APLICAÇÃO DE "SUPPORT VECTOR MACHINES" À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Rodrigo Abrunhosa Collazo

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:
Prof. Basílio de Bragança Pereira
Drª. Amália Faria dos Reis

Rio de Janeiro
Agosto de 2009

APLICAÇÃO DE "SUPPORT VECTOR MACHINES" À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Rodrigo Abrunhosa Collazo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA EM 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.

Aprovada por:

_____

Prof. Basílio de Bragança Pereira, Ph. D.

_____

Drª. Amália Faria dos Reis, D.Sc.

_____

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

_____

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

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2009

Collazo, Rodrigo Abrunhosa Collazo

Aplicação de “Support Vector Machines” à classificação do risco de morte de pacientes com síndrome coronariana aguda/ Rodrigo Abrunhosa Collazo. - Rio de Janeiro: UFRJ/COPPE, 2009

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

Orientadores: Basílio de Bragança Pereira e Amália Faria dos Reis

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2009.

Refrências Bibliográficas: p. 50-52.

1. Support Vector Machines. 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 em pensamento, José Angelo,
à minha mãe, Etelvina,
e ao meu irmão, Fabio.

AGRADECIMENTOS

Aos meus orientadores, professor Basílio e Drª. Amália, pelo apoio e contribuição ao trabalho desenvolvido.

À professora Laura pela colaboração e orientação no desenvolvimento desta dissertação.

Ao Dr. Nelson Albuquerque pela sua participação na banca.

Ao professor Adilson Elias Xavier, pelas conversas que me propriciaram *insights* para a estabilização do funcionamento do SVM com o emprego da função núcleo do tipo tangente hiperbólica.

A todos os meus professores ao longo da minha trajetória acadêmica, em especial os do COPPE/UFRJ e da Escola Naval, pelos conhecimentos transmitidos.

Ao professor Antonio Luiz Porto e Albuquerque pela confiança, amizade e incentivo ao estudo contínuo, crítico e livre.

À Andréia, secretária da área de PO, pela boa-vontade e pronto-atendimento em sempre me atender e resolver os assuntos burocrático-administrativos inerentes ao curso.

Ao funcionário Pedrinho da secretária do Programa de Engenharia de Produção pelo apoio na execução do processo administrativo correspondente ao depósito desta dissertação.

Aos meus companheiros da Marinha do Brasil, especialmente da Turma Almirante Lúcio Meira e o amigo Leonardo Pessoa, e colegas do curso pelo apoio no estudo e confiança em mim depositada.

Ao Comando da Marinha do Brasil, mormente ao Centro de Análises de Sistemas Navais (CASNAV), pela oportunidade dada em realizar o presente curso, facultando a mim dedicação integral aos estudos. Em especial, agradeço aos Contra-Almirante Bernardo José Pierantoni Gambôa, meu ex-Comandante, e Contra-Almirante Liseo Zampronio, meu atual Comandante., pelo apoio e confiança incondicionais. Ao Capitão-de-Mar-e-Guerra João Augusto Gomes de Queiroz, meu atual Vice-Diretor, e ao Capitão-de-Corveta, meu orientador técnico, agradeço as orientações seguras e incentivos constantes. A toda tripulação do CASNAV, em especial à Divisão de Capacitação representada pela funcionária civil Suzanna, Capitão-de-Mar-e-Guerra

Azevedo e Primeiro-Sargento Cruz, pelo tempo despendido a fim de me atender e auxiliar.

Ao meu pai Angelo, em memória, e a minha mãe Etelvina pela abnegação incondicional e pelos incentivos diuturnos para minha formação e, principalmente, para minha realização pessoal e profissional.

Ao meu irmão Fabio, fiel e constante amigo, pelos créditos de lealdade e capacidade sempre depósitos na minha pessoa.

Finalmente agradeço a Deus, fonte da vida e da sabedoria, por iluminar e proteger meu caminho.

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 "SUPPORT VECTOR MACHINES" À CLASSIFICAÇÃO DO RISCO DE MORTE DE PACIENTES COM SÍNDROME CORONARIANA AGUDA

Rodrigo Abrunhosa Collazo

Agosto/2009

Orientadores: Prof. Basílio de Bragança Pereira
Drª. Amália Faria dos Reis

Programa: Engenharia de Produção

No Brasil, a principal causa de óbitos é em decorrência das doenças associadas ao aparelho circulatório, entre elas a síndrome coronariana aguda. Assim, desenvolver ferramentas cognitivas que contribuam para a redução das mortes devido a esta doença torna-se relevante para os indivíduos e para o Sistema de Saúde Pública do país.

Este estudo desenvolveu uma máquina de aprendizado capaz de classificar o risco de morte de pacientes internados com síndrome coronariana aguda em alto e baixo, a partir da construção de critérios de seleção das variáveis de entrada mais importantes para se inferir o desfecho e da elaboração de um modelo matemático baseado no classificador *Support Vector Machines*.

Os resultados computacionais indicam a prevalência das variáveis de entrada idade, creatinina, qualquer revascularização prévia e hipertensão arterial sistêmica como as mais relevantes para a predição do desfecho, e permitem a obtenção de classificadores com acurácia maior que 90%.

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 "SUPPORT VECTOR MACHINES" TO CLASSIFICATION OF THE RISK OF DEATH OF HOSPITALIZED PACIENT WITH ACUTE CORONARY SYNDROME

Rodrigo Abrunhosa Collazo

August/2009

Advisor: Prof. Basílio de Bragança Pereira
Drª. Amália Faria dos Reis

Department: Program of Production Engineering

In Brazil, the leading cause of death is due to diseases of the circulatory system, including acute coronary syndrome. Thus, to develop cognitive tools that help to minimize the deaths from this disease is important for individuals and for the public health system in the country.

This study developed a machine learning capable to classify the risk of death of hospitalized patients with acute coronary syndrome at high and low, from the construction of the criteria for selection of the most important input variables to infer the outcome and the development of a mathematical model based on Support Vector Machines classifier.

The computational results indicate the prevalence of input variables age, creatinine, any prior revascularization and hypertension as the most relevant for the prediction of the outcome, and allow to obtain classifiers with accuracy greater than 90%.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

1^{o}AM	Tempo para o Primeiro Atendimento Médico
a	Acurácia (%)
AF	Atividade Física
AGT	Angiotensinogênio
Apo E	Apolipoproteína E
CZCL	Chen-Zhang-Chen-Li
Col-HDL	Colesterol-HDL
DAC	Doença Arterial Coronariana
e	Especificidade (%)
ECA	Enzima Conversora da Angiotensina
FBR	Função Base Radial
FC	Freqüência Cardíaca
h	Dimensão Vapnik-Chervonenkis
HAS	Hipertensão Arterial Sistêmica
IAM	Infarto Agudo do Miocárdio
IMC	Índice de Massa Corporal
IMP	Infarto do Miocárdio Prévio
KKT	Karush-Kuhn-Tucker
L	Função de Perda
MIFS-U	Mutual Information Feature Selector under Uniform Information Distribution
PCA	Principal Component Analysis
PCQ	Problema Convexo Quadrático
QRP	Qualquer Revascularização Prévia
R	Risco Verdadeiro
R_{emp}	Risco Empírico
R_{reg}	Risco Regularizado
RNA	Redes Neurais Artificiais
s	Sensibilidade
SCA	Síndrome Coronariana Aguda
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
VC	Vapnik-Chervonenkis
Λ	Equação Lagrangeana Generalizada
$\eth$	Coeficiente de Fragmentação

Ω	Função de Estabillização
Ψ	Grau de Confiança

1 INTRODUÇÃO

A questão da possibilidade de se produzir sistemas capazes de aprender a partir de um banco de dados iniciais tem sido objeto de intensas investigações e debates filosóficos e técnicos.

Do ponto de vista filosófico, o problema é conhecido sob o rótulo de "inferência indutiva", sendo constatado um ponto de inflexão na sua abordagem a partir da obra POPPER (2007). Nela, refuta-se a existência da indução pura, ou seja, da indução que ocorre no vazio. Para o autor, o *insight* indutivo só é possível a partir de um conhecimento prévio e/ou um acúmulo de experiências fenomênicas.

Do lado quantitativo, uma excelente síntese histórica pode ser encontrada em CRISTIANINI, SHAWE-TAYLOR (2006). Segundo estes autores, as primeiras abordagens associadas à questão do aprendizado podem ser referenciadas à regressão dos mínimos quadrados, proposta por Gauss. Contudo, foram as abordagens de classificação, sugeridas por Fisher nos anos 30, que propiciaram o embasamento cognitivo inicial para o desenvolvimento da maioria das análises e métodos que hoje existem nesta subárea do conhecimento humano. De acordo com eles, a representação mental, consubstanciada em um sistema capaz de aprender, só floresceu, em 1950, com Alan Turing. Desta forma, a partir da segunda metade do século passado, com o advento dos computadores digitais e do avanço da micro-eletrônica, associada à mecânica e à robótica, o debate técnico, tanto teórico quanto aplicado, em torno dessa problemática ganhou enorme ímpeto e valorosos resultados foram alcançados.

Nos anos 90, um novo sistema de aprendizado foi proposto por Vapnik e co-autores (BOSER et al., 1992, VAPNIK, 1995, CORTES, VAPNIK, 1995), o *Support Vector Machines* (SVM), mesclando teoria da dualidade, projeção em espaços de dimensão superior, teoria do aprendizado, otimização, algoritmos e estrutura de dados. Existem hoje métodos de programação matemática que viabilizam a resolução eficiente de instâncias com até 110.000 dados de treinamento (OSUNA et al., 1997).

Segundo BURGES (1998) e CRISTIANINI, SHAWE-TAYLOR (2006), o SVM tem se mostrado como uma técnica classificatória de elevado poder distintivo, de custo computacional relativamente baixo e de fácil implementação se comparada a outros recursos teóricos existentes, como, por exemplo, as redes neurais artificiais (RNA).

O treinamento do SVM consiste na seleção de um hiperplano que minimize o risco estrutural, a partir da resolução de um problema convexo quadrático (PCQ). Esta técnica de aprendizado quando associada à função núcleo permite a construção de classificadores não-lineares, através do mapeamento dos dados iniciais num espaço de dimensão superior ao original. Estas características garantem ao SVM uma boa generalização, ou seja, uma boa capacidade de predizer corretamente dados não relacionados na amostra de treinamento.

Os recentes avanços vivenciados com o SVM, no campo estatístico do aprendizado, agregado aos necessários desenvolvimentos na área da programação matemática, descortina um amplo horizonte de aplicações práticas que inclui, por exemplo, modelos para classificação financeira de carteiras de investimentos, pesquisas genéticas, sistemas de combates, modelagem de reações químicas complexas, categorização de textos, estimação de regressões, reconhecimento de padrões em imagens e estudos na área médica.

A relevância desta pesquisa se faz notar quando se constata que a principal causa de mortes no Brasil é devido às doenças do aparelho circulatório. As Doenças Isquêmicas do Coração (DIC) e as Doenças Cerebrovasculares (DCBV) são as principais causas de morte entre as Doenças do Aparelho Circulatório. Em 2004, as DIC causaram 86.791 óbitos no Brasil, 46.326 (53,4%) deles ocorreram na região Sudeste. Nessa região os estados com maior mortalidade por DIC foram São Paulo com 26.235 óbitos (56,6%) e Rio de Janeiro com 10.718 óbitos (23,1%). No estado do Rio de Janeiro, os municípios com maior número de óbitos por DIC em 2004 foram: 1º) Rio de Janeiro (5.175), 2º) Nova Iguaçu (491), 3º) Niterói (469) e 4º) São Gonçalo (449). No município de Niterói, segundo dados da Fundação Municipal de Saúde, as doenças do aparelho circulatório foram a principal causa de internação nos últimos 10 anos, assim como a principal causa de morte de 2003 a 2005 (REIS, 2007).

Dentre as doenças isquêmicas do coração, destacam-se as Síndromes Coronarianas Agudas (SCA), cuja, a mortalidade ainda é elevada, apesar dos recentes avanços terapêuticos, atingindo pessoas em idade produtiva, com graves conseqüências psicossociais e econômicas (REIS, 2007). A SCA inclui o infarto agudo do miocárdio (IAM), com e sem supradesnível do ST, e a angina instável. Esta doença se caracteriza pela “oclusão total ou parcial da artéria coronariana, acarretando na isquemia e/ou

necrose da área do miocárdio irrigada por aquela artéria" (REIS, 2007), a partir da "ruptura de uma placa coronariana instável" (REIS, 2007).

Com os avanços das pesquisas na área da genética, foram descobertos polimorfismos em vários genes responsáveis pela codificação de proteínas importantes na fisiologia cardiovascular, os quais têm sido apontados na literatura como possíveis marcadores de risco para a ocorrência de doença isquêmica do coração ou para uma maior mortalidade nestes pacientes já acometidos pela doença. Entre os inúmeros polimorfismos estudados em todo o mundo, em relação às doenças cardiovasculares, foram analisados na tese de doutorado de REIS (2007) pela UFRJ, os polimorfismos DI do gene da enzima conversora da angiotensina, M235T do gene do angiotensinogênio e E2/E3/E4 do gene da apolipoproteina E. Os pacientes avaliados no presente trabalho são oriundos do banco de dados desta tese e as variáveis referentes a estes genótipos foram incluídas entre as variáveis de entrada, juntamente com as demais variáveis clínicas e laboratoriais.

Assim, o objeto de estudo desta dissertação é o SVM aplicado à classificação do risco de morte de pacientes internados com síndrome coronariana aguda (SCA).

1.1 METODOLOGIA

A fim de obter um razoável rigor teórico-científico, a dissertação está alicerçada na teoria de aprendizado, a partir da interação de métodos indutivos e dedutivos. A indução se descortina ao objetivarmos predizer o risco de morte por síndrome coronariana aguda usando a ferramenta abstrata SVM. Por outro lado, a aproximação ao fenômeno médico em estudo nos levou a interessantes e relevantes resultados no campo teórico amplo dos métodos quantitativos, caracterizando o aspecto dedutivo mencionado.

O banco de dados utilizado nesta dissertação é o mesmo que foi empregado na tese de doutorado de REIS (2007). Estes dados foram coletados sob a coordenação desta autora e a partir de um estudo de coorte prospectivo com pacientes internados com SCA no município de Niterói, RJ. A coleta das informações foi feita no período entre julho/agosto de 2004 e junho/julho de 2005, a partir de pacientes internados em cinco hospitais, três públicos e dois privados, que tivessem idade superior a 20 anos e não

apresentassem sinais de doenças neoplásicas em fase terminal, politraumatismos e demência.

Este trabalho foi desenvolvido dentro do programa de pós-graduação em Engenharia de Produção da COPPE, sob um enfoque multidisciplinar, com colaborações do curso de doutorado em Clínica Médica – Área de Pesquisa Clínica da UFRJ e da Pós-Graduação em Ciências Cardiovasculares da Universidade Federal Fluminense (UFF).

1.2 OBJETIVOS

Os objetivos e os resultados esperados com a realização desta dissertação são os seguintes:

i. construir uma ferramenta computacional capaz de classificar o risco de morte, em alto ou baixo, de pacientes internados com SCA, empregando o SVM, a partir de variáveis clínicas, laboratoriais, genéticas e sociais do paciente;
ii. avaliar se fatores genéticos, particularmente os polimorfismos da enzima conversora da angiotensina (ECA), do angiotensinogênio (AGT) e da apolipoproteína E (Apo E), em conjunto com outras variáveis disponíveis no banco de dados, constituem fatores determinantes para a classificação do risco de morte de pacientes internados com SCA;
iii. desenvolver um algoritmo que permita a seleção das variáveis mais relevantes a serem utilizadas no SVM, de forma a reduzir o custo computacional e a produzir resultados consistentes e classificadores eficazes;
iv. selecionar uma função núcleo que quando associada ao classificador SVM minimize as taxas de erro; e
v. comparar as conclusões obtidas com os resultados atingidos a partir da aplicação de RNA em REIS (2007).

1.3 ESTRUTURA TEXTUAL

O texto a seguir está dividido em sete capítulos, que podem ser agrupados em três blocos. No primeiro bloco (capítulos 2, 3, e 4), são desenvolvidos os conceitos

necessários à fundamentação matemática e estatística do SVM. Assim, nos capítulos 2, 3 e 4 são analisadas, respectivamente, a técnica de nuclearização, a teoria de generalização e a metodologia lagrangeana para resolução de problemas quadráticos. No segundo bloco (capítulo 5), são discutidas algumas formulações matemáticas do SVM e sua implementação computacional. No último bloco (capítulos 6, 7 e 8), o capítulo 6 é dedicado à apresentação das variáveis disponíveis no banco de dados e dos métodos de seleção de variáveis empregados. No capítulo seguinte, são sintetizados e discutidos os resultados computacionais alcançados. No último capítulo, é feito um breve resumo do trabalho e das conclusões obtidas.

2 NÚCLEO

O núcleo permite um mapeamento implícito de uma variável $x=(x_1,...,x_n) \epsilon R^n$ em um espaço Hilbert H de dimensão superior. O espaço Hilbert é qualquer espaço linear, que atenda as seguintes condições:

1. tem que possuir produto interno definido; e
2. deve ser completo em relação à sua norma, ou seja, toda seqüência de Cauchy em H tem que convergi para um ponto pertencente ao próprio H.

Quando se considera espaços Hilbert de dimensão infinita, várias coisas interessantes e inopinadas podem ocorrer, inclusive algumas indesejáveis (SCHÖLKOPF, SMOLA, 2002). Para evitar estas últimas, vários autores exigem ainda que o espaço Hilbert seja separável (BURGES, 1998), isto é, que exista um subconjunto contável S C H tal que S seja denso. Um subconjunto S é denso quando cada elemento de H é o limite de uma seqüência em S. Em síntese, o espaço Hilbert H pode ser pensado como uma generalização do espaço euclidiano, pois pode ser usado qualquer produto interno e não apenas o escalar (SCHÖLKOPF e SMOLA, 2002).

Assim, podemos definir dois tipos de espaço Hilbert de especial interesse para este trabalho. O espaço Hilbert $l_2(\mu)$ é conjunto de pontos de dimensão infinita $x=(x_1,...,x_n,...)$ tal que sua norma euclidiana seja mensurável ($\Sigma_i\ x_i^2 < \infty$) e o produto interno entre x e y seja definido por $\langle x,y\rangle = \Sigma_i\ \mu_i \cdot x_i \cdot y_i < \infty$. Por outro lado, o espaço Hilbert L_2 (X) é o conjunto das funções contínuas f,g: X C $R^n \rightarrow$ R para as quais $||f|| = \int_X f^2(x)\cdot dx < \infty$, $||g|| = \int_X g^2(x)\cdot dx < \infty$ e $\langle f,g\rangle = \int_X f(x)\cdot g(x)\cdot dx$.

O núcleo pode ser definido como sendo uma função **K**: XxX → R, tal que para $x,z \epsilon X C R^n$, $\mathbf{K}(x,z) = \langle \Phi(x),\Phi(z)\rangle$, onde Φ: X → F é o mapeamento de X em um espaço Hilbert, chamado no contexto do processo de nuclearização do aprendizado como espaço característico F.

Assim, funções da forma $f(x) = \Sigma_{i=1}^{l} (\rho_i \cdot \langle \Phi(x_i),\Phi(x)\rangle) + b$ (l podendo ser infinito podem ser descritas por $f(x) = \Sigma_{i=1}^{l} \rho_i \cdot \mathbf{K}(x_i,x) + b$. Isto permite evitar a "maldição da dimensionalidade" no cálculo de f, porque este cálculo não exige que se conheça o mapeamento Φ propriamente dito. Vejamos um exemplo de mapeamento de R^2 em R^3. Sejam $x = (x_1,x_2)$, $y = (y_1,y_2) \epsilon R^2$ e $z = (z_1,z_2,z_3)$, $w = (w_1,w_2,w_3) \epsilon R^3$, tal que $z = \Phi(x) = (x_1^2, 2^{1/2}\cdot x_1 \cdot x_2, x_2^2)$ e $w = \Phi(y) = (y_1^2, 2^{1/2}\cdot y_1 \cdot y_2, y_2^2)$. Desta forma,

$$<w,z> = x_1^2 \cdot y_1^2 + 2 \cdot x_1 \cdot x_2 \cdot y_1 \cdot y_2 + x_2^2 \cdot y_2^2 = (x_1 \cdot y_1 + x_2 \cdot y_2)^2 = (<x,y>)^2 \equiv \mathbf{K}(x,y).$$

Ou seja, se $f(x) = \Sigma_{i=1}^{3}\ 5 \cdot <z^i,z> + 3$, onde $x^i = (i,i+1)$ e $x = (1,1)$, não precisaríamos calcular explicitamente os valores de z^i e z para depois obtermos o valor de f(x). Poderíamos simplesmente utilizar a função núcleo obtida:

$$f(x) = \Sigma_{i=1}^{3}\ 5 \cdot <z^i,z> + 3 = \Sigma_{i=1}^{3}\ 5 \cdot \mathbf{K}(x^i,x) + 3 = \Sigma_{=1}^{3}\ 5 \cdot [(1 \cdot 1 + 2 \cdot 1)^2 + (2 \cdot 1 + 3 \cdot 1)^2 + (3 \cdot 1 + 4 \cdot 1)^2] + 3 = 418.$$

Este mapeamento implícito efetuado pelo núcleo é extremamente importante, pois permite reduzir o custo computacional, tornando os problemas tratáveis, mesmo quando se trabalha em espaço de dimensões elevadas. Para se ter idéia da simplificação conseguida com o uso do núcleo, basta observar que, para núcleos polinomiais, o espaço característico F teria dimensão (r+n-1)! / [(r-1)!·n!], onde n é a dimensão da variável de entrada x. Este fato tornaria o cálculo de f rapidamente intratável. Por exemplo, para n = 30 e r = 5, temos que trabalhar num espaço característico de 46.376 dimensões.

Um fato importante a ser destacado é que para um dado núcleo **K** o mapeamento Φ não é único (BURGES, 1998). Pode-se obter outros mapeamentos mantendo-se ou não a dimensão do espaço característico F. Desta forma, no exemplo numérico apresentado anteriormente, poderíamos ter:

$$\Phi: R^2 \rightarrow R^3;\ \Phi(x) = 2^{-1/2}((x_1 - x_2)^2,\ 2 \cdot x_1 \cdot x_2,\ (x_1 + x_2)^2);\ \text{ou}$$

$$\Phi: R^2 \rightarrow R^4;\ \Phi(x) = (x_1^2,\ x_1 \cdot x_2,\ x_1 \cdot x_2,\ x_2^2).$$

Nos dois casos, o núcleo **K** continuaria igual a $\mathbf{K}(x,y) = (<x,y>)^2$. Todavia, ressaltamos mais uma vez que quando se trabalha com processos de nuclearização aplicados ao cálculo de funções não precisamos determinar este mapeamento, pois o mesmo já está embutido no próprio núcleo.

Agora, temos que caracterizar quando uma função f(x,y) é um núcleo **K**(x,y). O teorema de Mercer (CRISTIANINI, SHAWE-TAYLOR, 2006, SCHÖLKOPF, SMOLA, 2002) nos fornece esta definição.

Teorema de Mercer

Seja $X \subset R^N$ um conjunto compacto. Suponha que **K** seja uma função simétrica contínua tal que $\int_{XxX} f(x) \cdot f(y) \cdot dx \cdot dy \geq 0$, para todo $f \in L_2(X)$. Então pode-se expandir **K**(x,y) em uma série uniformemente convergente em XxX

$$\mathbf{K}(x,y) = \Sigma_{i=1}^{\infty}\ \lambda_i \cdot \Phi_i(x) \cdot \Phi_i(x),$$

onde Φ_i são as auto-funções do operador $(T_k f)(\cdot) = \int_X \mathbf{K}(\cdot,x) \cdot f(x) \cdot dx$, tal que $||\Phi_i|| = 1$ e λ_i são os auto-valores não-negativos associados. ■

A condição de Mercer no caso finito corresponde a afirmar que $\mathbf{K}(x,y)$ é um núcleo se e somente se a matriz $\mathbf{K}=(\mathbf{K}(x_i,y_j))^n_{i,j=1}$ for positiva semi-definida.

A proposição a seguir (CRISTIANINI, SHAWE-TAYLOR, 2006) permite a obtenção de núcleos complexos a partir de algumas operações simples.

Proposição 1

Sejam K_1 e K_2 núcleos em XxX, X C R^n, K_3 um núcleo em R^mxR^n, $\alpha \epsilon R^+$, f: X → R, Φ: X → R^m, M uma matriz semi-definida positiva nxn e p(x) um polinômio em X com coeficientes positivos. Então podemos afirmar que as seguintes operações resultam em funções núcleos **K**:

i. $\mathbf{K}(x,y) = \mathbf{K_1}(x,y) + \mathbf{K_2}(x,y)$
ii. $\mathbf{K}(x,y) = \mathbf{K_1}(x,y)\cdot\mathbf{K_2}(x,y)$
iii. $\mathbf{K}(x,y) = f(x)\cdot f(y)$
iv. $\mathbf{K}(x,y) = \mathbf{K_3}(\Phi(x),\Phi(y))$
v. $\mathbf{K}(x,y) = x^T My$
vi. $\mathbf{K}(x,y) = p(\mathbf{K_1}(x,y))$
vii. $\mathbf{K}(x,y) = \exp(\mathbf{K_1}(x,y))$ ■

Desta forma, consegue-se construir facilmente os núcleos usualmente encontrados em aplicações de SVM:

i. núcleo linear: $\mathbf{K}(x,y) = \langle x,y \rangle$
ii. núcleo polinomial: $\mathbf{K}(x,y) = (\text{escala}\cdot\langle x,y \rangle + \text{coef})^{\text{grau}}$
iii. núcleo tangente hiperbólica: $\mathbf{K}(x,y) = \tanh(\text{escala}\cdot\langle x,y \rangle + \text{coef})$
iv. núcleo função base radial (FBR) de Laplace: $\mathbf{K}(x,y) = \exp(-\sigma\cdot\|x - y\|)$
v. núcleo função base radial (FBR) de Gauss: $\mathbf{K}(x,y) = \exp(-\sigma\cdot\|x - y\|^2)$
vi. núcleo de Bessel do 1º tipo: $\mathbf{K}(x,y) = \text{Bessel}^n_{(v+1)}(\sigma\cdot\|x - y\|)/(\|x-y\|)^{-n(v+1)}$
vii. núcleo função Anova: $\mathbf{K}(x,y) = (\sum_{j=1}{}^n \exp(-\sigma\cdot\|x_j - y_j\|^2))^{\text{grau}}$

Nos núcleos apresentados, os termos "escala", "coef", "grau", σ, n e ν são hiperparâmetros que devem ser ajustados durante o treinamento do classificador por meio de técnicas estatísticas, como, por exemplo, validação cruzada.

Em KARATZOGLOU et al. (2006) e GUNN (1998), encontramos algumas sugestões de uso destas funções conforme o problema que se quer resolver. O núcleo linear é recomendável quando se lida com dados esparsos, facilmente encontrados em categorização de textos. O núcleo polinomial é um método popular para a modelagem não-linear, muito útil em processamento de imagens. Neste caso, prefere-se, em geral,

empregar um “coef” diferente de zero de forma a se evitar os problemas decorrentes da matriz hessiana se tornar zero. O núcleo tangente hiperbólica é normalmente usado como uma aproximação para RNA. Os núcleos FBR de Laplace e Gauss e o núcleo de Bessel atendem a propósitos gerais, quando não se dispõe de conhecimento a priori sobre os dados. Mais particularmente, o núcleo FBR de Laplace produz uma solução poligonal que é útil quando descontinuidades são admissíveis. Por fim, o núcleo FBR Anova é tipicamente empregado em problemas de regressão.

3 TEORIA DE GENERALIZAÇÃO

A teoria da generalização desloca o objetivo de um sistema de aprendizado da obtenção de uma hipótese consistente, ou seja, perfeitamente ajustada aos dados de treinamento, para a busca de uma hipótese que seja capaz de classificar corretamente dados que não fazem parte do conjunto de treinamento. Desta forma, procura-se limitar o grau de complexidade da hipótese de classificação, a fim de se evitar o seu sobreajustamento aos dados de treinamento.

Em todo modelo clássico de aprendizado, considera-se que as observações de treinamento e de teste são geradas independentemente e aleatoriamente de acordo com uma função de classificação fixa e desconhecida φ. No caso de classificadores binários, cada observação é representada por um par (x_j,y_j) ϵ Xx{-1,1}, X C R. A máquina de aprendizado deve aprender, portanto, o mapeamento $x_j \rightarrow y_j = f(x_j,\alpha)$, onde α é o parâmetro de ajuste da função f. O desejável é que esta máquina seja capaz de minimizar o risco verdadeiro, isto é, o erro médio verdadeiro ou a esperança do erro de teste:

$$R(f) = \int_{Xx\{-1,1\}} 0{,}5\cdot|y - f(x,\alpha)|\cdot\varphi(x,y)\cdot dx\cdot dy. \qquad \text{(eq. 1)}$$

Contudo, como φ é desconhecido, não se consegue calcular esta quantidade. A alternativa é utilizar um estimador conhecido como risco empírico, definido pela média da função de perda no conjunto de treinamento:

$$R_{emp}(f) = (1/l)\cdot\sum_{i=1}^{l} L(x_i,y_i,f(x_i)). \qquad \text{(eq. 2)}$$

Existem diversos tipos de função de perda. Neste trabalho, empregaremos a função de perda do erro absoluto

$$L(x_i,y_i,f(x_i)) = 0{,}5\cdot|y_i - f(x_i)|. \qquad \text{(eq. 3)}$$

Observamos facilmente que L assume valor 1 se houver erro no treinamento e 0, caso contrário.

Todavia, se procurarmos minimizar apenas o risco empírico provavelmente obteremos uma hipótese consistente e sobreajustada. Para evitar este efeito indesejável, devemos introduzir o conceito de risco regularizado:

$$R_{reg}(f) = C\cdot R_{emp}(f) + \Omega(f), \qquad \text{(eq. 4)}$$

onde Ω é uma função de estabilização que limita a complexidade da hipótese a ser encontrada ao final do treinamento e C o parâmetro não-negativo de regularização que

efetua o balanceamento entre o risco empírico $R_{emp}(f)$ e a função de suavizamento $\Omega(f)$. Por conseguinte, é através da minimização do risco de regularização que se consegue concretizar a estratégia de busca de uma alta capacidade de generalização da máquina treinada.

VAPNIK (1995) obteve um memorável resultado que estabelece um valor máximo para o risco verdadeiro. Assim, para funções de perda assumindo valores 0 e 1, numa amostra de treinamento de tamanho l, o limite abaixo é válido com probabilidade de 1-ψ:

$$R(f) \leq R_{emp}(f) + \{[h\cdot(\log(2\cdot l/h) +1) + \log(4/\psi)]/l\}^{1/2}, \qquad \text{(eq. 5)}$$

onde h é um inteiro não-negativo denominado de dimensão Vapnik-Chervonenkis (VC). O segundo termo do lado direito é conhecido como termo de confiança VC ou termo de capacidade. Por conseguinte, para uma dada amostra e estabelecido o grau de confiança ψ que se deseja obter, o termo de confiança VC passa ser função unicamente da dimensão VC. Se desejamos minimizar o risco verdadeiro podemos atuar sobre o seu limite superior, isto é , sobre o lado direito da desigualdade, conhecido como limite do risco. Assim, a função de suavizamento do risco regularizado pode ser escrita em termos da dimensão VC.

3.1 DIMENSÃO VC

A dimensão VC é um escalar que mede a capacidade de uma classe de conjuntos. Antes de a definirmos formalmente, é preciso que introduzamos o conceito de coeficiente de fragmentação. Sejam Δ uma classe de conjuntos, $T=\{x_1,...,x_n\}$ um conjunto dado qualquer e $N_\Delta(T)=\text{card}\{T\cap A \mid A \in \Delta\}$ o número de subconjuntos selecionados por Δ. Definimos coeficiente de fragmentação por

$$\vartheta(\Delta,n)=\max_{T\epsilon Tn} (N_\Delta(T)), \qquad \text{(eq. 6)}$$

onde T_n é a família de todos os conjuntos de tamanho n.

A dimensão VC representa a dimensão de uma classe Δ de conjuntos, VC(Δ), e é definida conforme a regra abaixo:

i. se $\vartheta(\Delta,n)=2^n$ para todo n, então $VC(\Delta)=\infty$;

ii. caso contrário, $VC(\Delta)=k$, tal que k é o maior k para o qual $s(\Delta,k)=2^k$.

Um exemplo seria de um conjunto de semi-espaços bidimensionais Δ. Qualquer três pontos não colineares podem ser selecionados por Δ, mas se n for igual a quatro não podemos afirmar o mesmo. Para um contra-exemplo, basta imaginar os vértices de um losango. Neste caso, os pares de vértices que definem as diagonais não podem ser selecionados por Δ. Logo VC(Δ)=3. Este resultado pode ser ampliado para uma família T C R^d de semi-espaços, que teria dimensão VC igual a d+1.

Além disso, Vapnik provou que, para classificadores lineares, o seguinte limite é válido (SCHÖLKOPF, SMOLA, 2002):

$$h \leq R^2 \cdot ||w||^2, \qquad \text{(eq. 7)}$$

onde w é o vetor normal ao hiperplano separador, tal que sua margem funcional seja 1, e R é o raio da menor esfera capaz de conter todos os dados de treinamento. Isto nos sugere definir a função suavizadora Ω(f) da seguinte maneira:

$$\Omega(f) = ||w||^2/2. \qquad \text{(eq.8)}$$

Notemos que a minimização do risco regularizado descrito por

$$R(f) \leq R_{emp}(f) + ||w||^2/2 \qquad \text{(eq. 9)}$$

nos leva obrigatoriamente a uma minimização do limite do risco. Ademais a função suavizada Ω(f) é convexa, fato fundamental para se garantir a obtenção de soluções ótimas globais no treinamento do SVM, conforme veremos a seguir.

3.2 MINIMIZAÇÃO DO RISCO ESTRUTURAL

Devemos distinguir risco regularizado e risco estrutural. A diferença é de natureza conceitual. O risco regularizado é um funcional que estabelece um compromisso entre a quantidade estatística que deveria ser minimizada e a quantidade que pode ser minimizada de maneira eficiente (SCHÖLKOPF e SMOLA, 2002). Por outro lado, a minimização do risco estrutural (BURGES, 1998, CRISTIANINI, SHAWE-TAYLOR, 2006, SCHÖLKOPF, SMOLA, 2002) é um princípio indutivo que estabelece uma estrutura sobre uma classe de funções Γ candidatas à hipótese classificadora e objetiva a minimização do risco regularizado durante o processo de treinamento.

A estrutura é construída de tal forma que se constitua uma seqüência aninhada de conjunto de hipóteses Γ_1 C ...C Γ_i C C $\Gamma_M = \Gamma$. Seja g_i a hipótese com menor erro de treinamento k_i no conjunto Γ_i e h_i a dimensão VC de Γ_i. Logo, como os conjuntos são

aninhados, $k_i \geq k_{i+1}$ e $h_i \leq h_{i+1}$. Desta forma, a minimização do risco estrutural busca indutivamente em cada conjunto Γ_i a hipótese g_i que apresenta o menor risco empírico e soma este funcional à quantidade suavizadora representada pela dimensão VC da respectiva classe, obtendo o menor risco regularizado para este subconjunto i de hipóteses. Se iniciarmos do conjunto de menor cardinalidade, o risco regularizado irá diminuir até um conjunto Γ_p, quando então começara a aumentar. Portanto, a solução buscada é a função $f_p \in \Gamma_p$ que minimiza o risco empírico para este conjunto de hipóteses.

É exatamente esta estratégia indutiva que propicia ao SVM uma boa generalização nas várias áreas em que tem sido aplicado.

4 TEORIA DE OTIMIZAÇÃO

A teoria de otimização é uma parte da matemática que se preocupa em caracterizar as soluções ótimas dos seguintes problemas:

minimizar	$f(x)$,	$x \in \Theta \subseteq R^n$
sujeito a	$g_i(x) = 0$,	$i = 1, ..., k$,
	$h_j(x) \leq 0$,	$j = 1, ..., m$,

onde f é chamada de função objetivo e as demais relações de restrições. Ela ainda se destina ao desenvolvimento de algoritmos que propiciem a obtenção eficiente e robusta dos pontos ótimos.

O objetivo deste capítulo restringe-se a descrever as propriedades básicas inerentes às soluções ótimas dos problemas cuja função objetivo é quadrática e convexa e as restrições são lineares. Esta classe de problemas matemáticos é conhecida como classe dos problemas quadráticos convexos (PQC) e corresponde à modelagem matemática que constitui o treinamento do SVM.

Na solução ótima x*, as restrições $h_j(x^*)$ que são iguais a zero são ditas ativas, caso contrário elas são denominadas inativas. Além disso, a formulação do problema, conforme o apresentamos acima, é conhecida como forma primal. Contudo, a solução dos PQC é mais fácil a partir da sua formulação dual, que é construída a partir da função lagrangeana generalizada. Basicamente, o que fazemos é simplificar as restrições, adicionando-as à função objetivo multiplicadas pelos multiplicadores de Lagrange, que são escalares irrestritos no caso de igualdade e não-negativos no caso de desigualdade do tipo "≤". Assim, a função lagrangeana generalizada é dada por

$$\Lambda(x,\lambda,\mu) = f(x) + \Sigma_{i=1}^{k} \lambda_i \cdot g_i(x) + \Sigma_{i=1}^{k} \mu_i \cdot h_i(x), \qquad \text{(eq. 10)}$$

com $\mu_i \geq 0$, $i = 1, ..., m$. Agora estamos em condições de definirmos o problema dual como sendo

maximizar $\beta(\lambda,\mu) = \inf_{x \in \Theta} \Lambda(x,\lambda,\mu)$

sujeito a $\mu_i \geq 0$, $i = 1, ..., m$.

De acordo com o teorema da dualidade fraca (BERTSEKAS,2008, BOYD, VANDENBERGHE, 2003, LUENBERGER, YE, 2006, NOCEDAL, WRIGHT, 2006), o valor da solução do primal é limitado inferiormente pelo valor da solução do dual, ou seja,

$$\inf\{f(x) \mid g_i(x) = 0,\ i = 1,...,k \text{ e } h_j(x) \leq 0,\ j = 1,...,m\} \geq \sup\{\beta(\lambda,\mu) \mid \mu_j \geq 0,\ j = 1,..., m\}. \quad \text{(eq. 11)}$$

A diferença entre os dois lados desta relação é chamado de *gap* de viabilidade.

Por outro lado, o teorema de Karush-Kuhn-Tucker (KKT) estabelece as condições necessárias para que um ponto seja solução ótima de um problema geral de otimização. São elas as seguintes:

i. $\partial\Lambda/\partial x\ (x^*,\lambda^*,\mu^*) = 0$

ii. $\mu_i \cdot h_i(x) = 0, \qquad i = 1,...,m$

iii. $g_i(x) = 0, \qquad i = 1,...,k$

iv. $h_i(x) \leq 0, \qquad i = 1,...,m$

v. $\mu_i \geq 0, \qquad i = 1,..., m.$

A primeira condição propicia a obtenção de um valor de x em função de λ e μ, que deverá ser substituído na função lagrangeana generalizada, resultando na função objetivo do problema dual. A segunda condição é conhecida na literatura como condição de complementaridade de folga e estabelece que se uma restrição de desigualdade é ativa seu correspondente multiplicador lagrangeano tem que ser nulo, caso contrário ele poderá assumir qualquer valor não-negativo.

Finalmente, nos PQC, as condições KKT são suficientes para se garantir que a solução é ótima global, anulando o *gap* de dualidade.

5 SUPPORT VECTOR MACHINE

Existem diversos tipos de SVM de acordo com o tipo de aplicação e de formulação da máquina de aprendizado. Trataremos aqui do C-SVM, que foi o primeiro SVM a ser desenvolvido, e do ν-SVM, que é uma variante do anterior. Eles se destinam a reconhecimento de padrões de dois tipos, contudo podem ser estendidos para multi-classes e problemas de regressão. Primeiramente, trataremos dos conceitos inerentes à taxonomia dos classificadores lineares, para depois tratarmos da formulação do C-SVM linear e não-linear, e, em seguida, do ν-SVM. Por último analisaremos uma técnica de solução de problemas duais para classificadores de duas classes conhecida como *Sequential Minimal Optimisation* (SMO).

5.1 CLASSIFICADORES LINEARES

Os classificadores lineares pressupõem a existência de um hiperplano que separa os dados em duas classes distintas, identificadas por -1 e 1, conforme podemos observar na Figura 1.

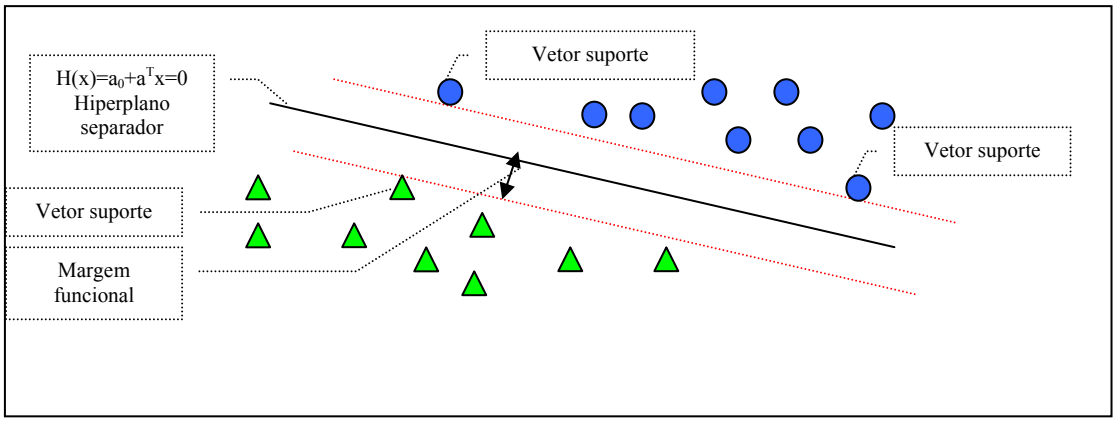

Figura 5.1: Esquema de um classificador linear

Obtida a expressão do hiperplano separador, a regra de classificação é feita de acordo com o seguinte mapeamento:

$f_{w,b}$: $X \subset R^n \rightarrow \{-1,1\}$

$x \rightarrow f_{w,b}(x) = \mathrm{sgn}(\langle w,x \rangle + b)$,

$$\text{onde } \operatorname{sgn}(t) = \begin{cases} 1, & \text{se } t \geq 0 \text{ e} \\ 0, & \text{caso contrário.} \end{cases}$$

Assim, a margem funcional de uma observação (x_i,y_i) em relação a um hiperplano (w,b) é definida como sendo a quantidade

$$\gamma_i = y_i \cdot (\langle w,x_i \rangle + b). \qquad \text{(eq. 12)}$$

Logo, o módulo da margem funcional de um dado representa a distância deste dado ao hiperplano separador. Dizemos módulo, pois esta medida assume valores negativos quando a classificação dada pelo SVM é errônea. A margem funcional de um hiperplano em relação a um conjunto de treinamento $T = \{(x_1,y_1),...,(x_i,y_i),...,(x_l,y_l)\}$ é dada pela menor margem encontrada neste conjunto de treinamento. Esta quantidade é chamada de margem funcional do hiperplano ou simplesmente margem funcional. Como podemos constatar, para um mesmo hiperplano poderíamos ter infinitas representações dependendo da escala λ que usamos para representá-lo: $\lambda \cdot \langle w,x \rangle + \lambda \cdot b = 0$, onde λ é um escalar positivo. Para contornarmos este inconveniente, fixaremos o fator de escala de tal forma que a margem funcional do hiperplano seja um. As observações de treinamento que tenham margem funcional unitária são chamadas de vetores suportes.

5.2 C-SVM LINEAR

5.2.1 DADOS LINEARMENTE SEPARÁVEIS

Inicialmente consideraremos um conjunto de treinamento com dados perfeitamente separáveis linearmente. Isto implica que, ao utilizarmos um classificador linear para separá-los, o erro empírico deve ser zero. Conforme estudado no Capítulo 3, para termos uma boa capacidade de generalização devemos minimizar o risco regularizado. Neste caso, esta quantidade reduz-se à função de suavizamento expressa em termos do vetor normal ao hiperplano separador com margem funcional unitária. Ou seja, devemos

minimizar $0,5 \cdot \|w\|^2$

sujeito a $y_i \cdot (\langle w,x_i \rangle + b) \geq 1$, $i=1,...,l$.

Esta formulação corresponde ao primal de um problema de formulação quadrática. A sua solução é obtida a partir do seu dual. Assim, devemos obter primeiramente a equação lagrangeana generalizada:

$$\Lambda(w,b,\alpha) = 0{,}5\cdot\|w\|^2 - \Sigma_{i=1}^{l} \alpha_i\cdot[y_i\cdot(<w,x_i> + b) - 1] \qquad \text{(eq. 13)}$$

Em seguida, impondo a primeira condição KKT, obtemos que:

$$\partial\Lambda/\partial w\ (w,b,\alpha) = w - \Sigma_{i=1}^{l} \alpha_i\cdot y_i\cdot x_i = 0 \qquad \text{e} \qquad \text{(eq. 14)}$$

$$\partial\Lambda/\partial b\ (w,b,\alpha) = -\Sigma_{i=1}^{l} \alpha_i\cdot y_i = 0. \qquad \text{(eq. 15)}$$

Logo,

$$w = \Sigma_{i=1}^{l} \alpha_i\cdot y_i\cdot x_i \qquad \text{e} \qquad \text{(eq. 16)}$$

$$\Sigma_{i=1}^{l} \alpha_i\cdot y_i = 0. \qquad \text{(eq. 17)}$$

Substituindo estes valores em $\Lambda(w,b,\alpha)$, temos que:

$$\beta(\alpha) = \inf_{w,b} \Lambda(w,b,\alpha) = \Sigma_{i=1}^{l} \alpha_i - 0{,}5\cdot\Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i\cdot\alpha_j\cdot y_i\cdot y_j\cdot<x_i,x_j>. \qquad \text{(eq. 18)}$$

Além disso, ainda de acordo com as condições KKT, os multiplicadores lagrangeanos devem ser não-negativos:

$$\alpha_i \geq 0,\ i = 1,...,l. \qquad \text{(eq. 19)}$$

Estamos, então, em condições de estruturarmos o problema dual como sendo:

maximizar $\Sigma_{i=1}^{l} \alpha_i - 0{,}5\cdot\Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i\cdot\alpha_j\cdot y_i\cdot y_j\cdot<x_i,x_j>$

sujeito a $\Sigma_{i=1}^{l} \alpha_i\cdot y_i = 0$

$\alpha_i \geq 0,\ i = 1,...,l.$

Ao resolvermos o dual, obteremos o valor de α e, por conseguinte, o valor de w através da equação 16. Resta, portanto, calcularmos o valor de b para termos a expressão do hiperplano separador. Esta quantidade pode ser obtida a partir das condições KKT de complementaridade de folga:

$$\alpha_i\cdot[y_i\cdot(<w,x_i> + b) - 1] = 0, \qquad i=1,...,l. \qquad \text{(eq. 20)}$$

Para isto, devemos selecionar uma observação de treinamento k, tal que seu correspondente multiplicador lagrangeano seja não-nulo. Ou seja, devemos selecionar um ponto do conjunto de treinamento que seja vetor suporte. Assim, teremos que:

$$b = y_i - <w,x_i>, \qquad \text{(eq. 21)}$$

onde i é um dado de treinamento tal que $\alpha_i \neq 0$. Segundo BURGES (1998), por questões de confiança numérica, é recomendável utilizar a média dos valores de b obtidos a partir do cálculo efetuado para todos os dados que sejam vetores suportes.

Finalmente, a regra de classificação é dada por

$$\tilde{y} = \text{sgn}\left(\Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \langle x_i, x \rangle + (1/\|SV\|) \cdot \Sigma_{i \in SV} \left(y_i - \Sigma_{j=1}^{l} \alpha_j \cdot y_j \cdot \langle x_j, x_i \rangle\right)\right), \quad \text{(eq. 22)}$$

onde $SV = \{i \mid \alpha_i \neq 0\}$ e $\|A\|$ representa a cardinalidade do conjunto A.

Na figura 2, temos um exemplo de SVM linear com dados perfeitamente separáveis, onde os quadrados representam os vetores suporte.

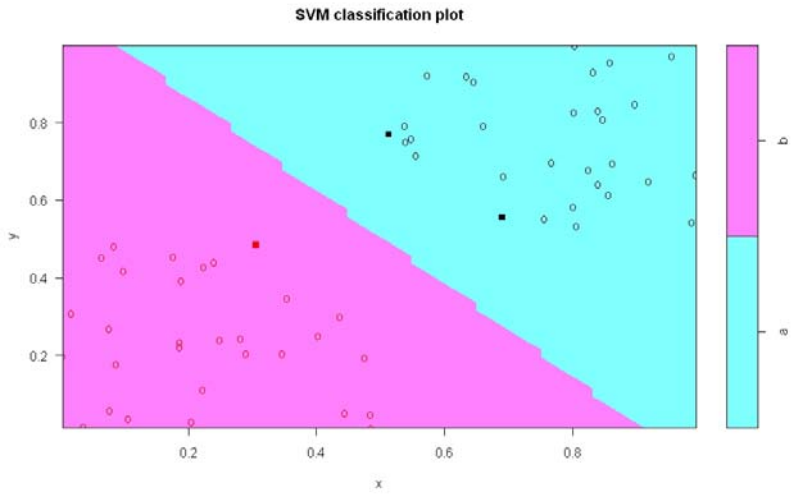

Figura 5.2: C-SVM linear com dados linearmente separáveis

5.2.2 DADOS NÃO-SEPARÁVEIS

No caso de uma amostra com dados não-separáveis, ao utilizarmos o C-SVM linear não é possível anular o risco empírico. Por outro lado, esta quantidade assume valores binários. Se a utilizássemos no cálculo do risco regularizado, a formulação primal do C-SVM se tornaria um problema quadrático misto, isto é, com variáveis inteiras e contínuas. Isto é um problema *NP-hard* e, portanto, de difícil solução. Por outro lado, sabemos que o risco regularizado é um funcional que aproxima as quantidades estatísticas que se desejar otimizar, a fim de deixá-las numa forma mais tratável. Assim, através de variáveis de folga ξ_i, que permitem a sobreposição entre a duas classes, pode-se aproximar o risco empírico pelo seguinte somatório:

$$R_{emp}(f) = (1/l) \cdot \sum_{i=1}^{l} \xi_i, \quad \text{(eq. 23)}$$

onde ξ_i é um escalar não nulo que representa a folga marginal da observação i, isto é:

$$y_i \cdot (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \text{ com } \xi_i \geq 0, \; i=1,...,l. \quad \text{(eq. 24)}$$

Desta forma, o problema primal do C-SVM continua um PQC, constituído apenas de variáveis contínuas e dado por:

minimizar $\quad 0{,}5 \cdot \|w\|^2 + (C/l) \cdot \sum_{i=1}^{l} \xi_i$

sujeito a $y_i \cdot (\langle w, x_i \rangle + b) \geq 1 - \xi_i$, i=1,...,l

$\xi_i \geq 0$, i=1,...,l.

A equação lagrangeana generalizada para este problema é a seguinte:

$$\Lambda(w,b,\xi,\alpha,\rho) = 0{,}5 \cdot \|w\|^2 + (C/l) \cdot \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i \cdot [y_i \cdot (\langle w, x_i \rangle + b) - 1 + \xi_i] - \sum_{i=1}^{l} \rho_i \cdot \xi_i \quad \text{(eq. 25)}$$

Aplicando as condições KKT, temos que:

$$\partial\Lambda/\partial w\ (w,b,\xi,\alpha,\rho) = w - \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot x_i = 0 \Rightarrow w = \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot x_i \quad \text{(eq. 26)}$$

$$\partial\Lambda/\partial b\ (w,b,\xi,\alpha,\rho) = -\sum_{i=1}^{l} \alpha_i \cdot y_i = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i \cdot y_i = 0 \quad \text{(eq. 27)}$$

$$\partial\Lambda/\partial\xi_i\ (w,b,\xi,\alpha,\rho) = C/l - \alpha_i - \rho_i = 0 \Rightarrow \rho_i = C/l - \alpha_i,\ i = 1,...,l \quad \text{(eq. 28)}$$

$$\alpha_i \cdot [y_i \cdot (\langle w, x_i \rangle + b) - 1 + \xi_i] = 0,\ i = 1,\ldots,l \quad \text{(eq. 29)}$$

$$\rho_i \cdot \xi_i = 0,\ i = 1,\ldots,l \Rightarrow^{\text{eq.28}} (C/l - \alpha_i) \cdot \xi_i = 0,\ i = 1,\ldots,l \quad \text{(eq. 30)}$$

$$\alpha_i \geq 0,\ i=1,...,l \quad \text{(eq. 31)}$$

$$\xi_i \geq 0,\ i=1,...,l. \quad \text{(eq. 32)}$$

$$\rho_i \geq 0,\ i=1,...,l \quad \text{(eq. 33)}$$

As equações 28 e 33 implicam que os multiplicadores lagrangeanos α_i são limitados superiormente a C/l. Além disso, a equação 30 estabelece que uma observação só terá folga não nula quando $\alpha_i = C/l$. Ainda, a partir das equações 26, 27 e 28, obtemos que:

$$\beta(\alpha,\rho) = \inf_{w,b,\xi} \Lambda(w,b,\xi,\alpha,\rho) = \sum_{i=1}^{l} \alpha_i - 0{,}5 \cdot \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle x_i, x_j \rangle. \quad \text{(eq. 34)}$$

Assim, a formulação dual é representada por:

maximizar $\sum_{i=1}^{l} \alpha_i - 0{,}5 \cdot \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle x_i, x_j \rangle$

sujeito a $\sum_{i=1}^{l} \alpha_i \cdot y_i = 0$

$0 \leq \alpha_i \leq C/l$, i = 1,...,l.

Como se pode constatar, as variáveis de folga desaparecem no dual e, portanto, podem ser determinadas de forma a sempre assegurarem uma solução viável para o primal.

O valor de b é determinado da mesma forma que foi obtido no caso do C-SVM linear com dados linearmente separáveis: calcula-se o valor médio de b a partir da equação 29 correspondente aos vetores suportes, tais que $0 < \alpha_i < C/l$ e $\xi_i = 0$. Assim,

$$b = (1 \backslash \|SV\|) \cdot \Sigma_{i \epsilon SV} \left(y_i - \sum_{j=1}^{l} \alpha_j \cdot y_j \cdot \langle x_j, x_i \rangle\right), \quad \text{(eq. 35)}$$

onde $SV = \{i |\ 0 < \alpha_i < C/l \text{ e } \xi_i = 0\}$.

Portanto, a predição é efetuada a partir da seguinte regra de classificação:

$$\tilde{y} = \text{sgn}\left[\sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \langle x_i, x \rangle + (1 \backslash \|SV\|) \cdot \Sigma_{i \epsilon SV} \left(y_i - \sum_{j=1}^{l} \alpha_j \cdot y_j \cdot \langle x_j, x_i \rangle\right)\right], \quad \text{(eq. 36)}$$

Esta formulação do C-SVM é sempre preferível, pois ela evita que se sobreajuste o classificador aos dados do conjunto de treinamento. Na figura abaixo, temos um exemplo de C-SVM linear com dados não-separáveis, onde os quadrados cheios representam os vetores suporte. Pode-se observar que houve três erros de treinamento: dois quadrados vermelhos sob o fundo branco e um quadrado verde sob o fundo azul.

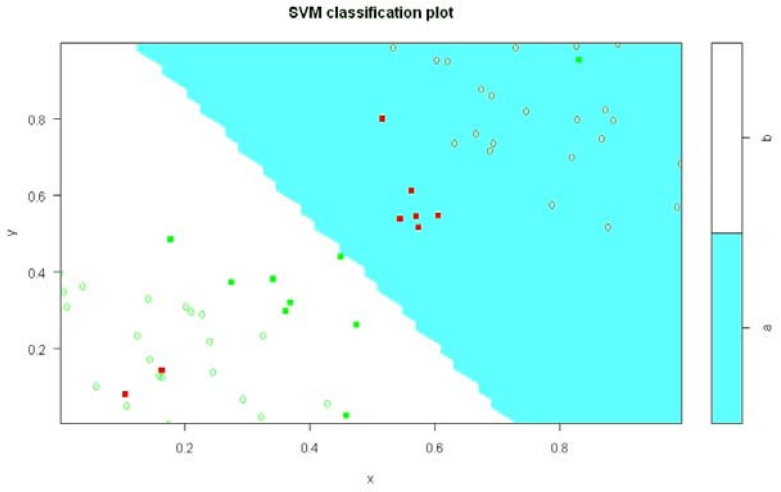

Figura 5.3: C-SVM linear com dados não separáveis

5.3 C-SVM NÃO LINEAR

Na maioria das aplicações práticas, dificilmente consegue-se obter resultados satisfatórios com classificadores lineares. A solução é mapear os dados de treinamento num espaço de dimensão superior (Φ: $R^n \rightarrow R^p$, $x \rightarrow \Phi(x)$, com $p > n$) e neste novo espaço característico aplicar um classificador linear. Neste caso, o classificador linear em um espaço de dimensão superior corresponderá a um classificador não-linear no espaço original (WASSERMAN, 2004). Esta estratégia é corroborada pelo teorema da separabilidade de Cover (PAN,2003, SCHÖLKOPF e SMOLA, 2002), que afirma que um problema de difícil classificação é mais provável de ser linearmente separável num espaço de dimensões mais elevadas.

O processo de projeção é facilitado no caso do SVM, pois toda manipulação das variáveis de entrada se dá através de produtos internos. Ou seja, é possível empregarmos a técnica de nuclearização, que permite o mapeamento implícito num espaço de dimensão mais elevada, sem incorrer na "maldição da dimensionalidade" e, por conseguinte, sem um custo computacional excessivo. Tem-se, então, que no espaço

característico de dimensão superior a formulação dual do problema de treinamento do SVM é a seguinte:

maximizar $\Sigma_{i=1}^{l} \alpha_i - 0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \Phi(x_i), \Phi(x_j) \rangle$

sujeito a $\Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$

$0 \leq \alpha_i \leq C/l,\ i = 1,...,l.$

Substituindo o produto interno pelo núcleo **K**, temos que:

maximizar $\Sigma_{i=1}^{l} \alpha_i - 0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i, x_j)$

sujeito a $\Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$

$0 \leq \alpha_i \leq C/l,\ i = 1,...,l.$

Notemos ainda que a condição de Mercer de semi-definição positiva para o núcleo garante que o problema continua um PQC e, portanto, de fácil solução e implementação relativamente simples.

A predição também pode ser feita sem a necessidade de se saber a função de mapeamento Φ. Esta novamente fica a cargo do núcleo. Portanto,

$$\tilde{y} = \operatorname{sgn}\left[\Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \langle \Phi(x_i), \Phi(x_j) \rangle + (1 \backslash \|SV\|) \cdot \Sigma_{i \epsilon SV} \left(y_i - \Sigma_{j=1}^{l} \alpha_j \cdot y_j \cdot \langle \Phi(x_i), \Phi(x_j) \rangle\right)\right]$$

$$= \operatorname{sgn}\left[\Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \mathbf{K}(x_i, x_j) + (1 \backslash \|SV\|) \cdot \Sigma_{i \epsilon SV} \left(y_i - \Sigma_{j=1}^{l} \alpha_j \cdot y_j \cdot \mathbf{K}(x_i, x_j)\right)\right], \qquad \text{(eq. 37)}$$

Na figura 4, encontramos um exemplo de SVM não linear obtido com um núcleo polinomial de grau dois, onde os quadrados cheios representam os vetores suportes. Neste treinamento, pode-se constatar que o erro empírico foi nulo.

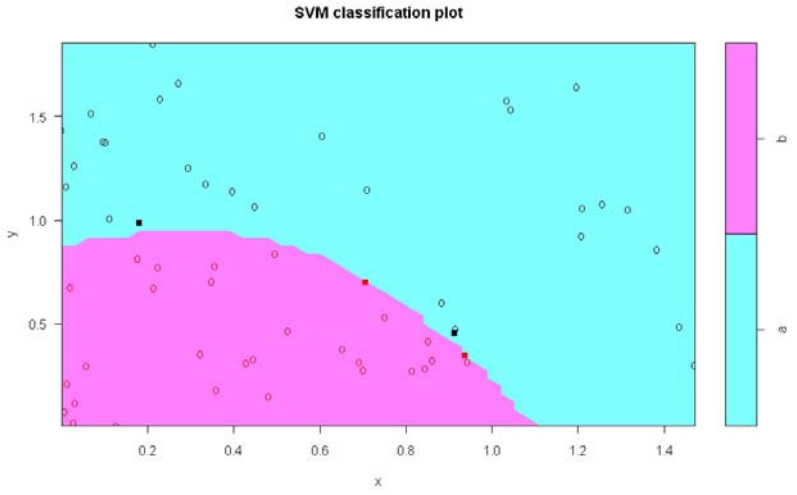

Figura 5.4: C-SVM não linear

5.4 ν-SVM

No treinamento do C–SVM, uma dificuldade tácita é o ajuste do parâmetro C, que não possui nenhum método de ajuste prévio e nenhuma interpretação do seu significado que facilite a intuição do seu valor. Para contornar este inconveniente, foi proposta uma formulação alternativa denominada ν-SVM (SMOLA et al., 2000) , que trabalha com um plano de margem funcional γ e substitui o parâmetro C pelo parâmetro *v*. No ν-SVM, o problema primal é colocado da seguinte forma:

minimizar $0{,}5\cdot\|w\|^2 - \nu\cdot\gamma + (1/l)\cdot\sum_{i=1}^{l}\xi_i$

sujeito a $y_i\cdot(\langle w,\Phi(x_i)\rangle + b) \geq \gamma - \xi_i,\ i=1,...,l$

$\gamma \geq 0$

$\xi_i \geq 0,\ i=1,...,l.$

Embora a construção da função objetivo do primal do ν-SVM não tenha uma intuição tão direta quanto a do C-SVM, prova-se que o ν-SVM e o C-SVM possuem a mesma função de decisão no ótimo, desde que a margem funcional ótima γ* do hiperplano ν-SVM seja positiva e que $C = 1/l\cdot\gamma^*$. Este resultado estabelece uma equivalência forte entre os dois tipos de classificadores. Ademais, a proposição abaixo, estabelecida em CHEN et al. (2005) e em SCHÖLKOPF, SMOLA (2002), apresenta um interessante significado para o parâmetro ν, que torna o ajuste do mesmo durante a fase de treinamento do classificador mais fácil de ser realizado em comparação com o parâmetro de regularização C.

Proposição 2

Se o valor ótimo γ* da margem funcional for estritamente positivo após se treinar um classificador ν-SVM com um núcleo **K** em uma amostra de treinamento T, então:

i. o parâmetro ν representa um limite superior para fração dos erros de treinamento;
ii. o parâmetro ν representa um limite inferior na fração dos vetores suportes; e
iii. com probabilidade 1, o parâmetro ν iguala assintoticamente as frações dos erros de treinamento e dos vetores suporte. ■

Deduzamos agora a formulação dual do ν-SVM. Nesse sentido, a equação lagrangeana correspondente é dada por:

$$\Lambda(w,b,\gamma,\xi,\alpha,\eta,\rho) = 0{,}5\cdot\|w\|^2 - \nu\cdot\gamma + (1/l)\cdot\sum_{i=1}^{l}\xi_i - \Sigma_{i=1}^{l}\alpha_i\cdot[y_i\cdot(\langle w, \Phi(x_i)\rangle + b) - \gamma + \xi_i] - \eta\cdot\gamma - \Sigma_{i=1}^{l}\rho_i\cdot\xi_i \quad \text{(eq. 38)}$$

Aplicando as condições KKT, obtemos que:

$\partial\Lambda/\partial w\ (w,b,\gamma,\xi,\alpha,\eta,\rho) = w - \Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \Phi(x_i) = 0 \Rightarrow w = \Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \Phi(x_i)$ (eq. 39)

$\partial\Lambda/\partial b\ (w,b,\gamma,\xi,\alpha,\eta,\rho) = -\Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0 \Rightarrow \Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$ (eq. 40)

$\partial\Lambda/\partial\gamma\ (w,b,\gamma,\xi,\alpha,\eta,\rho) = -\nu + \Sigma_{i=1}^{l} \alpha_i - \eta = 0 \Rightarrow \Sigma_{i=1}^{l} \alpha_i - \eta = \nu$ (eq. 41)

$\partial\Lambda/\partial\xi_i\ (w,b,\gamma,\xi,\alpha,\eta,\rho) = 1/l - \alpha_i - \rho_i = 0 \Rightarrow \alpha_i + \rho_i = 1/l,\ i = 1,...,l$ (eq. 42)

$\alpha_i \cdot [y_i \cdot (\langle w, \Phi(x_i)\rangle + b) - \gamma + \xi_i] = 0,\ i = 1,\ldots,l$ (eq. 43)

$\eta \cdot \gamma \geq 0$ (eq. 44)

$\rho_i \cdot \xi_i = 0,\ i = 1,\ldots,l \Rightarrow^{eq.28} (1/l - \alpha_i) \cdot \xi_i = 0,\ i = 1,\ldots,l$ (eq. 45)

$\alpha_i \geq 0,\ i=1,...,l$ (eq. 46)

$\gamma \geq 0$ (eq. 47)

$\eta \geq 0$ (eq. 48)

$\xi_i \geq 0,\ i=1,...,l.$ (eq. 49)

$\rho_i \geq 0,\ i=1,...,l$ (eq. 50)

A partir das equações 39, 40, 41 e 42, temos que:

$$\beta(\alpha,\rho) = \inf_{w,b,\gamma,\xi} \Lambda(w,b,\gamma,\xi,\alpha,\eta,\rho) = -0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \langle \Phi(x_i), \Phi(x_j) \rangle =$$
$$-0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i,x_j). \quad \text{(eq. 51)}$$

Assim, a formulação dual é dada por:

maximizar $\Sigma_{i=1}^{l} \alpha_i - 0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i,x_j)$

sujeito a $\Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$

$\Sigma_{i=1}^{l} \alpha_i \geq \nu$

$0 \leq \alpha_i \leq 1/l,\ i = 1,...,l.$

Para calcular o valor de b, usa-se a equação 43 correspondente aos vetores suportes, tais que $0 < \alpha_i < 1/l$ e $\xi_i = 0$. Assim, obtemos que:

$\Sigma_{j=1}^{l} \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i,x_j) + b - \gamma = 0$, se $y_i = 1$ (eq. 52)

$\Sigma_{j=1}^{l} \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i,x_j) - b - \gamma = 0$, se $y_i = -1$ (eq. 53)

Resolvendo este sistema e empregando a média do valor de $\Sigma_{j=1}^{l} \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i,x_j)$ em relação aos conjuntos SVP = {i| $0 < \alpha_i < 1/l$, $\xi_i = 0$ e $y_i = 1$} e SVN = {i| $0 < \alpha_i < 1/l$, $\xi_i = 0$ e $y_i = -1$} para evitar erros numéricos, temos que:

$$b = 0{,}5 \cdot [(1\backslash\|SVN\|) \cdot \Sigma_{i \epsilon SVN} \Sigma_{j=1}^{l} \alpha_j \cdot y_i \cdot y_j \cdot K(x_i,x_j) - (1\backslash\|SVP\|) \cdot \Sigma_{i \epsilon SVP} \Sigma_{j=1}^{l} \alpha_j \cdot y_i \cdot y_j \cdot K(x_i,x_j)]$$
(eq. 54)

A regra de classificação é dada, então, por:

$\tilde{y} = \text{sgn}\ (\Sigma_{i=1}^{l} \alpha_i \cdot y_i \cdot \mathbf{K}(x_i,x_j) + b).$ (eq. 55)

5.5 SMO

Existem diversos algoritmos para a resolução do PQC, que podem ser utilizados para se encontrar a solução ótima do problema dual referente ao treinamento do ν-SVM. Neste trabalho empregaremos o pacote **kernlab** do *software* **R**, que utiliza o algoritmo SMO para otimização do problema quadrático do ν-SVM. Este algoritmo encontra analiticamente a solução global, prescindindo do uso de métodos numéricos. Isto é possível levando a técnica de decomposição ao extremo e, por conseguinte, otimizando a cada passo apenas dois multiplicadores lagrangeanos. Assim, consegue-se manter sempre a viabilidade da solução, ou seja, $\Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$, $\Sigma_{i=1}^{l} \alpha_i \geq \nu$ e $\leq \alpha_i \leq 1/l$, i = 1,...,l.

A cada iteração, escolhe-se heuristicamente dois multiplicadores lagrangeanos para serem otimizados, mantendo todos os demais constantes. Os dois multiplicadores selecionados são otimizados analiticamente e o processo prossegue até que não haja mais melhorias a serem realizadas, ou seja, quando o *gap* de dualidade se anular ou todas as condições KKT de complementaridade de folga tiverem sido satisfeitas.

Para se aplicar o algoritmo SMO no treinamento do ν-SVM, torna-se mister recorremos à afirmação que para qualquer ν, existe pelo menos uma solução ótima que satisfaz $\Sigma_{i=1}^{l} \alpha_i = \nu$ (CHANG e LI, 2001). Assim, o problema dual pode ser simplificado por:

maximizar $\quad \Sigma_{i=1}^{l} \alpha_i - 0{,}5 \cdot \Sigma_{i=1}^{l} \Sigma_{j=1}^{l} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \mathbf{K}(x_i, x_j)$

sujeito a $\quad \Sigma_{i=1}^{l} \alpha_i \cdot y_i = 0$

$\Sigma_{i=1}^{l} \alpha_i = \nu$

$0 \leq \alpha_i \leq 1/l$, i = 1,...,l.

Neste caso, quando se otimiza apenas dois multiplicadores lagrangeanos α_i e α_j, temos que:

$$\alpha_i^N \cdot y_i + \alpha_j^N \cdot y_j = \alpha_i^A \cdot y_i + \alpha_j^A \cdot y_j \qquad \text{(eq. 56)}$$

$$\alpha_i^N + \alpha_j^N = \alpha_i^A + \alpha_j^A, \qquad \text{(eq. 57)}$$

onde o N sobrescrito indica os novos valores de α_i e α_j calculados e o A sobrescrito representa os antigos valores de α_i e α_j. Analisando o sistema acima, chega-se a conclusão que se $y_i \neq y_j$, os valores de α_i e α_j permanecem constantes. Portanto, devemos selecionar multiplicadores lagrangeanos i e j, tais que $y_i = y_j$. Neste particular, a equação 56 se torna idêntica à equação 57. Além disso, α_i e α_j devem satisfazer à

restrição de caixa, isto é, $0 \leq \alpha_i, \alpha_j \leq 1/l$. A partir desta condição e da equação 57, é fácil constatar que para i e j tais $y_i = y_j$

$$LI \leq \alpha_j^N \leq LS, \qquad \text{(eq. 58)}$$

onde

$$LI = \max(0, \alpha_i^A + \alpha_j^A - 1/l), \qquad \text{(eq. 59)}$$

$$LS = \min(1/l, \alpha_i^A + \alpha_j^A). \qquad \text{(eq. 60)}$$

Prova-se (CRISTIANINI, SHAWE-TAYLOR, 2006 e SCHÖLKOPF, SMOLA, 2002) que, atuando-se em apenas dois multiplicadores lagrangeanos, a função objetivo é maximizada através do procedimento abaixo:

i. calcular as quantidades E_i e E_j:

ii. $E_k = f(x_k) - y_k = \Sigma_{r=1}^{l} \alpha_r \cdot y_r \cdot \mathbf{K}(x_r, x_k) + b - y_k$, $k = i,j$ e (eq. 61)

iii. $\varsigma = K(x_i,x_i) + K(x_j,x_j) - 2 \cdot K(x_i,x_j)$ (eq. 62)

iv. obter um valor inicial $\alpha_j^{N\_i}$ para α_j^N segundo a fórmula:

v. $\alpha_j^{N\_i} = \alpha_j^A + y_j \cdot (E_1 - E_2)/\varsigma$ (eq. 63)

vi. restringir α_j^N aos limites anteriormente calculados:

$$\alpha_j^N = \begin{cases} LS, & \text{se } \alpha_j^{N\_i} > LS \\ \alpha_j^{N\_i}, & \text{se } LI \leq \alpha_j^{N\_i} \leq LS \\ LI, & \text{se } \alpha_j^{N\_i} < LI. \end{cases}$$

vii. calcular α_i^N:

$\alpha_i^N = \alpha_i^A + y_i \cdot y_j \cdot (\alpha_i^A - \alpha_j^N)$. (eq. 65)

Apesar de precisar de maior número de iterações para convergir, o SMO é extremamente rápido, uma vez que realiza poucas operações de fácil implementação. Além disso, o SMO não precisa que a matriz núcleo fique armazenada, embora a sua manutenção na memória do computador resulte em acréscimo de velocidade do algoritmo.

6 AMOSTRA

O banco de dados dispõe de 411 indivíduos observados sob o enfoque de 25 variáveis, sendo que apenas 37 destes evoluíram para óbito. Porém nem todas as variáveis foram coletadas em todos os indivíduos. Assim, o emprego de todas as variáveis resulta num diminuto número de dados de treinamento: 231 indivíduos, sendo que apenas 16 evoluíram para óbito. Assim, torna-se mister selecionar as variáveis que sejam determinantes para ocorrência do desfecho morte, de forma a aumentar o número de dados de treinamento e, mormente, auxiliar na elaboração de medidas terapêuticas mais individualizadas, na realização de prognósticos eficientes e na redução da mortalidade.

Dada à explosão de possibilidades combinatórias que existem ao selecionarmos as variáveis que devem compor o modelo a ser construído, é fundamental aplicar e/ou desenvolver um algoritmo que forneça de maneira consistente e relativamente rápida as variáveis mais significativas a serem usadas no SVM.

Desta forma, o presente capítulo terá duas subseções. Na primeira, serão descritas as variáveis disponíveis no banco de dados. Na segunda, discutiremos os critérios de seleção de variáveis utilizados.

6.1 VARIÁVEIS

O banco de dados do presente trabalho dispõe de 25 variáveis selecionadas do banco de dados de REIS (2007) que podem ser agrupadas em seis categorias: variáveis antropométricas, sociais e hábitos de vida, variáveis de história prévia cardiovascular, variáveis clínicas e laboratoriais na admissão hospitalar, variáveis de diagnóstico, variáveis genéticas e variável de desfecho.

6.1.1 VARIÁVEIS ANTROPOMÉTRICAS, SOCIAIS E HÁBITOS DE VIDA

6.1.1.1 IDADE

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

6.1.1.2 ÍNDICE DE MASSA CORPORAL (IMC)

O IMC (Kg/m^2) expressa a 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.

6.1.1.3 SEXO

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

6.1.1.4 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.

6.1.1.5 ATIVIDADE FÍSICA (AF)

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

6.1.1.6 TABAGISMO

Os indivíduos foram categorizados em três classes:

i. não tabagistas: os fumantes passivos e as pessoas que nunca fumaram ou fumaram menos de cinco cigarros por menos de cinco anos;
ii. ex-tabagistas: pessoas que pararam de fazer uso do tabaco há mais de seis meses; e
iii. tabagistas atuais e eventuais: pessoas que fizeram uso regular do tabaco durante os seis meses que antecederam a coleta dos seus dados.

6.1.2 VARIÁVEIS DE HISTÓRIA PRÉVIA CARDIOVASCULAR

6.1.2.1 INFARTO DO MIOCÁRDIO PRÉVIO (IMP)

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

6.1.2.2 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 por escrito dos procedimentos ou relatos detalhados dos pacientes, que deveriam possuir evidência de cicatriz cirúrgica compatível no caso de CRM.

6.1.2.3 HISTÓRIA FAMILIAR DE DOENÇA ARTERIAL CORONARIANA (DAC)

A história familiar de DAC é uma variável bi-classe, que foi coletada em parentes de primeiro e segundo graus (pais, filhos, irmãos, tios e avós). Era considerada história positiva no caso de parentes com relato de diagnóstico de infarto do miocárdio ou angina pectoris ou relato de realização de angioplastia coronariana ou CRM.

6.1.3 VARIÁVEIS CLÍNICAS E LABORATORIAIS NA ADMISSÃO HOSPITALAR

6.1.3.1 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.

6.1.3.2 TEMPO PARA O PRIMEIRO ATENDIMENTO MÉDICO (1ºAM)

Esta variável (1ºAM) 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.

6.1.3.3 FREQÜÊNCIA CARDÍACA (FC)

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

6.1.3.4 CLASSE KILLIP

Os indivíduos foram categorizados em quatro classes na sua admissão hospitalar. Segundo REIS (2007), os critérios para esta classificação foram os seguintes:

i. KILLIP I: ausência de sinais de insuficiência cardíaca;
ii. KILLIP II: presença de estertores em bases pulmonares ou de terceira bulha na ausculta cardíaca;
iii. KILLIP III: presença de edema agudo de pulmão (estertores pulmonares maiores que 50%); e
iv. KILLIP IV: choque cardiogênico.

6.1.3.5 CREATININA

A creatinina é uma variável contínua que foi coleta a partir do primeiro exame de sangue realizado pelo paciente após dar entrada no hospital.

6.1.4 VARIÁVEIS DE DIAGNÓSTICO

6.1.4.1 HIPERTENSÃO ARTERIAL SISTÊMICA (HAS)

A HAS é uma variável bi-classe, na qual foram consideradas como portadores de HAS os 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 medicação anti-hipertensiva até a data em que foram admitidos no hospital. Os demais indivíduos foram classificados como não tendo HAS.

6.1.4.2 COLESTEROL TOTAL ELEVADO

O colesterol total elevado é representado por uma variável bi-classe, que caracteriza indivíduos com colesterol total acima de 200 mg/dl. Os demais são considerados como tendo colesterol total dentro dos limites normais.

6.1.4.3 TRIGLICERÍDEOS ELEVADOS

O triglicerídeo constitui uma variável bi-classe, na qual os indivíduos com mais de 150 mg/dl são considerados como tendo triglicerídeos elevados e separados dos demais, avaliados como tendo valores dentro dos limites normais.

6.1.4.4 COLESTEROL-HDL BAIXO (Col-HDL)

O colesterol-HDL baixo é uma variável bi-classe, que distingue os indivíduos com colesterol-HDL inferior a 40 mg/dl dos demais, que teriam valores iguais ou superiores ao normal.

6.1.5 VARIÁVEIS GENÉTICAS

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

i. polimorfismo genético: são mutações presentes em mais de 1% da população, tornando possível a observação de diferentes formas alélicas de um mesmo lócus gênico (REIS, 2007);
ii. gene: constitui a peça central da hereditariedade. Ele é um segmento da cadeia do DNA, localizado no lócus do cromossomo, e é responsável pela produção de uma proteína ou pela determinação de uma característica do indivíduo (MONÇORES et al., 2008);
iii. 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 alelos diferentes cuja combinação determinará a cor azul, verde, castanha ou preta dos olhos (BROWN, 1999); e
iv. 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 (BROWN, 1999).

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 e ao imaginarmos cada alelo como um vetor. Ou seja, se, no exemplo acima, o alelo R contribuir para uma determinada característica em grau elevado, o alelo S em grau reduzido e o alelo T a desfavorecer, o SVM atribuirá pesos a cada um destes alelos. Assim, cada alelo pode ser visto como um vetor de mesma direção, porém com intensidade e sentido que é dependente do sinal do peso atribuído pelo classificador. Por exemplo, uma pessoa que possua o genótipo RS tem uma propensão maior a apresentar a referida característica do que o indivíduo que tenha o genótipo ST. Esta situação seria capturada pelo modelo, pois os alelos R e S podem ser vistos como vetores que se somam em módulo e os alelos S e T como vetores que se subtraem em módulo. Assim, as variáveis genéticas são, na verdade, variáveis bi-classes representativas da presença ou não nos indivíduos dos alelos referentes aos polimorfismos observados em cada gene em estudo. Foram consideradas sete variáveis genéticas:

i. alelo D do polimorfismo do gene da enzima conversora da angiotensina I (ECA);
ii. alelo I do polimorfismo do gene da enzima conversora da angiotensina I (ECA);
iii. alelo M do polimorfismo M235T do gene do angiotensinogênio (AGT);
iv. alelo T do polimorfismo M235T do gene do angiotensinogênio (AGT);
v. alelo E2 do polimorfismo do gene da apolipoproteína E (APO E);
vi. alelo E3 do polimorfismo do gene da apolipoproteína E (APO E); e
vii. alelo E4 do polimorfismo do gene da apolipoproteína E (APO E).

6.1.6 VARIÁVEL DE DESFECHO

A variável de desfecho possui duas classes e identifica se ocorreu o óbito ou não do paciente.

6.2 CRITÉRIOS PARA SELEÇÃO DE VARIÁVEIS

Foram empregados neste trabalho quatro métodos para seleção de variáveis: o método desenvolvido por CHEN et al. (2008) e três modificações dele. No restante

desta seção, chamaremos o primeiro critério de critério CZCL (a abreviação corresponde às iniciais dos sobrenomes dos seus proponentes) e os outros três de critérios CZCL Adaptado, CZCL Dual e CZCL Adaptado Dual. Além disso, consideraremos uma amostra de treinamento $T = \{(x_1,y_1),...,(x_l,y_l)\}$, com $x_i \in R^n$ e $y_i \in \{-1,1\}$, onde a notação x_{ij} indica o valor que a variável j assume na observação i.

6.2.1 CRITÉRIO CZCL

O critério CZCL destina-se a otimizar o balanceamento entre a experiência humana e a sensibilidade dos dados. Desta forma, propõe-se que as variáveis a serem selecionadas sejam ordenadas de acordo com os pesos que as mesmas obtenham na seguinte formulação:

$F_k = g_1 \cdot VA_k(x_k,y) + g_2 \cdot S_k$, onde (eq. 66)

g_1 e g_2 → parâmetros positivos para efetuar o compromisso entre o conhecimento humano e a informação presente na amostra de treinamento;

VA_k → valor atribuído a cada variável k por especialista da área; e

S_k → critério de sensibilidade dos dados para cada variável k.

Como neste estudo, não há um consenso estabelecido quanto à importância absoluta e/ou relativa de cada variável para o desfecho considerado e o que queremos é exatamente destacar as prevalências a partir das informações imanentes ao banco de dados considerado, desprezaremos a parcela VA_k da formulação proposta e, por conseguinte, os parâmetros g_1 e g_2.

O valor de S_k deve, segundo os autores proponentes do método, atender a duas hipóteses básicas:

i. se uma pequena variação da variável de entrada corresponde a uma grande variação da variável de saída, então a variável é considerada sensível; e
ii. se uma pequena alteração da variável de entrada corresponde a uma pequena alteração da variável de saída, então a variável é considerada insensível.

Assim, eles propõem que:

$S_k = [\max_{k\in\{1,...,n\}}(T_k) - T_k] / [\max_{k\in\{1,...,n\}}(T_k) - \min_{k\in\{1,...,n\}}(T_k)]$, onde (eq. 67)

$T_k = \Sigma_{r=1}^{l}\Sigma_{s=1;s\neq r}^{l} d(y_r,y_s) / d_k'(x_r,x_s)$, (eq. 68)

$d_k'(x_r,x_s) = [d^2(x_r,x_s) - d_k^2(x_r,x_s)]^{1/2}$, (eq. 69)

$d^2(x_r,x_s) = \Sigma_{i=1}^{n}(x_{ri} - x_{si})^2$ (eq. 70)

$d_k^2(x_r,x_s) = (x_{rk} - x_{sk})^2$, e (eq. 71)

$$d(y_r,y_s) = |y_r - y_s| \quad \text{(eq. 72)}$$

Como podemos perceber, S_k corresponde apenas a uma normalização de T_k e se destina também a compatibilizar a tendência de T_k com a tendência da parcela VA_k. Assim, como a parcela VA_k já foi descartada para fins deste trabalho, não há a necessidade de se calcular S_k e, por conseguinte, podemos nos concentrar apenas em T_k. Portanto, quanto menor o valor de T_k, mais relevante é a variável k para o desfecho.

Segundo os seus criadores, o método CZCL possui as seguintes vantagens:

i. considera, para seleção das variáveis mais relevantes, o relacionamento entre as variáveis de entrada e a variável de desfecho;
ii. emprega as variáveis de entrada primitivas e não uma transformação das mesmas, como no método *Principal Component Analysis* (PCA);
iii. não necessita de um grande número de dados;
iv. não requer que os dados obedeçam a qualquer distribuição estatística;
v. é capaz de capturar relações não lineares entre as variáveis de entrada e saída; e
vi. fácil de ser implementado e de baixo custo computacional.

6.2.2 CRITÉRIO CZCL ADAPTADO

No caso em estudo, a variável de saída é bi-classe. Desta forma, só interessam para o cômputo do valor do somatório correspondente a T as parcelas calculadas entre os dados de treinamento que tenham classificação distinta, pois do contrário o valor da parcela será zero. Ficamos então com

$$T_k = \Sigma_{r|y=1}\Sigma_{s|y=-1} |y_r - y_s| / d_k'(x_r,x_s). \quad \text{(eq. 73)}$$

Por outro lado, as hipóteses feitas pelos autores do método CZCL não seriam violadas se T_k fosse calculado pela seguinte formulação:

$$T_k = \Sigma_{r|y=1}\Sigma_{s|y=-1} [d_k'(x_r,x_s)]^2 = \Sigma_{r|y=1}\Sigma_{s|y=-1} [d^2(x_r,x_s) - d_k^2(x_r,x_s)]. \quad \text{(eq. 74)}$$

Com esta modificação, só estaríamos invertendo a tendência de T_k e tornando-a mais sensível à distância entre os dados por elevá-la ao quadrado, uma vez que a parcela $|y_r - y_s|$ é sempre igual. Por outro lado, como a parcela $d^2(x_r,x_s)$ está presente no cálculo da sensibilidade de todas as variáveis, a sua omissão não afetaria o ordenamento relativo das variáveis e tornaria o critério ainda mais sensível à parcela realmente determinante: $d_k^2(x_r,x_s)$. Assim, chegamos à seguinte formulação:

$$T_k^M = \Sigma_{r|y=1}\Sigma_{s|y=-1} d_k^2(x_r,x_s). \quad \text{(eq. 75)}$$

Agora, quanto menor o valor de T_k^M mais relevante é variável k para a classificação de um determinado dado. Ao final, as variáveis serão ordenadas em valor crescente de T_k^F, dado por

$$T_k^F = T_k^M / \max_{k\epsilon\{1,\ldots,n\}} (T_k^F) \qquad \text{(eq. 76)}$$

O critério adaptado preserva as hipóteses do método original e traz consigo as seguintes vantagens adicionais:

i. maior sensibilidade à quantidade principal para a determinação do grau de importância de cada variável para a classificação de uma observação amostral; e
ii. maior facilidade e simplicidade para implementação computacional.

6.2.3 CRITÉRIO DUAL

Nos dois critérios anteriores, para o cálculo das variáveis de ordenamento T_k e T_k^F só foram consideradas, na realidade, as distâncias entre as observações com classificações distintas. Por outro lado, é também correto pensar que o oposto das hipóteses básicas é o que se objetiva, ou seja, quanto maior a distância entre as variáveis de entrada, uma melhor classificação elas propiciarão, pois a classificação no caso em estudo é bipolar. Desta forma, devemos selecionar as variáveis que apresentem os menores T_k e T_k^F, dando origem, respectivamente, aos critérios CZCL Dual e CZCL Adaptado Dual.

7 RESULTADOS COMPUTACIONAIS E DISCUSSÃO

A implementação dos critérios de seleção de variáveis e do ν-SVM foi feita no *software* **R**, empregando o pacote **kernlab** e utilizando o banco de dados em **Excel**. Os testes foram realizados em um processador AMD Athlon 2.1 GHz, 512 MB de memória RAM e sistema operacional **Windows**. Para evitar problemas de escala, os dados foram todos normalizados.

Os resultados dos métodos CZCL, CZCL Adaptado, CZCL Dual, CZCL Adaptado Dual para seleção de variáveis estão sintetizados, respectivamente, nas Tabelas 7.1 e 7.2. Como os métodos duais apenas invertem o ordenamento dos seus respectivos métodos primais, a ordem dual é apresentada entre parênteses nestas tabelas. Os métodos foram empregados numa amostra de 231 indivíduos, sendo que destes 16 evoluíram para óbito. Todos os pacientes deste conjunto amostral possuíam as 25 variáveis de entrada observadas.

Tabela 7.1: Critérios CZCL e CZCL Dual (entre parêntesis)

Ordem	**Variável**	$\mathbf{T_k/max(T_k)}$
1 (25)	QRP	0.98325
2 (24)	Alelo E3	0.98424
3 (23)	1ºAM	0.98517
4 (22)	Alelo E2	0.98699
5 (21)	IMP	0.98721
6 (20)	HAS	0.98769
7 (19)	Alelo E4	0.98887
8 (18)	IMC	0.98900
9 (17)	Alelo I	0.98921
10 (16)	Alelo T	0.98964
11 (15)	Alelo D	0.98982
12 (14)	Alelo M	0.99037
13 (13)	Colesterol	0.99105

Ordem	**Variável**	$\mathbf{T_k/max(T_k)}$
14 (12)	História DAC	0.99151
15 (11)	Triglicerídeos	0.99159
16 (10)	AF	0.99234
17 (9)	Tipo SCA	0.99241
18 (8)	Sexo	0.99292
19 (7)	Killip	0.99380
20 (6)	Escolaridade	0.99431
21 (5)	Tabagismo	0.99647
22 (4)	Col-HDL	0.99531
23 (3)	Idade	0.99853
24 (2)	FC	0.99863
25 (1)	Creatinina	1.00000

Tabela 7.2: Critérios CZCL Adaptado e CZCL Adaptado Dual (entre parênteses)

Ordem	Variável	T_k
1 (25)	QRP	0.43671
2 (24)	HAS	0.47359
3 (23)	1ºAM	0.51345
4 (22)	Alelo I	0.52098
5 (21)	IMP	0.52485
6 (20)	AF	0.54477
7 (19)	IMC	0.55521
8 (18)	História DAC	0.56697
9 (17)	Alelo M	0.57137
10 (16)	Alelo T	0.57503
11 (15)	Alelo E2	0.58567
12 (14)	Colesterol	0.60205
13 (13)	Triglicerídeos	0.61366

Ordem	Variável	T_k
14 (12)	Alelo E4	0.61412
15 (11)	Tipo SCA	0.61768
16 (10)	Escolaridade	0.62307
17 (9)	Sexo	0.64378
18 (8)	Tabagismo	0.67185
19 (7)	Col-HDL	0.67700
20 (6)	Alelo D	0.692605
21 (5)	Alelo E3	0.76387
22 (4)	Idade	0.76690
23 (3)	Killip	0.84072
24 (2)	FC	0.92736
25 (1)	Creatinina	1.00000

A primeira constatação que podemos fazer é que o critério CZCL Adaptado propiciou uma classificação das variáveis de entrada mais bem definida do que o critério original, pois a distância entre a primeira e a última variável no critério CZCL Adaptado é de 0.56 enquanto que no critério original é de apenas 0.02. Isto já era esperado, conforme exposto na seção 6.2.2, uma vez que o critério CZCL Adaptado leva em consideração para o ordenamento relativo das variáveis apenas a quantidade realmente determinante.

Na tabela 7.3, construída a partir das tabelas acima, podemos observar as primeiras seis variáveis selecionadas a partir dos quatro métodos de seleção de variáveis utilizados no presente trabalho. Constata-se que os critérios CZCL e CZCL Adaptado coincidiram em quatro variáveis (QRP, 1ºAM, IMP e HAS). A maior diferença ocorreu na comparação entre os critérios duais para as seis variáveis selecionadas, onde só houve a sobreposição de três (creatinina, FC e idade).

Com relação às variáveis genéticas, elas foram selecionadas: no critério CZCL, com os alelos E2 e E3 do polimorfismo do gene da APO E ocupando, respectivamente, a terceira e a segunda posições; no critério CZCL Adaptado, em que o alelo I do polimorfismo do gene da ECA ficou em quarta posição; e no critério CZCL Adaptado

Dual, segundo o qual o alelo E3 ficou na quinta posição e o alelo D do polimorfismo do gene da ECA ocupou o sexto lugar. No método CZCL Dual, as variáveis genéticas não foram selecionadas entre as seis variáveis mais importantes.

Tabela 7.3: Seis melhores variáveis em cada critério

Ordem	CZCL	CZCL Adaptado	CZCL Dual	CZCL Adaptado Dual	MIFS-U
1	QRP	QRP	Creatinina	Creatinina	Idade
2	Alelo E3	HAS	FC	FC	QRP
3	1°AM	1°AM	Idade	Killip	Creatinina
4	Alelo E2	Alelo I	HDL	Idade	IMC
5	IMP	IMP	Tabagismo	Alelo E3	Genótipo DD
6	HAS	AF	Escolaridade	Alelo D	Genótipo E4E4

Na tabela 7.3, são mostradas também as seis variáveis mais relevantes de acordo com o critério *Mutual Information Feature Selector Under Uniform Information Distribution* (MIFS-U), empregado por REIS (2007) sobre o mesmo banco de dados. Em linhas gerais, o MIFS-U é um algoritmo guloso que ordena as variáveis segundo o grau de contribuição para o índice de informação mútua conjunta (KWAK, CHOI, 2002), sendo muito eficiente e de implementação relativamente simples. Em relação ao estudo de REIS (2007), a principal diferença foi na manipulação das variáveis genéticas, que foram estudadas de acordo com os seus genótipos e não com os alelos. Em relação ao método do MIFS-U, o critério CZCL Adaptado Dual foi o que apresentou resultados mais próximos. Isto porque além das variáveis creatinina e idade, também selecionadas pelo CZCL Dual, o CZCL Adaptado Dual também destacou o alelo D que pode ser considerado coincidente com o genótipo DD selecionado pelo método MIFS-U. Contudo, o que mais chama a atenção foi a variável QRP ter ficado em primeiro lugar pelo critério CZCL Adaptado, exatamente o correspondente do CZCL Adaptado Dual, e também pelo critério CZCL, homólogo do CZCL Adaptado. Ou seja, os critérios que apresentaram maior coincidência com o MIFS-U (CZCL Dual e CZCL Adaptado Dual)

ordenaram a variável QRP, destacada pelo MIFS-U como a segunda variável mais importante, na última posição.

Para o treinamento do classificador ν-SVM, foram realizados experimentos computacionais para as três, as quatro, as cinco e as seis primeiras variáveis selecionadas por cada um dos critérios apresentados na subseção anterior. Em todas estas máquinas de aprendizado, foi utilizado o núcleo tangente hiperbólica de forma a se aproximar às redes neurais, já que um dos objetivos desta dissertação é comparar os resultados aqui alcançados com os apresentados em REIS (2007), onde foi empregada a RNA *feedforward* sob o mesmo banco de dados. Além disso, foram testados, numa etapa de avaliação inicial, os desempenhos dos núcleos polinomiais, FBR de Laplace e FBR de Gauss. Todos estes tiveram um aproveitamento bastante inferior se comparado ao núcleo tangente hiperbólica. A calibração do parâmetro *v* da função objetivo e do hiperparâmetro *escala* da função núcleo foi realizado sobre o conjunto X = {(ν,escala) | ν ϵ {0.13;0.14;...;0.25} e escala ϵ {0.1;0.2;...;1.5}. O hiperparâmetro *coef* foi ajustado para zero ao longo de todo o processo.

Para se avaliar os resultados em pesquisa médica costuma-se utilizar os seguintes conceitos (BALDI et al., 2000):

i. acurácia (a): probabilidade de predizer corretamente as observações;
ii. sensibilidade (s): probabilidade de predizer corretamente as observações positivas; e
iii. especifidade (e): probabilidade de predizer corretamente as observações negativas.

Para se estimar estas medidas, foi utilizado o recurso de validação cruzada *leave-one-out*, onde se treina o classificador numa amostra de tamanho "l-1" e faz ele predizer o dado que ficou de fora. Percebe-se que nesta metodologia, devemos realizar "l" treinamentos para cada máquina de aprendizado. Ao final, as estimativas das quantidades anteriores são dadas por:

$$a = acerto\_total / l, \quad \text{(eq. 77)}$$

$$s = acerto\_positivo / num\_positivo, \text{ e} \quad \text{(eq. 78)}$$

$$e = acerto\_negativo / num\_negativo, \quad \text{(eq. 79)}$$

onde:

acerto_total: número total de acerto que o ν-SVM obteve em um treinamento;

acerto_positivo: número de dados positivos classificados corretamente;

num_positivo: número de dados positivos;

acerto_negativo: número de dados negativos classificados corretamente; e

num_negativo: número de dados negativos.

Para cada conjunto de variáveis, foi empregado um conjunto de treinamento maximal de tal forma que os indivíduos possuíssem todas as informações em relação as mesmas. As Tabelas 7.4, 7.5, 7.6 e 7.7 apresentam os resultados encontrados. Nelas estão indicados apenas os melhores ajustes de parâmetro e hiperparâmetro alcançado para cada conjunto de variáveis.

Tabela 7.4: ν-SVM segundo as variáveis selecionadas pelo critério CZCL

Variáveis	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
QRP, Alelo E3, 1ºAM	0.15	0.6	63.1	62.2	74.1
QRP, Alelo E3, 1ºAM, Alelo E2	0.18	1	55.7	55.8	81.5
QRP, Alelo E3, 1ºAM, Alelo E2, IMP	0.16	0.6	85.6	87.5	60.0
QRP, Alelo E3, 1ºAM, Alelo E2, IMP, HAS	0.18	1.4	53.2	52.1	68.0

Tabela 7.5: ν-SVM segundo as variáveis selecionadas pelo critério CZCL Adaptado

Variáveis	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
QRP, HAS, 1ºAM	0.15	0.9	93.5	94.9	75.9
QRP, HAS, 1ºAM, Alelo I	0.14	1.2	95	96.1	81.5
QRP, HAS, 1ºAM, Alelo I, IMP	0.19	0.2	54.8	54.3	80
QRP, HAS, 1ºAM, Alelo I, IMP, AF	0.24	0.6	53.5	51.6	80

Tabela 7.6: ν-SVM segundo as variáveis selecionadas pelo critério CZCL Dual

Variáveis	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
Creatinina, FC, Idade	0.15	0.4	86.6	88.3	68.8
Creatinina, FC, Idade, Col-HDL	0.17	1.5	43.8	44.8	33.3
Creatinina, FC, Idade, Col-HDL, Tabagismo	0.19	0.5	59.9	57.9	81.5
Creatinina, FC, Idade, Col-HDL, Tabagismo, Escolaridade	0.14	1.5	91.5	93.9	60.9

Tabela 7.7: ν-SVM segundo as variáveis selecionadas pelo critério CZCL Adaptado Dual

Variáveis	ν	escala	a (%)	e (%)	s (%)
Creatinina, FC, Killip	0.16	0.4	56.9	56	66.7
Creatinina, FC, Killip, Idade	0.17	0.4	92.2	94.3	70.0
Creatinina, FC, Killip, Idade, Alelo E3	0.17	0.2	92.0	93.2	79.3
Creatinina, FC, Killip, Idade, Alelo E3, Alelo D	0.17	0.3	85.8	86.4	79.3

Como se pode constatar, o critério CZCL Adaptado e o seu correspondente dual selecionaram variáveis que permitiram o treinamento de classificadores com rendimentos muito próximos e superiores ao método original proposto por CHEN et al. (2008). Isto porque ao ordenar as variáveis diretamente pela quantidade principal, os critérios adaptados se mostram mais capazes de capturar as habilidades classificatórias imanentes a cada variável de entrada. Os resultados similares apresentados pelos classificadores primais e duais explicitam que as duas lógicas subjacentes à construção dos dois métodos são válidas e que talvez a combinação das duas traria melhores resultados. Nesse sentido, teríamos nove variáveis selecionadas: creatinina, FC, classe Killip, idade, alelo E3, QRP, HAS e 1ºAM. Algumas observações são importantes com respeito a este conjunto de variáveis e o desempenho dos classificadores descritos nas tabelas anteriores:

i. o alelo I propiciou uma melhora percentual significativa, de 75,9% para 81,5%, apenas dos óbitos (Tabela 7.5). A variável I, ao ser inclusa na amostra de treinamento, elimina dois indivíduos que evoluíram para óbito, devido à falta de dados desta variável. Provavelmente, pelo menos um destes indivíduos era um *outlier*, que o ν-SVM não conseguia capturar. Dada a pequena quantidade de pacientes que evoluíram para óbito no espaço amostral utilizado, consideramos que o avanço obtido com a inclusão do alelo I no conjunto das variáveis classificatórias não foi devido às qualidades tácitas deste variável. Neste caso, o conjunto realmente determinante para eficiência do classificador é formado pelas três primeiras variáveis: QRP, HAS e 1ºAM;
ii. o alelo E3 proporcionou maior sensibilidade ao classificador, aumento de 9%, embora tenha contribuído para uma redução de cerca de 1% no valor da especificidade e de 0.2% no valor da acurácia (Tabela 7.7). A inclusão deste alelo como variável de entrada do conjunto de treinamento provoca a retirada

dos mesmos dois indivíduos que morreram e que foram retirados pelo alelo I do espaço amostral anterior. Assim, pelas mesmas razões expostas acima, consideramos que as variáveis realmente significativas para o desempenho do classificador são creatinina, FC, classe Killip e idade;

iii. a idade é uma variável que foi coletada em todos os indivíduos e, por conseguinte, sua inclusão em qualquer classificador não acarreta em perda de observações de treinamento. Além disso, ela é uma variável que tem uma baixa probabilidade de apresentar ruído, haja vista que o atual nível de desenvolvimento sócio-econômico alcançado nos grandes centros urbanos impede que as pessoas não tenham um perfeito conhecimento da sua idade;

iv. a inclusão da idade no conjunto de variáveis de treinamento (Tabela 7.7) permitiu uma substancial melhora do desempenho do classificador: acurácia (35.3%), especificidade (38.3%) e sensibilidade (3.3%); e

v. o classificador formado pelas variáveis creatinina, FC e classe Killip (Tabela 7.7) tem um desempenho muito inferior ao do classificador formado pelas variáveis creatinina, FC e idade (Tabela 7.6). Esta diferença é equivalente a 29.7% na acurácia, 32.3% na especificidade e 2.1% na sensibilidade. Isto nos indica que o poder de distinção da variável idade é aparentemente maior que da variável classe Killip.

A partir das observações acima relatadas, concluímos que as variáveis mais influentes na predição do desfecho seriam: creatinina, FC, Idade, QRP, HAS e 1ºAM. As três primeiras selecionadas pelo critério CZCL Adaptado Dual e as três últimas pelo critério CZCL Adaptado. Da mesma forma que anteriormente, efetuamos experimentos computacionais para o treinamento de classificadores com alguns subconjuntos destas seis variáveis, sendo a calibração do parâmetro v e do hiperparâmetro *escala* realizados sobre o mesmo conjunto X. Os resultados estão sintetizados na tabela 7.8. Em todos os subconjuntos selecionados decidimos manter as variáveis QRP (primeira selecionada pelo critério CZCL Adaptado), creatinina (primeira selecionada pelo critério CZCL Adaptado Dual), e idade. A decisão por esta última foi devido ao baixo nível de erro inerente à sua coleta e à melhora no desempenho do classificador, e por ela ter sido amostrada em todos os indivíduos.

Tabela 7.8: ν-SVM segundo as variáveis selecionadas pela combinação de critérios

Variáveis	ν	escala	a (%)	e (%)	s (%)
Creatinina, QRP, Idade	0.16	0.3	97.5	98.3	87.5
Creatinina, QRP, Idade, HAS	0.16	0.6	97.7	98.6	87.5
Creatinina, QRP, Idade, FC	0.16	0.1	96.1	97.6	79.3
Creatinina, QRP, Idade, 1ºAM	0.17	0.5	70.4	70	75
Creatinina, QRP, Idade, HAS, FC	0.16	0.3	96.4	97.9	79.3
Creatinina, QRP, Idade, HAS, 1ºAM	0.15	0.6	79.8	81.1	64.3

Em REIS (2007), a RNA associada ao método MIFS-U obteve um classificador ótimo com três variáveis (Idade, QRP, creatinina) cujos valores de acurácia, especificidade e sensibilidade foram, respectivamente, 70.7%, 69.8% e 78.3%. Como se pode constatar na tabela 7.8, todos os conjuntos de variáveis resultantes da combinação dos critérios CZCL Adaptado e CZCL Adaptado Dual, que não incluísse a variável 1ºAM, proporcionaram classificadores cujos desempenhos foram superiores ao da RNA. Além disso, os conjuntos de variáveis selecionadas pelos critérios CZCL Adaptado e CZCL Adaptado Dual aplicados isoladamente (Tabelas 7.5 e 7.7) também permitiram a construção de máquinas de aprendizado cujos desempenhos foram maiores que o da RNA. Por outro lado, o conjunto de variáveis neste estudo que apresentou melhor desempenho foi o mesmo que o encontrado em REIS (2007).

Por fim, verificamos a robustez dos quatro classificadores construídos, os de melhor desempenho segundo os critérios de seleção de variáveis CZCL Adaptado e CZCL Adaptado Dual e os dois mais eficientes obtidos a partir da combinação destes critérios conforme descrito anteriormente. Para isto, foram realizados dez treinamentos com cada conjunto de variáveis selecionadas, associado à validação cruzada *leave-one-out*. O dado inicial de cada treinamento foi alterado aleatoriamente. Os resultados estão apresentados nas tabelas abaixo.

Tabela 7.9: ν-SVM segundo as variáveis QRP, HAS, 1ºAM e alelo I (Critério CZCL Adaptado)

Treinamento	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
1	0.14	1.2	93.7	94.9	77.8
2	0.14	1.2	93.7	94.9	77.8
3	0.14	1.2	94	94.9	81.5
4	0.14	1.2	93.7	94.9	77.8
5	0.14	1.2	94	94.9	81.5
6	0.14	1.2	94	94.9	81.5
7	0.14	1.2	93.7	94.9	77.8
8	0.14	1.2	93.7	94.9	77.8
9	0.14	1.2	94	94.9	81.5
10	0.14	1.2	93	93.8	81.5
Média	0.14	1.2	93.75	94.79	79.65

Tabela 7.10: ν-SVM segundo as variáveis creatinina, FC, classe Killip, idade e alelo E3 (Critério CZCL Adaptado Dual)

Treinamento	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
1	0.17	0.2	93.8	95.5	75.9
2	0.17	0.2	93.2	94.2	82.8
3	0.17	0.2	93.2	94.2	82.8
4	0.17	0.2	93.2	94.2	82.8
5	0.17	0.2	92.6	93.8	79.3
6	0.17	0.2	94.1	95.5	79.3
7	0.17	0.2	92.6	94.2	79.3
8	0.17	0.2	92.9	94.2	75.9
9	0.17	0.2	93.5	94.8	79.3
10	0.17	0.2	94.7	95.8	79.3
Média	0.17	0.2	93.38	94.64	82.8

Tabela 7.11: ν-SVM segundo as variáveis creatinina, idade e QRP (Critério CZCL Adaptado e Critério CZCL Adapatado Dual combinados)

Treinamento	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
1	0.16	0.3	97.5	98.3	87.5
2	0.16	0.3	97.7	98.3	90.6
3	0.16	0.3	97.7	98.3	90.6
4	0.16	0.3	97.7	98.3	90.6
5	0.16	0.3	97.5	98.3	87.5
6	0.16	0.3	97.5	98.3	87.5
7	0.16	0.3	97.5	98.3	87.5
8	0.16	0.3	97.5	98.3	87.5
9	0.16	0.3	97.5	98.3	87.5
10	0.16	0.3	97.5	98.3	87.5
Média	0.16	0.3	97.56	98.30	88.43

Tabela 7.12: ν-SVM segundo as variáveis creatinina, idade, QRP e HAS (Critério CZCL Adaptado e Critério CZCL Adaptado Dual combinados)

Treinamento	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
1	0.16	0.6	97.5	98.3	87.5
2	0.16	0.6	97.2	98.3	84.4
3	0.16	0.6	97.2	98.3	84.4
4	0.16	0.6	97.2	98.3	84.4
5	0.16	0.6	97.2	98.3	84.4
6	0.16	0.6	97.5	98.3	87.5
7	0.16	0.6	97.5	98.3	87.5
8	0.16	0.6	97.5	98.3	87.5
9	0.16	0.6	97.5	98.3	87.5
10	0.16	0.6	97.2	98.3	84.4
Média	0.16	0.6	97.35	98.30	85.95

Os números nas tabelas acima nos indicam que os classificadores obtidos são bem robustos e estáveis. Além disso, poderíamos pensar que a variável HAS não acrescenta

informação relevante ao classificador de melhor resultado e que desta forma poderia ser descartada para a classificação do risco de morte do paciente internado com SCA. Contudo, antes de afirmarmos isto, decidimos testar um classificador de três variáveis homólogo ao classificador constituído pelas variáveis creatinina, idade e QRP, isto é, decidimos realizar o teste de calibração e robustez para um classificador formado pelas variáveis creatinina, idade e HAS. Neste caso, trocamos as variáveis selecionadas pelo mesmo critério e que são bi-classe em oposição as outras duas, creatinina e idade, que são contínuas. O ajuste ótimo foi obtido para um parâmetro $\nu = 0.17$ e hiperparâmetro *escala* = 0.3. Os valores de desempenho deste classificador estão apresentados na Tabela 7.13.

Tabela 7.13: ν-SVM segundo as variáveis creatinina, idade, HAS (Critério CZCL Adaptado e Critério CZCL Adapatado Dual combinados)

Treinamento	**ν**	**escala**	**a** (%)	**e** (%)	**s** (%)
1	0.17	0.3	96.8	98.4	80.0
2	0.17	0.3	96.8	98.4	80.0
3	0.17	0.3	96.8	98.4	80.0
4	0.17	0.3	96.8	98.4	80.0
5	0.17	0.3	96.8	98.4	80.0
6	0.17	0.3	96.8	98.4	80.0
7	0.17	0.3	96.5	98.4	77.1
8	0.17	0.3	96.5	98.4	77.1
9	0.17	0.3	96.8	98.4	80.0
10	0.17	0.3	96.5	98.4	77.1
Média	0.17	0.3	96.71	98.4	79.13

Antes de analisarmos os números da tabela anterior, é preciso mencionar que o conjunto de treinamento para as variáveis creatinina, idade e HAS (grupo 1) possui três indivíduos a mais que evoluíram para óbito em relação ao conjunto de treinamento para as variáveis creatinina, idade e QRP. Assim, dado o reduzido número de óbitos na amostra, ao não ser exposto a estes três pacientes, provavelmente de difícil classificação, o ν-SVM treinado com as variáveis creatinina, idade e QRP (grupo 2) obteve uma sensibilidade maior. Assim, consideramos que a variável HAS se sobrepõe

à variável QRP e, desta forma, ambas deveriam ser utilizadas para predição do risco de óbito, mas não num único classificador. Um exemplo de aplicação prática seria o emprego integrado de dois classificadores, um constituído pelas variáveis do grupo 1 e outro pelas variáveis do grupo 2.

8 CONCLUSÃO

Ao longo deste trabalho procuramos construir uma ferramenta computacional capaz de predizer o risco de morte dos pacientes internados com SCA. Para isto, elaboramos um modelo matemático a partir do uso da técnica do SVM, que combina otimização quadrática convexa, função núcleo e teoria de generalização. O resultado foi a construção de um classificador que superou o desempenho da RNA *feedforward* construída sobre o mesmo banco de dados e apresentada em REIS (2007), a partir do uso de uma função núcleo tangente hiperbólica.

Para seleção das variáveis, foram testados quatro métodos: o método proposto por CHEN et al. (2008) e três modificações do mesmo. As soluções encontradas indicaram que o melhor resultado era obtido a partir do uso combinado das duas alterações propostas. Neste sentido, foram selecionados dois conjuntos de variáveis:

i) creatinina, idade e hipertensão arterial sistêmica; e

ii) creatinina, idade e qualquer revascularização prévia.

Discutimos que, apesar do grupo 2 de variáveis ter tido uma sensibilidade mais alta, isto foi devido ao pequeno número de óbitos presente no banco de dados empregado e à presença de três óbitos a menos no grupo 2 em relação ao grupo 1. Esta redução de óbitos provavelmente permitiu a eliminação de *outliers* de difícil classificação pelas máquinas de aprendizado. Assim, é recomendável o uso simultâneo dos dois classificadores na prática médica. Desta forma, poderíamos classificar o risco de morte de pacientes internados com SCA em três classes: risco alto – os dois classificadores apontam risco alto; risco moderado – os dois classificadores apontam riscos divergentes, isto é, um indica risco baixo e o outro risco alto; e risco baixo – os dois classificadores apontam risco baixo.

Em comparação ao método do MIFS-U, a metodologia de seleção de variáveis aqui proposta chegou as mesmas variáveis selecionadas (grupo 2), porém também permitiu o apontamento da variável HAS (grupo 1) como uma variável importante para embasar futuras decisões médicas no caso de pacientes com SCA.

Com relação às variáveis genéticas, o atual estudo não indica que as mesmas sejam relevantes e/ou contribuam para predizer o risco de morte de pacientes com SCA na população estudada.

Concluímos, portanto, que a presente dissertação atingiu os objetivos pré-estabelecidos. Esperamos que os resultados aqui alcançados sirvam como auxílio na difícil e árdua missão que os médicos têm que enfrentar diariamente no tratamento dos pacientes internados com SCA.

REFERÊNCIAS BIBLIOGRÁFICAS

BALDI, P., BRUNAK, S., CHAUVIN, Y. et al., 2000, "Assessing the accuracy of prediction algorithms for classification: an overview", *Bioinformatics Review*, v. 16, n. 5, pp. 412-424.

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

BOSER, B. E., GUYON, I. M., VAPNIK, V. N., 1992, "A training algorithm for optimal margin classifiers". In: *Proceedings of the 5th Annual ACM Workshop on Computacioal Learning Theory*, pp. 144-152, Pittsburgh, Jul.

BOYD, S., VANDERNBERGHE, L., 2003, *Convex Optimization*. 6 ed. Cambridge, Cambridge University Press.

BROWN, T. A., 1999, "O Genoma Humano". In: Brown, T. A. (ed), *Genética. Um enfoque molecular*, 3 ed., capítulo 16, Rio de Janeiro, Editora Guanabara Koogan.

BURGES, C. J. C., 1998, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, v. 2, n. 2, pp. 145-163.

CHEN, P.H., L., C. J., SCHÖLKOPF, B., 2005, "A Tutorial on ν-Support Vector Machines", *Applied Stochastic Models in Business and Industry*, v. 21, n. 2., pp. 111-136.

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.

CORTES, C., VAPNIK, V., 1995, "Support Vector networks", Machine Learning, v. 20, n. 3, pp. 273-297.

CRISTIANINI, N., SHAWE-TAYLOR, J., 2006, *An Introduction to Support Vector Machines and other kernel-based learning methods*. 10 ed. Cambridge, Cambridge University Press.

GUNN, S. R., 1998, *Support Vector Machines for Classification and Regression*, Technical Report, University of Southampton, Southampton, UK.

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

KARATZOGLOU, A., MEYER, D, HORNIK, K., 2006, "Support Vector Machines in R", *Journal of Statistical Software*, vol.15, n. 9, pp. 1-28. Disponível em: <http://www.jstatsoft.org/v15/i09/paper>. Acesso em: 17 abr. 2009.

LIMA, E. L., 2000, *Curso de Análise, Volume 2*. 6 ed. Rio de Janeiro, IMPA.

LUENBERGER, D. G., YE, Y., 2008, *Linear and Nonlinear Programming*. 3ª ed. New York, Springer.

MONÇORES, M. W., PEREIRA, S. B., GOUVEA, L. S. F. et al., 2008, "Medicina indicidualizada aplicada à cardiologia", *Revista SOCERJ* vol. 21, n. 3 (Mai-Jun), pp. 184-193.

NOCEDAL, J., WRIGHT, S. J., 2006, *Numerical Optimization*. 2 ed. New York, Springer.

OSUNA, E., FREUND, R., GIROSI, F., 1997, "An improved training algorithm for support vector machines". In: *Proceedings of the IEEE Workshop on Neural Networks and Signal Processing*, pp. 267-285, New York, Sep.

PAN, F., 2003, *Efficient Proximal Support Vector Machine for Spatial Data*. M. Sc. Dissertation, North Dakota State University of Agriculture and Applied Science at Fargo, North Dakota, USA.

POPPER, K., 2007, *A Lógica da Pesquisa Científica*. 13 ed. São Paulo, Cultrix.

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.

SCHÖLKOPF, B., SMOLA, A. J., 2002, *Learning with Kernels. Support Vector Machines, Regulararization, Optimization, and Beyond*. 1 ed. Cambridge, MIT Press.

VAPNIK, V., 1995, *The Nature of Statiscal Learning Theory*, 1 ed. New York, Springer.

WASSERMAN, L., 2004, *All of statistics: a concise course in statistical inference*. 2 ed. New York, Springer.