recognition*. 1 ed. Oxford, Clarendon Press.

CALÔBA, L. P., REBELLO, J., M., A., SANGRILLO, L. V., S., SILVA, R. R., 2002, "Evaluation of the Relevant Characteristic Parameters of Welding Defects and Probability of Correct Classification sing Linear Classifiers", *Insight*, v. 44, n.10, (Oct), pp. 616-622.

CARVALHO JR., J. G., 2002, *Modelos Conexionistas – Redes Neurais*, Apostila da PUC-Rio, Rio de Janeiro, PUC.

CHEN, T., ZHANG, C., CHEN, X. et al, 2008, "An Input Variable Selection Method for the Artificial Neural Network of Shear Stiffness of Worsted Fabrics Statistical", *Analysis and Data Mining*, v. 1, n. 5, pp. 287-295.

CHTIOUI, Y., PANIGRAHI, S., MARSH, R., "Conjugate Gradient and Approximate Newton Methods for an Optimal Probabilistic Neural Network for Food Color Classification", 1998, *Optical Engineering -Bellingham- International Society for Optical Engineering*, v.37, n. 11 (Nov), pp. 3015-3023.

COLLAZO, R. A. , 2009, Aplicação de "Support Vector Machines" à Classificação do Risco de Morte de Pacientes com Síndrome Coronariana Aguda.Dissertação de de M. SC., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

DEMUTH, H., BEALE, M, 2001, *Neural Network Toolbox, Computation, Visualization, Programming, Matlab User's Guide*, Seventh printing – minor revisions, Massachusetts, The MathWorks, Inc.

DTREG, *User's Manual*. Disponível em: < http://dtreg.com >. Acesso em: 04 setembro. 2009.

EVERITT, B. S., 1987, Introduction to Optimization Methods and Their Application in Statistics, 1 ed., London, Chapman and Hall.

GORUNESCU, F. GORUNESCU, M. EL-DARZI, E. ENE, M. GORUNESCU, S., *2006, "Statistical Comparison of a Probabilistic Neural Network Approach in Hepatic Cancer Diagnosis",* Computer as a Tool, 2005. EUROCON 2005.The International Conference*, Dept. of Math., Biostat. & Comput. Sci., Univ. of Medicine & Pharmacy of Craiova, Belgrado, v.1, (Mai), pp. 237-240.*

GILL, P. E., MURRAY, W., WRIGHT, M. H., 1981, *Practical Optimization.* 1 ed. London, Academic Press.

GONÇALVES, P.A., F., F., A., C., SEABRA-GOMES, R. T., PURSUIT and GRACE "Scores: Sustained Prognostic Value and Interaction With Revascularization in NSTE-ACS". *Eur Heart J* , 2005, v.26, pp 865-872.

HASTIE T., TIBSHIRANI, R., 1990, *Generalized Additive Models*. 1 ed. New York, Chapman and Hall CRC.

HAYKIN, S., 2005, *Neural Networks a Comprehensive Foundation*. 2.ed. Patparganj, Delhi, India, Prentice Hall.

HINES, J. W., 1997, Matlab Suplement to Fuzzy and Neural Approaches in Engineering. 1 ed. Canada, John Wiley and Sons.

HUANG, C, LIAO W. C., 2003, "A Comparative Study of Feature Selection Methods for Probabilistic Neural Networks in Cancer Classification", *Tools with Artificial Intelligence, IEEE International Conference on*, 15th pp. 451.

IZENMANM, A. J., 2008, Modern Multivariate Statistical Techniques, Regression, Classification, and Manifold Learning. 1 ed. New York, Springer.

KWAK, N., CHOI, C.H., 2002, "Input Feature Selection for Classification Problems", *IEEE Transactions on Neural Networks*, v. 13, n. 1, pp. 143-159.

LEE, H., K., H., 2000, Model Selection for Neural Network Classification, Duke University

LEV, E.I., KORNOWSKI, R., VAKNIN-ASSA, H., PORTER, A., TEPLITSKY, I., BEN-DOR, I., BROSH, D., FUCHS, S., BATTLER, A., ASSALI, A. "Comparison of the Predictive Value of Four Different Risk Scores for Outcomes of Patients With ST-Elevation Acute Myocardial Infarction Undergoing Primary Percutaneous Coronary Intervention". *Am J Cardiol*, 2008, vol 102 (1), pp 06-11.

LUENBERGER, D. G., 1984, *Linear and Nonlinear Programming*. 2ª ed. Massachusetts, Addison-Wesley.

LO, A. W., CAMPBELL, J. Y., 1996, *The Econometrics of Financial Markets*. 1 ed. New Jersey, Princeton University Press.

MACKAY, D. J. C., 1991, "Bayesian Interpolation", *Computation and Neural Systems*, California, California Institute of Technology.

_____, 1991, “Bayesian Model Comparison and Backprop Nets”. *Compuatation and Neural Systems*, California Institute of Technology, California.

_____, 1992, “Bayesian Methods for Adaptative Models”, Phd Thesis, California Institute of Technology, California.

NASCIMENTO, E. M., PEREIRA, B., B., SEIXAS, J., M., 2009, “Redes Neurais Artificiais, Uma Aplicação no Estudo da Poluição Atmosférica e Seus Efeitos Adversos à Saúde”, *Revista Brasileira de Biometria*, São Paulo, v.27, n.1 (Jan-Mar), p.37-50**.**

PARZEN, E., 1962, “On Estimation of a Probability Function and Mode”. *The Annals of Mathematical Statistics*, pp.1065-1076, Ohaio.

PATTERSON, D. W., 1996, *Artificial Neural Networks, Theory and Applications*, 1 ed., New York, Prentice Hall

PEREIRA, B. B., RAO, C. R., 2009, *Data Mining Using Neural Networks: A Guide for Statisticians*, Disponível em: <http://textbookrevolution.org/index.php/Data_Mining_using_Neural_Networks: A_Guide_for_Staticians >. Acesso em: 14 de janeiro. 2010..

REIS, A. F., 2007, Modelo preditivo de mortalidade na Síndrome Coronariana Aguda utilizando Redes Neurais Artificiais com base em variáveis clínicas e genéticas.Tese de D. SC., Clínica Médica/Pesquisa Clínica/UFRJ, Rio de Janeiro, RJ, Brasil.

RIPLEY, B. D., 2004, *Pattern Recognition and Neural Networks*. 7ed. New York, Cambridge University Press.

SANTOS, A., M., PEREIRA, B. B., SEIXAS, J. M., MELLO, F. C. Q., 2007, “Neural Networks: an Application for Predicting Smear Negative Pulmonary Tuberculosis”, In: Balakrishnan, N; Auget, J.L.; Mesbah, M.; Molenberghs, G.. (Org.). *Advances in Statistical Methods for the Health Sciences*. Boston, Birkhäuser Boston, v. , p. 275-287.

SANTOS, A., M., PEREIRA, B. B., SEIXAS, J. M., MEDRONHO, R. A., 2005, “Usando Redes Neurais Artificiais e Regressão Logística na Predição da Hepatite A”, *Revista Brasileira de Epidemiologia*, v.8 (2), pp.117-26.

SANDDHYA, S., 2006, *Neural Networks for Applied Sciences and Engineering*. 1ed. New York , Taylor and Francis Group, LLC.

SECRETAN, J., GEORGIOPOULOS, M., CASTRO, J., 2007, A "Privacy Preserving Probabilistic Neural Network for Horizontally Partitioned Databases", *IJCNN*, pp. 1554-1559.

SHAN, Y., ZHAO, R., XU, G., LIEBICH, H., M., ZHANG, Y., 2002, "Application of probabilistic neural network in the clinical diagnosis of cancers based on clinical chemistry data", *Analytica Chimica Acta*, v. 471, Issue 1, n.23 (Out), pp.77-86.

SPECHT, D. F., 1967, "Generation of Polynomia Discriminant Functions for Pattern Recognition" , *IEEE Transactions on Eletronic Computers*, v.16, pp. 308-319.

_____, 1988, "Probabilistic neural networks for classification mapping, or associative memory", *Proceedings, IEEE International Conference on Neural Networks*, v.1, pp. 525-532.

_____, 1990, "Probabilistic Neural Networks", *Neural Networks*, v.3, n.1 (Jun), pp. 109-118.

_____, 1991, "A General Regression Neural Network", *IEEE Transactions on Neural Networks*, v.2, n. 6 (Nov) pp. 568-576.

SUYKENS, M. D., VANDEWALLE, B. D. M., 2002, "Improved Long term Temperature Prediction by Chaining of Neural Networks", *International journal of Neural Systems*, Special Issue on Issue's Topic.

VOSS, R., CULLEN, P., SCHULTE, H., ASSMAN, G., 2002, "Prediction of Risk of Coronary Events in Middle-Aged Men in The Prospective Cardiovascular Munster Study (PROCAM) Using Neural Networks", *International Journal of Epidemiology*, v.31, pp. 1253-1262.

XAVIER, A. E., 2005, "Uma Função Ativação Para Redes Neurais Artificiais Mais Flexível e Poderosa e Mais Rápida", *Revista da Sociedade Brasileira de Redes Neurais*, v.1, n. 5, pp. 276-282.

WASSERMAN, L., 2004, All of Statistics: a Concise Course in Statistical Inference. 2 ed. New York, Springer.

WASSERMAN, P., 1993, *Advanced Methods in Neural Computing*. 1 ed, New York, Van Nostrand Reinhold.