APLICAÇÃO DO MODELO SUPPORT VECTOR MACHINE PARA A PREDIÇÃO DO RISCO DE MORTE DE PACIENTES INTERNADOS COM SÍNDROME CORONARIANA AGUDA

Rodrigo Abrunhosa Collazo
CASNAV - COPPE/UFRJ
Praça Barão de Ladário, s/nº, Centro, RJ - Ilha do Fundão, Centro de Tecnologia, bl. F, sala 105
e-mail: rodrigocollazo@uol.com.br

Amália Faria dos Reis
HUAP/UFF
Av. Marques do Paraná 103, Centro, Niterói
e-mail: amaliafr@gmail.com

Basílio de Bragança Pereira
HUCFF-COPPE/UFRJ
Ilha do Fundão, Centro de Tecnologia, bl. F, sala F103
e-mail: basilio@hucff.ufrj.br

Laura Bahiense
COPPE/UFRJ
Ilha do Fundão, Centro de Tecnologia, bl. F, sala F103
e-mail: laura@pep.ufrj.br

RESUMO

Este artigo apresenta um modelo S*upport Vector Machine* para classificar, em alto ou baixo, o risco de morte de pacientes internados com síndrome coronariana aguda. Para a seleção das variáveis de entrada foram empregados, de forma conjugada, o filtro MIFS-U (*Mutual Information Feature Selector under Uniform distribution criteria*) e o próprio algoritmo de classificação. Os resultados computacionais indicaram as variáveis de entrada idade, qualquer revascularização prévia e creatinina como as mais relevantes. Os classificadores construídos apresentaram acurácia preditiva superior a 95%.

PALAVRAS CHAVE. Seleção de variáveis. Support Vector Machine. Síndrome Coronariana Aguda

ABSTRACT

This article presents a Support Vector Machine model to classify the risk of death of hospitalized patients with acute coronary syndrome at high or low. To select the input variables, it was employed the Mutual Information Feature Selector under Uniform distribution (MIFS-U) criteria associated with the classifier algorithm itself. The computational results indicated the input variables age, prior revascularization and creatinine as the most relevant. The classifiers presented predictive accuracy greater than 95%.

KEYWORDS. Variable Selection. Support Vector Machine. Acute Coronary Syndrome.

1. Introdução

As doenças do aparelho circulatório, que incluem as doenças isquêmicas do coração e as doenças cerebrovasculares, são as principais causas de morte no país. No ano de 2004, segundo dados disponibilizados para este estudo, as doenças isquêmicas do coração foram responsáveis por 86.791 óbitos, sendo que 46.326 (53,4%) ocorreram no Sudeste. Nessa região, os estados com maior mortalidade por doenças isquêmicas do coração foram São Paulo e Rio Janeiro com, respectivamente, 26.235 óbitos (56,6%) e 10.718 óbitos (23,1%). No Rio de Janeiro, os municípios com maior incidência de mortes por doenças isquêmicas do coração naquele ano foram: 1º) Rio de Janeiro (5.175), 2º) Nova Iguaçu (491), 3º) Niterói (469) e 4º) São Gonçalo (449). De acordo com dados da Fundação Municipal de Saúde, as doenças do aparelho circulatório foram a principal causa de internação no município de Niterói entre 1997 e 2006, assim como a principal causa de morte de 2003 a 2005.

Dentre as doenças isquêmicas do coração, as síndromes coronarianas agudas possuem papel de relevo, devido à persistência do seu alto nível de mortalidade, não obstante as conquistas terapêuticas dos últimos anos. As síndromes coronarianas agudas englobam o infarto agudo do miocárdio (IAM), com e sem supradesnível do ST, e a angina instável. Elas são caracterizadas pela oclusão total ou parcial da artéria coronariana a partir da ruptura de uma placa coronariana instável, levando à isquemia e/ou à necrose da área do miocárdio irrigada por aquela artéria.

Na literatura médica, presume-se que a síndrome coronariana aguda provém do amálgama de condicionamentos ambientais, clínicos, genéticos e sócio-culturais. Portanto, a identificação das variáveis mais relevantes para predição do risco de morte de pacientes internados com esta doença e a conseqüente elaboração de um modelo preditivo permitirão otimizar as ações de saúde pública do setor, ajudando a tornar mais eficazes as condutas terapêuticas nos indivíduos classificados como sendo de alto risco de morte.

Em razão da conjunção destas características multifatoriais da síndrome coronariana aguda, o uso de métodos de modelagem não-linear permite a construção de classificadores binários de maior capacidade preditiva. Neste particular, vários estudos relacionados a doenças coronarianas foram desenvolvidos empregando-se o modelo *Support Vector Machine* (SVM) (Millet-Roig et al, 2000, Conforti & Guido, 2005, Çomak et al, 2007, Jankowski et al, 2007 e Bathia et al, 2008). Contudo, apenas o trabalho de Conforti & Guido (2005) teve como foco o desenvolvimento do modelo SVM aplicado à síndrome coronariana aguda, mais especificamente ao diagnóstico antecipado do infarto agudo do miocárdio. Por outro lado, o presente estudo pretende identificar o risco de morte dado que o paciente já apresentou o desenvolvimento clínico dos sintomas desta doença.

Desta forma, o objetivo deste trabalho foi o de construir um classificador SVM para predizer o risco de morte (alto ou baixo) de pacientes internados com síndrome coronariana aguda a partir da ordenação das variáveis de entrada fornecidas pelo critério MIFS-U (*Mutual Information Feature Selector under Uniform information distribution*), aplicado a um banco de dados coletados na cidade de Niterói entre os anos de 2004 e 2005.

2. Seleção de Variáveis e o Modelo SVM

Nesta seção, discutiremos brevemente o processo de seleção de variáveis e o modelo SVM empregado neste estudo.

2.1. Seleção de Variáveis

Na construção de classificadores, é fulcral, conforme discutido em Guyon & Elisseeff (2003) e Sallapa et al (2007), se proceder à seleção das variáveis de entrada a fim de: i) evitar o sobre-ajustamento; ii) obter modelos de menor custo computacional; e iii) facilitar o

entendimento do processo tácito que é responsável pela geração dos dados. Ademais, o banco de dados utilizado neste trabalho tem 411 indivíduos observados sob o enfoque de 28 variáveis, sendo que apenas 37 destes evoluíram para óbito. Além disso, há variáveis de entrada que não foram coletadas em todos os indivíduos. Ou seja, um indivíduo X pode possuir informações referentes a apenas n variáveis, onde n é menor que 28. Assim, o emprego de todas as variáveis de entrada obriga à eliminação dos indivíduos que não possuem os dados completos, reduzindo a amostra para 264 indivíduos, sendo que apenas 17 evoluíram para óbito. Desta forma, neste estudo é fundamental selecionar as variáveis de entrada mais importantes, de forma a se reduzir o espaço de entrada e, por conseguinte, se aumentar o número de dados completos para o treinamento.

Segundo Blum & Langley (1997) e Guyon & Elisseeff (2003), os métodos de seleção de variáveis podem ser agrupados em três classes: filtro, *embedded* e *wrapper*. Os filtros são um pré-processamento, no qual um subconjunto das variáveis de entrada é selecionado a partir somente das informações que os dados trazem, independentemente do algoritmo de classificação a ser utilizado. As vantagens dos filtros são a facilidade e a rapidez de implementação. A sua principal desvantagem é ignorar o viés do classificador. Os *embedded* incorporam os métodos de seleção ao próprio treinamento do classificador. Por outro lado, os *wrappers* fazem uso dos algoritmos de classificação como se estes fossem pacotes fechados, para avaliar a capacidade preditiva de subconjuntos de variáveis de entrada. Estes subconjuntos normalmente são constituídos através da utilização de algum procedimento heurístico. A principal vantagem dos últimos dois métodos é exatamente a interação com o algoritmo de classificação, embora esta seja também a fonte de suas deficiências: maior tempo de processamento e dependência do próprio algoritmo de classificação.

Neste trabalho, optou-se pelo emprego do *wrapper* associado a duas aplicações sucessivas do filtro MIFS-U (Kwak & Choi, 2002). Através de uma estratégia gulosa e de *forward selection*, o filtro MIFS-U ordena as variáveis de acordo com o grau de informação mútua existente entre elas e o desfecho. Para isto, as variáveis independentes, contínuas ou categóricas, precisam ser aproximadas pelos seus histogramas. O algoritmo pode ser sumarizado como segue. Passo inicial: seleção da variável que possui o maior grau de informação mútua com o desfecho; Passos intermediários (k): seleção da k-ésima variável que, em conjunto com as $k-1$ variáveis selecionadas nos passos anteriores, possui o maior grau de informação mútua com o desfecho; Critério de Parada: o algoritmo termina quando não houver mais variáveis a serem selecionadas.

Primeiramente, o filtro MIFS-U foi utilizado sobre uma amostra com todas as 28 variáveis observadas: 264 indivíduos, sendo que destes 17 faleceram e 247 sobvreviveram. Em seguida, as últimas 12 variáveis foram excluídas do estudo e o filtro MIFS-U foi aplicado novamente, agora sobre as 16 variáveis remanescentes em uma amostra com 351 indivíduos (23 óbitos e 328 não óbitos). Este aumento de dados na amostra foi devido à redução do número de variáveis e, por conseguinte, à diminuição do número de indivíduos que não possuíam as informações de todas as variáveis que estavam sendo empregadas. Este último ordenamento serviu de entrada para o *wrapper* seguindo a estratégia *sequential forward selection*: o algoritmo de classificação (RNA ou SVM) avaliou cada subconjunto aninhado de variáveis até que o erro de classificação do desfecho começasse a aumentar. Além disso, foi definido que a busca não seria interrompida até o valor mínimo de k igual a seis.

Para análise comparativa da performance dos classificadores foi utilizado o coeficiente de correlação de Pearson (CP) entre o desfecho real e o desfecho predito (Baldi et al, 2000). Os erros de classificação foram sintetizados através dos seguintes conceitos: i) acurácia (a): probabilidade de predizer corretamente o desfecho; ii) sensibilidade (s): probabilidade de predizer corretamente o desfecho de risco alto de morte; e iii) especifidade (e): probabilidade de

predizer corretamente uma alta probabilidade de sobrevivência. Para se estimar estas medidas, foi utilizado o recurso de validação cruzada *leave-one-out* (Wasserman, 2004). Neste caso, em uma amostra de tamanho n são realizados n treinamentos. Em cada treinamento, a amostra é constituída por $n-1$ indivíduos e a amostra de teste é formada pelo indivíduo que fica de fora e que é distinto em cada treinamento. As medidas finais são obtidas pela média aritméticas das medidas que foram registradas em todos os n treinamentos.

2.2. Modelo SVM

O modelo SVM (*Support Vector Machine*) consiste na seleção de um hiperplano que minimize o risco estrutural, a partir da resolução de um problema convexo quadrático. Esta técnica de aprendizado quando associada à função núcleo permite a construção de classificadores não-lineares, através do mapeamento dos dados iniciais em um espaço de dimensão superior ao original. Neste caso, o classificador linear obtido em um espaço de dimensão superior corresponderá a um classificador não-linear no espaço original (Wasserman, 2004). Estas características garantem ao SVM uma boa generalização, ou seja, uma boa capacidade de predizer corretamente dados não relacionados na amostra de treinamento.

Em Cristianini & Shawe-Taylor (2006) e Schölkopf & Smola (2002), a formulação matemática dos modelos SVM é detalhadamente abordada. O modelo empregado neste estudo foi o ν-SVM (Schölkopf et al, 2000, Chen et al., 2005, Schölkopf & Smola, 2002), que é uma variante do primeiro modelo SVM a ser desenvolvido, o C-SVM.

No treinamento do C–SVM, uma grande dificuldade é a calibração do parâmetro C. Para contornar este problema, foi proposta a formulação alternativa ν-SVM, que trabalha com um plano de margem funcional γ e substitui o parâmetro C pelo parâmetro ν. Prova-se que o parâmetro ν representa um limite superior para a fração dos erros de treinamento e um limite inferior para a fração dos vetores suportes. Estes significados matemáticos facilitam a calibração do parâmetro ν durante o treinamento do classificador. Além disso, 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 $C = 1/l \cdot \gamma^*$.

No modelo ν-SVM, o problema primal é colocado da seguinte forma:

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

$$\text{sujeito a } \begin{cases} y_i \cdot (\ll w, \Phi(x_i) \gg + b) \geq \gamma - \xi_i, \quad i = 1, \dots, l \\ \gamma \geq 0 \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

onde $\Phi(x_i)$ corresponde à projeção implícita do vetor de entrada x_i em um espaço de dimensão superior através da função núcleo $K(x_i, x_j)$; ξ_i, i=1,...,l, são variáveis de folga; e o hiperplano separador procurado é dado por $\ll w, \cdot \gg + b$. Isto corresponde a um problema quadrático convexo, cuja solução é mais fácil a partir da sua formulação dual dada por:

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

$$\text{sujeito a} \begin{cases} \sum_{i=1}^{l} \alpha_i \cdot y_i = 0 \\ \sum_{i=1}^{l} \alpha_i \geq \nu \\ 0 \leq \alpha_i \leq 1/l, \qquad i = 1, \dots, l \end{cases}$$

onde α_i, $i = 1, \dots, l$, são os multiplicadores lagrangeanos.

As funções núcleo usualmente encontradas em aplicações de modelos SVM são:

- núcleo linear: $K(x, y) = \ll x, y \gg$
- núcleo polinomial: $K(x, y) = (\text{escala} \cdot \ll x, y \gg + \text{coef}) \cdot \text{grau}$
- núcleo tangente hiperbólica: $K(x, y) = \tanh(\text{escala} \cdot \ll x, y \gg + \text{coef})$
- núcleo função base radial (FBR) de Laplace: $K(x, y) = e^{-\sigma \cdot \|x-y\|}$
- núcleo função base radial (FBR) de Gauss: $e^{-\sigma \cdot \|x-y\|^2}$

Os hiperparâmetros "escala", "coef", "grau" e σ são ajustados durante o treinamento do classificador por meio de técnicas estatísticas.

Em Karatzoglou et al. (2006) e Gunn (1998), há algumas sugestões de uso destas funções conforme o problema que se deseja 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 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. Neste estudo, foi utilizado o núcleo tangente hiperbólica, já que, em uma etapa de avaliação inicial, os desempenhos dos núcleos polinomiais, FBR de Laplace e FBR de Gauss foram inferiores.

Para a resolução do problema quadrático convexo há diversos algoritmos propostos na literatura. Optou-se pelo pacote **kernlab** do *software* **R**, que utiliza o algoritmo *Sequential Minimal Optimisation* para solução do problema dual. Este algoritmo encontra analiticamente a solução global, sem recorrer a métodos numéricos. Isto é possível levando a técnica de decomposição ao limite: a cada passo são otimizados apenas dois multiplicadores lagrangeanos. A cada iteração, escolhem-se heuristicamente dois multiplicadores lagrangeanos que são otimizados analiticamente, mantendo todos os demais constantes. O processo prossegue até que não haja mais melhorias a serem realizadas. Apesar de precisar de maior número de iterações para convergir, o SMO é bastante rápido, pois realiza várias operações de baixo custo computacional.

3. Dados

Os dados foram coletados a partir de um estudo de coorte prospectivo de 12 meses, a partir de julho de 2004, com pacientes internados com síndrome coronariana aguda no município de Niterói, no estado do Rio de Janeiro. Os pacientes considerados no levantamento estavam hospitalizados em três hospitais públicos e em dois hospitais privados. Eles tinham idade superior a 20 anos e não podiam apresentar sinais de doenças neoplásicas em fase terminal,

politraumatismos e demência. A Tabela 1 sumariza as 28 variáveis de entrada e a variável de desfecho presentes no banco de dados.

Tabela 1- Variáveis de entrada e desfecho

Categoria	Variável	Tipo
Variáveis antropométricas, sociais e relacionadas a hábitos de vida	Idade	Inteira
	Índice de Massa Corporal (IMC)	Real
	Sexo	Categórica (2 categorias)
	Escolaridade	Categórica (2 categorias)
	Tabagismo	Categórica (3 categorias)
Variáveis relacionadas à história cardiovascular prévia	Infarto do Miorcádio Prévio (IMP)	Categórica (2 categorias)
	Qualquer Revascularização Prévia (QRP)	Categórica (2 categorias)
Variáveis clínicas e laboratoriais na admissão hospitalar	Tipo de Síndrome Coronariana Aguda (TSCA)	Categórica (3 categorias)
	Freqüência Cardíaca	Inteira
	Classe Killip	Categórica (4 categorias)
	Creatinina	Real
Variáveis de diagnóstico	Hipertensão Arterial Sistêmica (HAS)	Categórica (2 categorias)
	Diabete Melito	Categórica (2 categorias)
	Colesterol Total elevado	Categórica (2 categorias)
	Triglicerídeos elevados	Categórica (2 categorias)
	Colesterol HDL baixo	Categórica (2 categorias)
Variáveis Genéticas	Genótipo DD do polimorfismo do gene da Enzima Conversora da Angiotensina I (ECA)	Categórica (2 categorias)
	Genótipo DI do polimorfismo do gene da enzima conversora da angiotensina I (ECA)	Categórica (2 categorias)
	Genótipo II do polimorfismo do gene da enzima conversora da angiotensina I (ECA)	Categórica (2 categorias)
	Genótipo MM do polimorfismo M235T do gene do angiotensinogênio (AGT)	Categórica (2 categorias)
	Genótipo MT do polimorfismo M235T do gene do angiotensinogênio (AGT)	Categórica (2 categorias)
	Genótipo TT do polimorfismo M235T do gene do angiotensinogênio (AGT)	Categórica (2 categorias)
	Genótipo E2E2 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
	Genótipo E2E3 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
	Genótipo E2E4 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
	Genótipo E3E3 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
	Genótipo E3E4 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
	Genótipo E4E4 do polimorfismo do gene da apolipoproteína E (APO E)	Categórica (2 categorias)
Variável de desfecho	Óbito	Categórica (2 categorias)

Neste estudo, as variáveis de diagnóstico foram consideradas como variáveis categóricas de duas classes (os pacientes apresentavam diagnóstico positivo ou negativo), embora elas possuíssem uma componente contínua na sua avaliação. Esta opção foi feita, pois qualquer uma destas variáveis poderia apresentar parâmetro normal se o paciente estivesse fazendo uso de medicamentos para controlar um possível diagnóstico positivo prévio à internação associado à respectiva variável. Ou seja, o paciente tacitamente apresentaria diagnóstico positivo, contudo esta situação não seria capturada pela simples medição de uma variável contínua representada pela pressão arterial, pelo colesterol, pelo triglicerídeo ou pelo colesterol–HDL. Desta forma, de acordo com procedimentos consolidados na prática médica, foram usados os seguintes critérios para determinar se o paciente apresentava ou não diagnóstico positivo:

- Hipertensão Arterial Sistêmica - 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;
- Diabete Melito – indivíduos com glicemia de jejum maior ou igual a 126 mg/dl ou uso prévio de hipoglicemiante oral ou insulina;
- Colesterol Total elevado - indivíduos com colesterol total acima de 200 mg/dl ou que faziam uso de hipolipemiante prévio à internação (estatina e/ou fibrato);
- Triglicerídeos elevados - os indivíduos com mais de 150 mg/dlb ou que faziam uso de hipolipemiante prévio à internação (estatina e/ou fibrato); e
- Colesterol HDL-baixo - os indivíduos com colesterol-HDL inferior a 40 mg/dl ou que faziam uso de hipolipemiante prévio à internação (estatina e/ou fibrato).

4. Resultados e Discussão

A implementação do filtro MIFS-U foi feita no *software* **MATLAB**, versão 7.0, utilizando o banco de dados em **Excel**. O modelo ν-SVM foi executado no *software* **R**, versão 2.7.1, empregando o pacote **kernlab** e utilizando também o banco de dados em **Excel**. Os testes foram realizados em um computador com processador AMD Athlon 2.1 GHz, com 512 MB de memória RAM e sistema operacional **Windows XP**. Para evitar problemas de escala, os dados foram todos normalizado**s**.

O filtro MIFS-U foi usado duas vezes consecutivas, conforme descrito na seção 2.1. Os resultados de cada fase são mostrados na Tabela 2. Em ambas as fases, observamos que as três primeiras posições permaneceram inalteradas. Além disso, sete variáveis (idade, qualquer revascularização prévia, creatinina, genótipos E2E4 e E4E4 da APO E, genótipos MM e TT do AGT) não tiveram qualquer alteração nos seus posicionamentos, duas variáveis (genótipo DD da ECA, genótipo E3E4 da APO E) variaram apenas uma posição e três variáveis (genótipo E2E2 da APO E, diabete melito, genótipo E3E3 da APO E) tiveram seu ordenamento modificado em somente três colocações. Por outro lado, as variáveis que mais sofreram alterações entre as duas fases foram índice de massa corporal (3 posições), genótipo E2E3 da APO E (4 posições), hipertensão arterial sistêmica (5 posições) e sexo (8 posições). Desta forma, os resultados se mostraram bastante estáveis, com as maiores variações de ordenamento aparecendo a partir da oitava posição.

Tabela 2 – 1ª e 2ª fases do filtro MIFS-U

1ª Fase	
Posição	**Variável**
1	Idade
2	QRP
3	Creatinina
4	Genótipo DD
5	Genótipo E2E2
6	Genótipo E4E4
7	IMC
8	Genótipo E2E3
9	HAS
10	Genótipo E3E4
11	Genótipo E2E4
12	Diabete Melito
13	Genótipo MM
14	Genótipo E3E3
15	Genótipo TT
16	Sexo
17	Colesterol HDL
18	Genótipo MT
19	Genótipo ECA II
20	Tabagismo
21	Colesterol Total
22	TSCA
23	Genótipo DI
24	Classe Killip
25	Triglicerídeos
26	IMP
27	Escolaridade
28	Freqüência Cardíaca

2ª Fase	
Posição	**Variável**
1	Idade
2	QRP
3	Creatinina
4	IMC
5	Genótipo DD
6	Genótipo E4E4
7	Genótipo E2E2
8	Sexo
9	Genótipo E3E4
10	Diabete Melito
11	Genótipo E2E4
12	Genótipo E2E3
13	Genótipo MM
14	HAS
15	Genótipo TT
16	Genótipo E3E3

A Tabela 3 resume os resultados da implementação dos modelos ν-SVM. O classificador constituído com as três primeiras variáveis selecionadas pelo filtro MIFS-U (Idade, QRP e Creatinina) obteve os melhores resultados. Desta forma, o classificador mais eficiente utiliza as informações contidas em uma variável inteira (Idade), uma variável categórica (QRP) e uma variável contínua (Creatinina). Ademais, as variáveis genéticas e de diagnóstico não entraram na construção do classificador ótimo. Além disso, este classificador constituído por três variáveis mostrou-se claramente superior a qualquer outro subconjunto testado, seja em termos do coeficiente de Pearson quanto em relação aos conceitos de acurácia, sensibilidade e especificidade.

Tabela 3 – Resultados dos modelos de ν-SVM

Variáveis	s (%)	e (%)	a (%)	CP (%)
Idade	0,85	0,97	0,97	0,7696
Idade, QRP	0,91	0,97	0,98	0,7696
Idade, QRP, Creatinina	0,91	0,99	0,98	0,8483
Idade, QRP, Creatinina, IMC	0,83	0,99	0,97	0,7950
Idade, QRP, Creatinina, IMC, Genótipo DD	0,90	0,97	0,96	0,7561
Idade, QRP, Creatinina, IMC, Genótipos DD e E4E4	0,77	0,96	0,95	0,6479

5. Conclusão

Ao longo deste trabalho, procurou-se construir uma ferramenta computacional capaz de predizer o risco de morte dos pacientes internados com síndrome coronariana aguda. Para isto, elaborou-se um modelo matemático a partir do uso do modelo SVM, que combina otimização quadrática convexa, função núcleo e teoria de generalização.

Para seleção das variáveis de entrada, integrou-se a abordagem *wrapper* ao filtro MIFS-U. A partir desta proposta, foram construídos os modelos ν-SVM fazendo uso da função núcleo tangente hiperbólica. O treinamento e teste dos classificadores foram realizados através da técnica *leave-one-out*. Os resultados computacionais indicam que o subconjunto mais relevante de variáveis de entrada é constituído pelas variáveis Idade, Qualquer Revascularização Prévia e Creatinina, que permitem a construção de um classificador com sensibilidade de 91% e especificidade de 99%.

Para futuros trabalhos, sugere-se a comparação dos resultados aqui apresentados com o de outros algoritmos de classificação, como por exemplo, as RNA, de forma a se verificar se o viés do algoritmo de classificação é capaz de influenciar a escolha das variáveis mais relevantes.

Referências Bibliográficas

Baldi, P., Brunak, S., **Chauvin**, Y. (2000), Assessing the accuracy of prediction algorithms for classification: an overview, *Bioinformatics Review*, v. 16, n. 5, pp. 412-424.

Bathia, S., **Prakash**, P., **Pillai**, G.N. (2008), SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features, *Proceedings of the World Congress on Engineering and Computer Science*.

Blum, A. L. e **Langley**, P. (1997), Selection of relevant features and examples in machine learning, *Artificial Intelligence*, v. 97, n. 1-2, pp. 245-271.

Chen, P.H., **Lin**, 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.

Çomak, E., **Arslam**, A., **Türkoğlu**. I. (2007), A decision suppor system based on support vector machines for diagnosis of the heart valve diseases, *Computers in Biology and Medicine*, v. 37, pp. 21-27.

Conforti, D. e **Guido**, R. (2005), Kernel-based Support Vector Machine classifiers for early detection of myocardial infarction, *Optimization Methods and Software*, v. 20, n. 2-3, pp. 395-407.

Cristianini, N. e **Shawe-Taylor**, J., *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2006.

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

Guyon, I. e **Elisseeff**, A (2003), An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, v. 3, pp. 1157-1182.

Jankowski, S., **Szymanski**, Z., **Piatkowska-Janko**, E. e **Oreziak**, A. (2007), Improved recognition of sustained ventricular tachycardia from SAECG by support vector machine, *Anatol Journal Cardiology*, v. 7, pp. 112-115.

Karatzoglou, A., **Meyer**, D, **Hornik**, K. (2006), Support Vector Machines in R, *Journal of Statistical Software*, v. 15, n. 9, pp. 1-28. (Disponível em: <http://www.jstatsoft.org/v15/i09/paper>. Acesso em: 17 abr. 2009.)

Kwak, N. e **Choi**, C. H. (2002), Input feature selection for classification problems, *IEEE Transactions on Neural Networks*, v. 13, n. 1, pp. 143-159.

Millet-Roig, J., **Ventura-Galiano**, R., **Chorro-Gascó**, F. J. e **Cebrián**, A. (2000), Support Vector Machine for Arrythmia discrimination with Wavelet-Transform-Based Feature Selection, *Computers in Cardiology*, v. 27, pp. 407-410.

Salappa, A., **Doumpos**, M., **Zopounidis**, C. (2007), Feature selection algorithms in classification problems: an experimental evaluation, *Optimization Methods and Software*, v. 22, n. 1, pp. 199-214.

Schölkopf, B. e **Smola**, A. J., *Learning with Kernels. Support Vector Machines, Regulararization, Optimization, and Beyond*, MIT Press, Cambridge, 2002.

Schölkopf, B. e **Smola**, A. J., **Williamson**, R. C., **Bartlett**, P. L. (2000), New support vector algorithms, *Neural Computation*, v. 12, pp. 1207-1245.

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