UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Nicole Holsbach

MÉTODO DE MINERAÇÃO DE DADOS PARA DIAGNÓSTICO DE CÂNCER DE MAMA BASEADO NA SELEÇÃO DE VARIÁVEIS

Porto Alegre
2012

Nicole Holsbach

Método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientador: Flávio Sanson Fogliatto, *Ph.D.*

Co-Orientador: Michel José Anzanello, *Ph.D.*

Porto Alegre

2012

Nicole Holsbach

Método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Flávio Sanson Fogliatto, *Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. Michel Jose Anzanello, *Ph.D.*
Co-Orientador PPGEP/UFRGS

_____

Prof. Carla Schwengber ten Caten
Coordenadora PPGEP/UFRGS

Banca Examinadora:

Professora Márcia Elisa Soares Echeveste, Dra. (DEST/UFRGS)

Professor Eduardo Jorge Valadares Oliveira, Dr. (UEPB)

Professora Ana Rita Facchini, Dra. (CRETIES/DEPROT)

AGRADECIMENTOS

Agradeço a todos os colaboradores que contribuíram para essa dissertação, em especial:

A minha mãe, por tudo.

Ao Márcio, pela motivação.

Ao meu marido, Ricardo, pela paciência.

As minhas irmãs, Ilesca e Ingrid, pela torcida.

A grande amiga Ana Rita, pelo apoio incondicional.

Ao meu orientador, Prof. Flávio Sanson Fogliatto, *Ph.D.*, pela sua avaliação extremamente criteriosa que possibilitou meu crescimento profissional.

Ao meu co-orientador Prof. Michel José Anzanello, *Ph.D.*, pela fé em mim depositada.

A Dr. Letícia Funchal Terres, radiologista de mamografia, pela sua importante contribuição.

Aos professores, funcionários e colegas do Programa de Pós-Graduação em Engenharia de Produção (PPGEP) e do Centro de Referências em Tecnologias e Insumos estratégicos para a Saúde (CRETIES) da Universidade Federal do Rio Grande do Sul (UFRGS), pelo conhecimento que adquiri e pelas oportunidades.

HOLSBACH, Nicole Holsbach ***Método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis,*** 2012. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

A presente dissertação propõe métodos para mineração de dados para diagnóstico de câncer de mama (CM) baseado na seleção de variáveis. Partindo-se de uma revisão sistemática, sugere-se um método para a seleção de variáveis para classificação das observações (pacientes) em duas classes de resultado, benigno ou maligno, baseado na análise citopatológica de amostras de célula da mama de pacientes. O método de seleção de variáveis para categorização das observações baseia-se em 4 passos operacionais: (*i*) dividir o banco de dados original em porções de treino e de teste, e aplicar a ACP (Análise de Componentes Principais) na porção de treino; (*ii*) gerar índices de importância das variáveis baseados nos pesos da ACP e na percentagem da variância explicada pelos componentes retidos; (*iii*) classificar a porção de treino utilizando as técnicas KVP (*k*-vizinhos mais próximos) ou AD (Análise Discriminante). Em seguida eliminar a variável com o menor índice de importância, classificar o banco de dados novamente e calcular a acurácia de classificação; continuar tal processo iterativo até restar uma variável; e (*iv*) selecionar o subgrupo de variáveis responsável pela máxima acurácia de classificação e classificar a porção de teste utilizando tais variáveis. Quando aplicado ao WBCD (*Wisconsin Breast Cancer Database*), o método proposto apresentou acurácia média de 97,77%, retendo uma média de 5,8 variáveis. Uma variação do método é proposta, utilizando quatro diferentes tipos de *kernels* polinomiais para remapear o banco de dados original; os passos (*i*) a (*iv*) acima descritos são então aplicados aos *kernels* propostos. Ao aplicar-se a variação do método ao WBCD, obteve-se acurácia média de 98,09%, retendo uma média de 17,24 variáveis de um total de 54 variáveis geradas pelo *kernel* polinomial recomendado. O método proposto pode auxiliar o médico na elaboração do diagnóstico, selecionando um menor número de variáveis (envolvidas na tomada de decisão) com a maior acurácia, obtendo assim o maior acerto possível.

Palavras-chave: Seleção de variáveis, Diagnóstico de câncer de mama, k-vizinhos mais próximos, Análise discriminante, *Kernel*.

HOLSBACH, Nicole Holsbach ***A data mining method for breast cancer diagnosis based on selected features,*** 2012. Dissertation (Master in Engineering) - Federal University of Rio Grande do Sul, Brazil.

ABSTRACT

This dissertation presents a data mining method for breast cancer (BC) diagnosis based on selected features. We first carried out a systematic literature review, and then suggested a method for feature selection and classification of observations, i.e., patients, into benign or malignant classes based on patients' breast tissue measures. The proposed method relies on four operational steps: (*i*) split the original dataset into training and testing sets and apply PCA (Principal Component Analysis) on the training set; (*ii*) generate attribute importance indices based on PCA weights and percent of variance explained by the retained components; (*iii*) classify the training set using KNN (*k*-Nearest Neighbor) or DA (Discriminant Analysis) techniques, eliminate irrelevant features and compute the classification accuracy. Next, eliminate the feature with the lowest importance index, classify the dataset, and re-compute the accuracy. Continue such iterative process until one feature is left; and (*iv*) choose the subset of features yielding the maximum classification accuracy, and classify the testing set based on those features. When applied to the WBCD (Wisconsin Breast Cancer Database), the proposed method led to average 97.77% accurate classifications while retaining average 5.8 features. One variation of the proposed method is presented based on four different types of polynomial *kernels* aimed at remapping the original database; steps (*i*) to (*iv*) are then applied to such *kernels*. When applied to the WBCD, the proposed modification increased average accuracy to 98.09% while retaining average of 17.24 features from the 54 variables generated by the recommended *kernel*. The proposed method can assist the physician in making the diagnosis, selecting a smaller number of variables (involved in the decision-making) with greater accuracy, thereby obtaining the highest possible accuracy.

Keywords: Feature selection, Breast cancer diagnosis, *k*-nearest Neighbor, Discriminant, *kernel*.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 Introdução

1.1 Considerações Iniciais

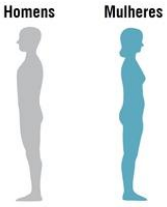
O quadro de saúde da população brasileira apresenta avanços na redução significativa de alguns problemas (por exemplo, reduções nas Taxas de Mortalidade Infantil (TMI), na taxa de desnutrição em crianças e na ocorrência das doenças imunopreveníveis), reduções menos significativas ou mesmo estabilidade em relação a outros, ou ainda problemas que apresentam tendência ao crescimento (reaparecimento do cólera, dengue e tuberculose), que tornam a população vulnerável a doenças que deveriam estar extintas (BRASIL, 2012). Observa-se ainda no país o desenvolvimento de novas tecnologias médicas, o acesso de maior parte da população a tais tecnologias e a redução das taxas de mortalidade de algumas doenças do grupo das Doenças Crônicas Não Transmissíveis (DCNT). Mesmo assim, outros problemas relacionados a essas doenças e alguns dos seus fatores de risco apresentam crescimento, podendo ocasionar uma cadeia de eventos que levam a efeitos danosos na saúde da população e à crescente demanda por serviços de saúde. Dentre os principais problemas estão: (*i*) o aumento da prevalência da obesidade em adultos; (*ii*) as Doenças Cardiovasculares (DCVs); (*iii*) as doenças respiratórias crônicas; (*iv*) as doenças neuropsiquiátricas, e (*v*) os diferentes tipos de câncer que vêm apresentando tendências diferentes de mortalidade. Entre as mulheres, destacam-se os aumentos em câncer de mama (CM) que responde por 22% dos casos novos a cada ano e de pulmão, sendo que as taxas de incidência dos cânceres de mama e cervical entre as brasileiras estão entre as mais altas do mundo. Além disso, as taxas de sobrevivência estão abaixo das observadas em países desenvolvidos, que é de uma sobrevida média após 5 anos de 61% na população mundial, refletindo diagnóstico tardio e falhas nos tratamentos. Para a redução da ocorrência dos cânceres, será necessário o fortalecimento do sistema de saúde, incluindo diagnóstico precoce, acesso ao tratamento adequado e implementação de medidas de redução e controle de fatores de risco (BRASIL, 2012). A estimativa de novos casos para 2012 de CM é de 52.680, e o número de óbitos em 2010 foi de 12.852, sendo 147 homens e 12.705 mulheres (BRASIL, 2004).

Câncer consiste em um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células que invadem tecidos e órgãos. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de

tumores malignos, que podem espalhar-se para outras regiões do corpo. Mesmo com os avanços na detecção e tratamento precoce, o câncer está evoluindo para uma condição crônica em muitos países. Estatísticas indicam aumento de sua incidência tanto nos países desenvolvidos quanto nos em desenvolvimento. O CM é o segundo tipo mais frequente no mundo e é o mais comum entre as mulheres (IARC, 2002; WHO, 2007). A doença consiste no crescimento desordenado de células do tecido da mama, formando nódulos que podem ser malignos (tumores) ou benignos.

A epidemiologia referente ao CM no Brasil, com seus elevados índices de incidência e mortalidade, justificam a implementação de ações nacionais voltadas para a prevenção e o controle do câncer (promoção, prevenção, diagnóstico, tratamento, reabilitação e cuidados paliativos). Ações de detecção precoce, como a adoção de estratégias (padronização de procedimentos e de condutas que garantam a qualidade dos processos técnicos e operacionais para o controle do câncer), são fundamentais para o controle da doença. Políticas públicas nessa área vêm sendo desenvolvidas no Brasil desde os anos 80, quando tiveram início ações para estruturação da rede assistencial na detecção precoce do CM (BRASIL, 2011). No entanto, as taxas de mortalidade por CM continuam elevadas, sobretudo porque a doença ainda é diagnosticada em estados avançados. Quando diagnosticado precocemente, as taxas de sobrevivência em pacientes com CM aumentam, e a morbidade associada ao tratamento diminui (BRASIL, 2011; IARC, 2002; WHO, 2007). O rastreamento do câncer de mama em condições de sucesso (cobertura da população alvo, qualidade dos exames de rastreamento e garantia de acesso ao diagnóstico e tratamento) reduz em 30% a sua mortalidade (BRASIL, 2009). Na Figura 1.1 são apresentadas as estimativas de incidência de câncer no Brasil para o ano de 2012.

Localização primária	casos novos	percentual
Próstata	60.180	30,8%
Traqueia, Brônquio e Pulmão	17.210	8,8%
Cólon e Reto	14.180	7,3%
Estômago	12.670	6,5%
Cavidade Oral	9.990	5,1%
Esôfago	7.770	4,0%
Bexiga	6.210	3,2%
Laringe	6.110	3,1%
Linfoma não Hodgkin	5.190	2,7%
Sistema Nervoso Central	4.820	2,5%



Localização primária	casos novos	percentual
Mama Feminina	52.680	27,9%
Colo do Útero	17.540	9,3%
Cólon e Reto	15.960	8,4%
Glândula Tireoide	10.590	5,6%
Traqueia, Brônquio e Pulmão	10.110	5,3%
Estômago	7.420	3,9%
Ovário	6.190	3,3%
Corpo do Útero	4.520	2,4%
Sistema Nervoso Central	4.450	2,4%
Linfoma não Hodgkin	4.450	2,4%

Figura 1.1 - Distribuição proporcional dos dez tipos de câncer mais incidentes no Brasil estimados para 2012 por sexo, exceto pele não melanoma (números arredondados para 10 ou múltiplos de 10). Fonte: BRASIL, 2011.

A presente dissertação apoia-se em um método de mineração de dados para o diagnóstico de câncer de mama baseado na seleção de variáveis, oriundas de exames clínicos, para fins de classificação das observações (pacientes) em duas classes de resultado, benigno ou maligno, baseado na análise citopatológica de amostras de célula da mama de pacientes. O método de classificação proposto testa seu desempenho no *Wisconsin Breast Cancer Database* (WBCD), obtido da universidade de Wisconsin e disponibilizado *online*. O WBCD é composto por 699 observações (16 delas incompletas) obtidas a partir da punção aspirativa por agulha fina (PAAF) de células da mama. A PAAF é um procedimento médico direcionado à investigação de pacientes com massas, que permite a investigação da malignidade em nódulos mamários (ALBRECHT *et al*., 2002). A técnica consiste na retirada de pequena porção de tecido por aspiração através de uma agulha fina e posterior coloração e análise microscópica. A classe (benigna ou maligna) a que cada observação pertence é conhecida. Na amostra de 683 valores completos utilizada nesta análise, há 239 casos malignos e 444 casos benignos. O banco de dados está disponível em http://www.ics.uci.edu/~mlearn/MLRepository. O WBCD foi selecionado para testar o desempenho do método proposto porque a literatura apresenta um grande número de pesquisas sobre técnicas de classificação utilizando esse banco de dados.

Esta dissertação é composta por três artigos. No primeiro artigo é apresentada uma revisão sistemática sobre seleção de variáveis para fins de classificação. O segundo artigo propõe um método para seleção de variáveis com vistas à classificação de observações do WBCD em duas categorias distintas. O método proposto integra uma técnica multivariada (ACP) a dois métodos de classificação (KVP e AD). O terceiro artigo apresenta uma variação do método proposto no segundo artigo, aplicando funções *kernel* do tipo polinomial ao banco de dados original com vistas ao remapeamento das observações e aumento da acurácia de classificação. O software MATLAB foi utilizado na operacionalização das técnicas do método proposto.

1.2 Objetivos

O principal objetivo do trabalho é propor um método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis.

Como objetivos específicos têm-se:

- Apresentar os principais métodos de seleção de variáveis para fins de classificação de pacientes a partir de diagnósticos médicos por meio de revisão sistemática da literatura;
- Propor um método para a seleção de variáveis para classificação dos casos em duas classes de resultado (benigno e maligno);
- Propor a integração de uma técnica multivariada (ACP) a dois métodos de classificação, KVP e AD, valendo-se de um novo índice de importância baseado em parâmetros oriundos da ACP;
- Aplicar transformações *kernel* nos dados originais, de maneira a melhorar o desempenho de classificação do método proposto;
- Avaliar os métodos propostos em um banco de dados real contendo informações sobre câncer de mama, mensurando seu desempenho através de indicadores de acurácia de classificação, número de variáveis retidas, sensibilidade e especificidade para diferentes proporções de porções de treino e teste do banco de dados.
- O método proposto pode auxiliar o médico na elaboração do diagnóstico, selecionando um menor número de variáveis (envolvidas na tomada de decisão) com a maior acurácia, obtendo assim o maior acerto possível.

1.3 Justificativa do Tema e dos Objetivos

O diagnóstico do CM depende da avaliação do médico a partir das informações obtidas dos pacientes através de exames. Esses exames incluem exame clínico da mama (autopalpação), mamografia e análise de tecido da mama. Em termos de análise, o exame clínico da mama fornece dados univariados para interpretação, enquanto que exames laboratoriais (citopatológicos) e de imagem produzem dados multivariados, os quais demandam maior processamento de informações.

Grandes bancos de dados oriundos de diagnóstico médico são usualmente constituídos por centenas de variáveis correlacionadas e ruidosas, o que compromete o

desempenho de várias técnicas estatísticas, como a análise de regressão e métodos de classificação de observações (ANZANELLO, 2009). De tal forma, a eliminação de variáveis que não contribuem para uma conclusão é importante, já que variáveis ruidosas tendem a produzir um modelo preditivo errôneo ou incompleto. A seleção de variáveis para classificação tem por objetivo identificar um subgrupo de variáveis com maior poder de discriminação das observações em classes distintas (ANZANELLO; FOGLIATTO, 2011).

Considerando as abordagens existentes na literatura para seleção de variáveis com vistas à classificação de pacientes a partir de diagnóstico médico, nota-se que não há método genérico para tal finalidade e existe espaço para o desenvolvimento de métodos de seleção mais eficientes e robustos. Tais modelos devem ser úteis na tomada de decisão médica, como por exemplo, no prognóstico ou diagnóstico. Um objetivo recorrente nesses estudos é selecionar um modelo que se ajuste bem aos dados, de fácil interpretação e que seja útil na prática clínica. Para tanto, é necessário decidir quais variáveis incluir no modelo (ROYSTON; SAUERBREI, 2003). Entende-se que os métodos de seleção devem classificar corretamente os dados, gerando a maior acurácia possível e retendo o menor número de variáveis. É importante lembrar que, para o rastreamento do CM, o método deve ser o mais sensível para que se consiga detectar o maior número de casos possível da doença.

Desta forma, justifica-se o desenvolvimento de métodos para seleção de variáveis com finalidade de classificação em duas categorias distintas a partir de exames médicos, tema dos três artigos dessa dissertação.

1.4 Procedimentos Metodológicos

A pesquisa pode ser classificada como de natureza aplicada, uma vez que o conteúdo teórico é explorado com vistas à solução de problemas genéricos, de abordagem quantitativa, em função das análises numéricas realizadas, de objetivo exploratório, uma vez que busca construir hipóteses para resolver os problemas a partir da sua análise, e de procedimentos bibliográficos, uma vez que foi feita uma revisão sistemática da literatura. A revisão sistemática é um método de síntese da literatura, a qual permite extrapolar achados de estudos

independentes, avaliar a consistência de cada um deles e explicar as possíveis inconsistências e conflitos (MULROW, 1996; GIL, 2002; SILVA; MENEZES, 2000).

1.5 Estrutura da Dissertação

A dissertação está organizada em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos, justificativas e o método de pesquisa adotado, bem como a delimitação e estrutura do trabalho.

O segundo capítulo traz o primeiro artigo, que apresenta uma revisão da literatura sobre métodos de seleção de variáveis com vistas à classificação de observações no diagnóstico médico. A fundamentação dos métodos mais relevantes para tal propósito é apresentada, bem como as aplicações mais recentes para classificação em bancos de dados da área médica e suas limitações. Também são descritas as principais medidas de desempenho de classificação utilizadas nas abordagens, tais como matriz de confusão, acurácia, especificidade, sensibilidade e análise ROC (*Receiver Operating Characteristic*).

O terceiro capítulo apresenta o segundo artigo, onde é proposto um método para seleção de variáveis oriundas de exames clínicos com vistas à classificação de observações em duas categorias. O método de seleção de variáveis para a categorização das observações do WBCD baseia-se em 4 passos operacionais: (*i*) dividir o banco de dados original em porções de treino e de teste, e aplicar a ACP (Análise de Componentes Principais) na porção de treino; (*ii*) gerar índices de importância das variáveis baseados nos pesos da ACP e na percentagem da variância explicada pelos componentes retidos; (*iii*) classificar a porção de treino utilizando KVP (*k*-vizinhos mais próximos) e AD (Análise Discriminante) separadamente. Em seguida, eliminar a variável com o menor índice de importância, classificar o banco de dados novamente, e calcular a acurácia de classificação. Continuar tal processo iterativo até restar uma variável; e (*iv*) selecionar o subgrupo de variáveis que gera a máxima acurácia de classificação e classificar a porção de teste utilizando essas variáveis. Uma contribuição importante deste artigo é a integração de uma técnica multivariada (ACP) a dois métodos de classificação, KVP e AD, e a proposição de um novo índice de importância baseado em parâmetros da ACP, o qual guia a eliminação recursiva de variáveis.

O quarto capítulo traz o terceiro artigo, onde é apresentado um método para melhorar o desempenho do método proposto no segundo artigo. Para tanto, os dados originais são transformados utilizando quatro tipos de *kernel* polinomial; tais transformações permitem o remapeamento das observações e potencialmente conduzem a arranjos passíveis de melhor categorização. Os passos de (*i*) a (*iv*) descritos no segundo artigo são então repetidos para as diferentes escolhas de função polinomial *kernel*. Uma contribuição significativa deste trabalho consiste na utilização de *kernels* no adensamento de bancos de dados, com vistas à melhoria do desempenho de classificação dos métodos de mineração de dados.

O quinto capítulo apresenta a conclusão do trabalho, onde são avaliados os resultados em acordo com os objetivos desejados e as delimitações do trabalho. Essa seção também traz sugestões para desenvolvimentos futuros.

1.6 Delimitações do Estudo

As restrições do presente estudo são:

- O banco de dados WBCD (*Winsconsin Breast Cancer Dataset*), que contém observações de pacientes com CM, foi escolhido para avaliar o desempenho do método proposto porque a literatura apresenta um grande número de pesquisas sobre técnicas de classificação utilizando este banco de dados.
- O banco de dados usado para aplicação é limitado ao número de amostras e às características dos pacientes (conforme sua nacionalidade) que compõem o banco; e
- As variáveis estudadas, compostas por características das células da mama, são selecionadas apenas com o objetivo de classificação em duas classes de resultado (benigno e maligno).

1.7 Referências

ALBRECHT, A. A.; LAPPAS, G.; VINTERBO, S. A.; WONG, C.K.; OHNO-MACHADO, L. Two applications of the LSA macrine. In: **Proceedings of the 9th**

International Conference on Neural Information Processing, Nov 18-22, Singapore, p.184-189, 2002.

ANZANELLO, M. J. **Seleção de Variáveis com vistas à Classificação de Bateladas de Produção em duas Classes**. Gestão da Produção, São Carlos, v. 16, n. 4, p. 526-533, out.-dez. 2009.

ANZANELLO, M. J.; FOGLIATTO, F. S . **Identificação de Variáveis Relevantes para Categorização de Bateladas de Produção com Base em Critérios de Sensibilidade E especificidade**. In: XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011, Ubatuba. XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011.

BRASIL. Ministério da Saúde/Ipea/Secretaria de Assuntos estratégicos da Presidência da República. A saúde no Brasil em 2030: diretrizes para a prospecção estratégica do sistema de saúde brasileiro (2012). 1 ed. Rio de Janeiro - RJ. 323p.

BRASIL. Ministério da Saúde. Programa Nacional de Controle do Câncer de Mama. Versão revista e ampliada do Programa Viva Mulher, desmembrado em Programa Nacional de Controle do Câncer do Colo do Útero e Programa Nacional de Controle do Câncer de Mama (INCA, 2010), elaborado pela Divisão de Apoio à Rede de Atenção Oncológica em abril de 2011. Brasília: Ministério da Saúde, 2011. 15p http://www2.inca.gov.br/wps/wcm/connect/521d4900470039c08bd8fb741a182d6f/pncc_mama.pdf?MOD=AJPERES&CACHEID=521d4900470039c08bd8fb741a182d6f. Acesso em 16/12/2012.

BRASIL. Ministério da Saúde. Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação Geral de Ações Estratégicas. Coordenação de Prevenção e Vigilância. Estimativa 2012: incidência de câncer no Brasil / Instituto Nacional de Câncer José Alencar Gomes da Silva, Coordenação Geral de Ações Estratégicas, Coordenação de Prevenção e Vigilância. – Rio de Janeiro: Inca, 2011. 118p. http://www.inca.gov.br/estimativa/2012/estimativa20122111.pdf. Acesso em 08/01/2012.

BRASIL. Ministério da Saúde. Instituto Nacional de Câncer (INCA). Controle do Câncer de Mama. Documento de Consenso. ***INCA*** [site na Internet]. 2004 Abr [acessado

2012 jul 23]; [cerca de 39 p.]. Disponível em: http://www1.inca.gov.br/publicacoes/ Consensointegra.pdf

GIL, A. C. (2002). **Como elaborar projetos de pesquisa**. 4 ed. São Paulo - Atlas. 176p.

International Agency for Research on Cancer (IARC). IARC Handbooks of Cancer Prevention Volume 7: Breast Cancer Screening. Lyon: ***IARC*** [serial on the Internet] 2002 [cited 2012 jul 23]; [about 243 p.]. Available from: http://www.iarc.fr/en/ publications/pdfs-online/prev/handbook7/index.php

MULROW, C. Rationale for systematic reviews. In: I Chalmes & D. G. Altman, Systematic Reviews. 3 ed. Bmj Publishing Group. London, 1996.

ROYSTON, P.; SAUERBREI, W. Stability of Multivariable Fractional Polynomial Models with Selection of Variables and Transformations: A Bootstrap Investigation. **Statistics in Medicine**, v. 22, p. 639–659, 2003.

SILVA, E. L.; MENEZES, E. M.(2000) – Metodologia da pesquisa e elaboração de dissertação. UFSC/PPGEP/LED, Florianópolis – SC.

World Health Organization (WHO). Cancer control: knowledge into action: WHO guide for effective programmes: early detection. ***WHO 2007*** [serial on the Internet] 2007 [cited 2012 Aug 09]; [about 50 p.]. Available from: http://www.who.int/cancer/ publications/cancer_control_detection/en/

2 Primeiro Artigo: Métodos de seleção de variáveis para fins de classificação de indivíduos a partir de diagnósticos médicos

Nicole Holsbach

Flávio Sanson Fogliatto

Michel José Anzanello

Resumo

Os bancos de dados oriundos de diagnóstico médico são constituídos por centenas de variáveis correlacionadas e ruidosas (ou incertas), o que compromete o desempenho de técnicas estatísticas. A eliminação de variáveis que não contribuem para uma conclusão é importante, visto que variáveis ruidosas tendem a produzir um modelo preditivo errôneo ou incompleto. Na seleção de variáveis para predição o objetivo é encontrar um subgrupo de variáveis de entrada que assegure a predição precisa de uma ou mais variáveis de resposta. Na seleção de variáveis para classificação o objetivo é identificar um subgrupo de variáveis de entrada a fim de categorizar as observações em classes. O câncer de mama (CM) se tornou uma das maiores causas de mortalidade no mundo, e a disponibilidade de um diagnóstico preciso é imprescindível para a sua detecção precoce. A literatura apresenta um grande número de pesquisas sobre técnicas de classificação em bancos públicos de dados, tais como o WBCD (*Winsconsin Breast Cancer Dataset*), que contém observações de pacientes com CM. A maioria das abordagens revelou um aumento da acurácia quando as variáveis mais relevantes são utilizadas. Este artigo apresenta uma revisão da literatura sobre métodos de seleção de variáveis com vistas à classificação de observações no diagnóstico médico. A fundamentação dos métodos mais relevantes e as aplicações para classificação em bancos de dados são apresentadas. Também são descritas as principais medidas de desempenho de classificação utilizadas nas abordagens, tais como matriz de confusão, acurácia, especificidade, sensibilidade e análise ROC (*Receiver Operating Characteristic*).

Palavras-chave: Seleção de variáveis, Classificação, Diagnóstico.

Abstract

The databases from medical diagnostics consist of hundreds of correlated and noisy variables, which compromises the performance of statistical techniques. The elimination of variables that do not contribute to a conclusion is important, since noisy variables tend to produce a predictive model erroneous or incomplete. In selecting variables for prediction the goal is to find a subset of input variables to ensure accurate prediction of one or more response variables. In selecting variables for classification the goal is to identify a subset of input variables to categorize notes in classes. Breast cancer (BC) has become a major cause of mortality worldwide, and the availability of an accurate diagnosis is essential for early detection. The literature contains a large number of researches on classification techniques in public databases, such as WBCD (Wisconsin Breast Cancer Dataset), which contains observations of patients with BC. Most approaches revealed an increase in accuracy as the most important variables are used. This article presents a review of the literature on variable selection methods aiming at classifying observations in medical diagnosis. The reasons most relevant tools and applications for classification databases are presented. Also described are the key performance measures used in classification approaches, such as confusion matrix, accuracy, specificity, sensitivity analysis and ROC (Receiver Operating Characteristic).

Keywords: Feature selection, Classification, Diagnostic.

2.1 Introdução

O volume de dados produzido nos hospitais e centros médicos está crescendo rapidamente. A produção anual de dados gerados por exames de imagens nos grandes centros de radiologia é da ordem de 2 *terabytes* **(**OLIVEIRA *et al*., 2007). A crescente importância e utilização dos exames de imagem têm levado à obtenção e armazenamento de dados dos pacientes. Os dados armazenados e indexados são fundamentais no diagnóstico clínico, pois dão apoio ao diagnóstico, prognóstico e decisão terapêutica (OLIVEIRA *et al*., 2007). A interpretação dos resultados somada aos métodos computacionais de análise de imagens é bastante útil, especialmente nos casos onde as suspeitas clínicas de câncer, por exemplo, prevalecem depois de biópsias com resultados negativos pelos métodos tradicionais (ARAÚJO-FILHO *et al*., 2006).

Grandes bancos de dados, como os oriundos de diagnóstico médico, são usualmente constituídos por centenas de variáveis correlacionadas e ruidosas, o que compromete o desempenho de várias técnicas estatísticas, como a análise de regressão e métodos de classificação de observações (ANZANELLO, 2009).

A eliminação de variáveis que não contribuem para uma conclusão ou dedução é importante em diversos contextos, visto que variáveis ruidosas tendem a produzir um modelo preditivo errôneo ou incompleto. Além disso, a identificação de variáveis relevantes baseada em conhecimento empírico pode estar incorreta. O propósito da seleção de variáveis para predição consiste em encontrar um subgrupo de variáveis de entrada (variáveis independentes) que assegure a predição precisa de uma ou mais variáveis de resposta (variáveis dependentes) (ANZANELLO; FOGLIATTO, 2011). Já a seleção de variáveis para classificação tem por objetivo identificar um subgrupo de variáveis de entrada a fim de categorizar as observações em classes (em diagnósticos médicos, por exemplo, positivos significariam pacientes doentes, e negativos pacientes não doentes).

Na seleção de variáveis para predição, parte-se da relação entre as variáveis de entrada e de resposta, a qual pode ser formalizada através de modelos de Regressão Múltipla Linear (RML). O objetivo é identificar um grupo de variáveis de entrada capaz de explicar e predizer uma ou mais variáveis de resposta (GUO; TANAKA, 2006). Os métodos *Backward*, *Forward*, e *Stepwise* consistem-se nos procedimentos básicos para seleção de variáveis na RLM (ANZANELLO; FOGLIATTO, 2011), porém métodos de seleção mais robustos vêm sendo desenvolvidos com propósitos de predição, em diversas áreas do conhecimento.

Na seleção de variáveis para classificação, o objetivo é categorizar as observações em classes em relação a alguma especificação. A maioria das abordagens de seleção de variáveis para classificação apoia-se em métodos de mineração de dados (ANZANELLO; FOGLIATTO, 2011). O desempenho dos métodos de classificação é normalmente avaliado pela matriz de confusão, por medidas como acurácia, sensibilidade e especificidade e através da análise ROC (*Receiver Operating Characteristic*).

Um exemplo da importância da utilização de variáveis relevantes com vistas à classificação de observações vem da área médica, onde, por exemplo, o CM se tornou uma das maiores causas de mortalidade no mundo, e a disponibilidade de um diagnóstico preciso e

rápido se tornou uma questão importante para a comunidade científica, no que diz respeito à sua prevenção. A detecção precoce da presença de células cancerígenas com base em dados obtidos dos pacientes e em decisões dos especialistas são fatores fundamentais para o diagnóstico da doença (MARCANO-CEDEÑO *et al*., 2011). A literatura apresenta um grande número de pesquisas sobre técnicas de classificação em bancos públicos de dados, tais como o *Winsconsin Breast Cancer Dataset* (WBCD), que contém observações de pacientes com CM. A maioria das abordagens revelou um aumento da acurácia quando as variáveis mais relevantes são utilizadas nos procedimentos em questão (MARCANO-CEDEÑO *et al*., 2011; ABBASS, 2002; SETIONO, 2000).

Este artigo apresenta uma revisão da literatura sobre métodos de seleção de variáveis com vistas à classificação de observações no diagnóstico médico. A fundamentação dos métodos mais relevantes para tal propósito é apresentada, bem como as aplicações mais recentes para classificação em bancos de dados da área médica e suas limitações. Também são descritas as principais medidas de desempenho de classificação utilizadas nas abordagens, tais como matriz de confusão, acurácia, especificidade, sensibilidade e análise ROC (*Receiver Operating Characteristic*).

Este artigo está estruturado em quatro seções, incluindo a presente introdução. A seção 2 deste artigo apresenta os procedimentos metodológicos. A seção 3 apresenta uma revisão sobre métodos de seleção de variáveis com o propósito de classificação e predição e uma revisão sobre os seguintes critérios de classificação: matriz de confusão, acurácia, especificidade, sensibilidade e análise ROC. A seção 4 traz uma análise crítica da literatura abordada, a indicação de lacunas e oportunidades de pesquisa, além de conclusões.

2.2 Procedimentos metodológicos

Utilizando o site eletrônico da CAPES disponível em http://www.capes.gov.br/, foi feita a busca pelos artigos na base de dados *Scopus*.

A busca utilizou as palavras chave *'variable/feature selection/extraction'*, '*classification*', '*prediction*' e '*diagnostic*', que apareciam nos títulos dos artigos, *abstract* ou palavras-chave, a partir do ano de 1990. Limitando a pesquisa utilizando outros critérios como tipo do documento ('*article*'), tipo da fonte ('*journals*') e área de interesse ('*Medicine',*

'Biochemistry, Genetics and Molecular Biology', 'Computer Science', 'Mathematics', 'Engineer', 'Immunology and Microbiology', 'Neuroscience', 'Pharmacology, Toxicology and Pharmaceutics', 'Health Professions' e 'Nursing'), chegou-se a 491 documentos.

Refinando a busca, procurando por '*cancer diagnosis' e 'medical diagnosis'*, restaram 69 artigos. Após a leitura dos *abstracts*, 20 documentos diretamente relacionados ao tema desta pesquisa foram selecionados. Na Tabela 2.1 é apresentado o número de artigos, conforme os métodos utilizados na seleção de variáveis, para fins de classificação e predição. Os artigos que se encontram na coluna '*Forward, Backward, Stepwise'* utilizaram pelo menos um desses métodos.

A revisão sistemática da literatura também inclui outras referências complementares, que foram selecionadas na busca de autores que abordavam técnicas de seleção de variáveis para fins de classificação e predição.

Tabela 2.1 - Principais métodos encontrados

	Métodos utilizados na Seleção de Variáveis							
	Forward, Backward, Stepwise	*Network Pruning*	*F -Score*	*LASSO*	*Random Forest*	*Bayesian*	*Recursive Bootstrap Elimination*	Total
Classificação	0	1	6	2	1	0	0	10
Predição	8	0	0	0	0	1	1	10
Total	8	1	6	2	1	1	1	20
Periódicos	*Critical Care, Clinical Cancer Research, Artificial Intelligence in Medicine, Statistics in Medicine, Revista Brasileira de Epidemiologia, Revista de Psiquiatria Clínica*	*Neural Networks*	*Pattern Recognition, J Med Syst, Systems with Applications, Digital Signal Processing, Expert Systems with Applications, Communications in Computer and Information Science*	*Journal of Biomedicine and Biotechnology, Statistics in Medicine, BMC Medical Research Methodology*	*BMC Bioinformatics, Journal of Royal Statistical Society*	*Cancer Research*	*Clinical Cancer Research*	

Fonte: elaborado pelos autores.

Na Tabela 2.1, *Forward, Backward e Stepwise* correspondem aos métodos de seleção avançada (progressiva), retro-seleção (seleção regressiva) e seleção gradual (passo a passo) de variáveis. O método *Forward* acrescenta variáveis de forma sequencial, conforme a ordem do índice de importância. O método *Backward* faz a eliminação dessas variáveis de forma sequencial, em ordem crescente de índice de importância. Por fim, o método *Stepwise* analisa a significância das variáveis, seguindo um valor limítrofe de referência, incluindo-as uma a uma no modelo. O método *Network Prunning* (Purificação da rede) consiste em remover as variáveis cuja influência é menos importante (PASTOR-BÁRCENAS, *et al.*,2005). O método *F-Score* (medida F) é um critério estatístico que faz o *ranking* de cada variável (CHANG *et al*., 2010). O método *LASSO*, sigla para *Least Absolute Shrinkage and Selection Operator* (TIBSHIRANI, 1996), promove a otimização dos coeficientes de regressão e reduz a magnitude e o número de coeficientes necessários para o processo de classificação (PAOLUCCI, 2006). *Random Forest* (Floresta Randômica - FR) é um conceito de árvores de regressão que usa seleção de características randômicas no processo de indução de árvore (BREIMAN, 2001). A partir das árvores de decisão se estabelece a classificação dos dados por votos, e a seleção de variáveis é feita no instante de construção do modelo de classificação (PEREIRA *et al.*, 2009). *Baysesian* corresponde à abordagem bayesiana, onde todos os parâmetros do modelo são considerados como variáveis aleatórias, conforme o conceito subjetivo de probabilidade. A partir de uma generalização do Teorema de Bayes, informações sobre parâmetros são utilizadas em associação com os dados amostrais, possibilitando uma inferência sobre os parâmetros (PEREIRA *et al*., 2009). *Recursive Bootstrap Elimination* (Eliminação Recursiva via *bootstrap* - ERB) é um método de reamostragem *bootstrap* que, além de fornecer estimativas de parâmetros e seus desvios-padrão, permite obter intervalos de confiança para os parâmetros analisados, bem como a distribuição empírica de suas estimativas (EFRON; TIBSHIRANI, 1993).

2.3 Referencial teórico

As técnicas de seleção de variáveis não alteram a representação original das variáveis, apenas selecionam um subconjunto delas (SAEYS *et al*., 2007). Existem dois grandes grupos de abordagens para seleção de variáveis, conforme a sua finalidade: predição ou classificação. As subseções que se seguem foram organizadas agrupando trabalhos conforme seu objetivo.

2.3.1 Seleção de variáveis para predição

Na seleção de variáveis para predição, o objetivo é identificar um grupo de variáveis de entrada capaz de explicar e predizer uma ou mais variáveis de resposta (GUO; TANAKA, 2006). A inclusão de variáveis irrelevantes pode aumentar a variância das estimativas, o que representa uma perda na capacidade preditiva do modelo. A inclusão de muitas variáveis pode levar a modelos difíceis de interpretar (ROCA-PARDINAS *et al*., 2009), demandando grande tempo de análise e não oferecendo garantias em termos de predição (ANZANELLO, 2009). Torna-se assim, essencial identificar as variáveis mais relevantes a serem incluídas em um modelo de predição.

Os métodos *Backward, Forward*, e *Stepwise* são procedimentos básicos para seleção de variáveis na Regressão Múltipla Linear – RML, um dos métodos mais usados de modelagem de dados para fins de predição (ANZANELLO; FOGLIATTO, 2011).

O método da Regressão Múltipla *Backward* começa com o modelo completo e elimina somente variáveis que dificilmente teriam qualquer influência no modelo (ROYSTON; SAUERBREI, 2003). O teste de significância, baseado no valor limite de *p* escolhido de acordo com o objetivo do estudo e utilizado como critério de parada, é realizado nos coeficientes de regressão (ROYSTON; SAUERBREI, 2003; ANZANELLO; FOGLIATTO, 2011). A variável com o maior valor *p* é removida e um novo teste é realizado para avaliar a significância das variáveis remanescentes. O processo é repetido até que as variáveis restantes tenham um valor menor que o valor limite definido. Royston e Sauerbrei (2003) consideram o uso da eliminação *backward* como vantajosa, pois começa com o modelo completo e vai eliminando as variáveis que dificilmente terão influência no modelo.

No método da Regressão Múltipla *Forward*, também realiza-se um teste de significância, porém as variáveis são introduzidas uma a uma no modelo, de acordo com seus valores *p* (do menor até o valor limite definido pelo usuário).

No método da Regressão Múltipla *Stepwise*, as variáveis podem entrar e sair do modelo baseado em limiares específicos de adição e exclusão (ANZANELLO; FOGLIATTO; WEBER *et al*., 2004). O procedimento termina quando não há variáveis que satisfaçam a nenhum dos critérios estabelecidos (WEBER *et al*., 2004).

Um problema recorrente na estatística médica é a construção de modelos multivariados confiáveis, a partir de diversos preditores (isto é, variáveis de entrada). Preditores candidatos normalmente são de natureza diversa, incluindo variáveis binárias, categóricas e contínuas. Em aplicações médicas é considerado importante o efeito individual das variáveis com vistas a sua inclusão no modelo. Em estudos observacionais na área de Medicina, um grande número de variáveis é frequentemente considerado com o objetivo de construir modelos de regressão confiáveis. Tais modelos devem ser úteis na tomada de decisão médica como, por exemplo, no prognóstico ou diagnóstico. Um objetivo recorrente nesses estudos é selecionar um modelo que se ajuste bem aos dados, de fácil interpretação e que seja útil na prática clínica. Para tanto, é necessário decidir quais variáveis incluir no modelo (ROYSTON; SAUERBREI, 2003).

Os métodos mais usados na prática são a utilização do modelo completo (que não envolve seleção de variáveis), estratégias sequenciais, como seleção de variáveis *forward* e *backward* e procedimentos *stepwise*, ou modelos de seleção baseados em critérios de informação; por exemplo, critérios de informação de Bayes (ROYSTON; SAUERBREI, 2003).

Santos *et al.* (2005) desenvolveram em seu trabalho, um sistema para predição da soroprevalência da Hepatite A (doença do fígado) visando o apoio ao diagnóstico. Para isto, os autores consideraram os modelos de Regressão Logística (RL) e Redes Neurais Artificiais (RNA). Os autores acreditam que o sistema contribuirá na identificação de indivíduos com alto risco de contrair a doença, atenuando o risco de disseminação da doença para a população. Para identificar as variáveis relevantes ao problema em estudo, os autores construíram um modelo de RL. Após a identificação das variáveis relevantes, foi implementada uma rede neural artificial para processar as variáveis escolhidas, a partir de um treinamento supervisionado. A seleção das variáveis no modelo logístico foi realizada utilizando o método *stepwise*, considerando 10% e 20% os níveis de significância para inclusão e exclusão de variáveis, respectivamente. O modelo neural proposto se apoiou nas variáveis selecionadas para inclusão no modelo de RL.

Costa *et al*. (2007) tiveram como objetivo em seu trabalho examinar a prevalência da depressão pós-parto e as circunstâncias suscetíveis de predizer a sintomatologia depressiva, em 1 semana e em 3 meses após o parto. Segundo os autores, as consequências adversas da

depressão pós-parto justificam os estudos que procuram identificar os fatores de risco associados à patologia. Os preditores do grau de sintomatologia depressiva nos períodos investigados foram explorados através de análise de Regressão Múltipla Linear - RML, pelo método *stepwise*, considerando separadamente as características sociais e demográficas, variáveis de saúde física e psicológica, de antecipação do parto, do tipo de parto e do bebê, analgesia peridural, contato com o bebê e de experiência de parto. Os autores concluíram no trabalho que realizaram que a regressão múltipla pelo método *stepwise* não identificou nenhuma das variáveis investigadas como significativa na predição da sintomatologia depressiva pós-parto.

Com o objetivo de desenvolver um modelo de predição no diagnóstico de certas doenças, como alguns tipos de câncer, Weber *et al*. (2004) utilizaram o modelo de Regressão Logística (RL). O método *stepwise* e *backward* foi utilizado para selecionar as variáveis (genes que podem ser descritos como causas para certas doenças) no modelo de regressão. Os autores utilizaram dois bancos de dados que foram publicados em Khan *et al*. (2001) e Golub *et al*. (1999) na análise. Royston e Sauerbrei (2003) também utilizaram o método *backward* para selecionar as variáveis utilizadas no modelo de predição no diagnóstico de câncer de mama, utilizando o modelo de Regressão Logística (RL).

Looy *et al.* (2007) propuseram uma abordagem para predizer a concentração do medicamento *Tacrolimus* no sangue de pacientes com fígado transplantado. Segundo os autores, é importante monitorar a concentração do medicamento para avaliar o quadro clínico do paciente. Modelos de previsão para a concentração sanguínea do medicamento podem melhorar o atendimento clínico por meio do monitoramento destas concentrações, especialmente na fase inicial após o transplante, quando o paciente encontra-se na unidade de terapia intensiva (UTI). Os métodos Regressão por Suporte Vetorial - RSV (*Support Vector Regression*), Função de Base Radial - FBR (*Radial Basis Function*) e Regressão Múltipla Linear – RML (*Multiple Linear Regression*) foram utilizados depois da seleção das variáveis de entrada clinicamente relevantes para o modelo. Para cada método (RSV, FBR RSV e RML), foram utilizadas as seleções *backward*, *forward* e *stepwise* para definir as variáveis do modelo. A acurácia dos três métodos foi comparada depois da realização de uma validação cruzada (VC). A predição da concentração do medicamento no sangue com os modelos linear e não linear do RSV foi considerada satisfatória pelos autores, com desempenho melhor em

comparação com o modelo RLM. Os autores utilizaram na análise o banco de dados da Unidade de Cuidados Intensivos (*Intensive Care Unit database*) do Hospital Universitário Ghent (*Ghent University Hospital*) em Ghent na Bélgica.

A abordagem bayesiana também é utilizada na seleção de variáveis. Segundo essa abordagem, todos os parâmetros do modelo são considerados como variáveis aleatórias (PEREIRA *et al*., 2009). Erkanli *et al*. (2006) utilizaram a abordagem Bayesiana para a seleção de variáveis na detecção precoce de câncer de ovário, para determinar o melhor modelo de predição considerando modelos de Regressão Logística. Os autores utilizaram a abordagem bayesiana de simulações designada Cadeias de Markov de Monte Carlo - CMMC (*Monte Carlo Markov Chain*) para determinar as variáveis relevantes no modelo preditivo.

A eliminação recursiva via *bootstrap* (ERB) também é utilizada como método de seleção de variáveis. O *bootstrap* é uma técnica estatística que permite a avaliação da variabilidade de estatísticas, com base nos dados de uma única amostra existente. Operacionalmente, o procedimento *bootstrap* consiste na reamostragem de mesmo tamanho e com reposição dos dados da amostra original, e cálculo da estatística de interesse para cada reamostra (LAVORANTI, 2003). Roca-Pardinas *et al*.(2009) utilizaram a ERB para a seleção de variáveis na detecção precoce do câncer de mama, para determinar o melhor modelo de predição considerando modelos de regressão.

2.3.2 Seleção de variáveis para classificação

O objetivo da seleção de variáveis para classificação é categorizar as observações em classes em relação a alguma especificação. A maioria das abordagens de seleção de variáveis para classificação apoia-se em métodos de mineração de dados (ANZANELLO; FOGLIATTO, 2011). Selecionar o número ideal de variáveis para utilizar na classificação é uma tarefa complicada (DÍAZ-URIARTE; ANDRÉS, 2006); por isso, se torna importante encontrar o método adequado para determinar as variáveis relevantes dentro de um grupo de variáveis (WEBER *et al*., 2004).

O método de seleção de variáveis *LASSO* (*Least Absolute Shrinkage and Selection Operator* ou Operador de Menor Encolhimento Absoluto e Seleção; Tibshirani, 1996) é uma técnica utilizada em bancos de dados multicolineares, que promove a otimização das

estimativas obtidas pelo método dos mínimos quadrados e reduz a dimensionalidade dos parâmetros de entrada (PAOLUCCI, 2006). É baseado na máxima verossimilhança parcial penalizada, que faz com que alguns coeficientes da solução sejam zero, facilitando a interpretação do modelo (PARAÍBA, 2009; ROCA-PARDINAS *et al*., 2009). A utilização do LASSO pode promover a otimização dos coeficientes de regressão e reduzir a magnitude e o número de coeficientes necessários para o processo de classificação.

Com o objetivo de maximizar a acurácia de classificação no diagnóstico de câncer de próstata, Ghosh e Chinnaiyan (2005) desenvolveram regras de classificação baseada na consideração de medidas de acurácia de diagnóstico. A solução encontrada pelos autores foi combinar os problemas de seleção de variáveis e classificação, sugerindo uma abordagem para classificação baseada no método LASSO. Uma vantagem deste método, como descrito anteriormente, é que o efeito de algumas variáveis é anulado, de forma a facilitar a interpretação das variáveis remanescentes. Ghosh e Chinnaiyan (2005) descreveram a aplicação da abordagem proposta em dados simulados e em dados de um estudo recente sobre câncer divulgado por Dhanasekaran *et al.* (2001), onde o foco era a classificação de tumores (tecido com câncer *versus* tecidos sem câncer). Boulesteix e Strobl (2009) também utilizaram o método LASSO com o objetivo de maximizar a acurácia de classificação no diagnóstico de câncer do cólon e de próstata.

Random Forest (Floresta Randômica - FR) é um algoritmo para classificação que utiliza um conjunto de árvores de classificação. Cada árvore é construída usando uma amostra *bootstrap* dos dados, e em cada divisão o grupo de variáveis candidatas é um subgrupo aleatório do conjunto total de variáveis. Díaz-Uriarte e Andrés (2006) utilizaram o método *Random Forest* para selecionar variáveis com vistas à classificação de sobreviventes de câncer de bexiga.

O método *Network Prunning* (Purificação da rede) consiste em remover as variáveis menos importantes (Pastor-Bárcenas *et al.*,2005). Com o objetivo de maximizar a acurácia de classificação no diagnóstico do refluxo vesicoureteral (desordem anatômica e funcional que causa infecção renal), Mantzaris *et al*. (2011) introduziu a purificação da rede neural probabilística (*Probabilistic Neural Networks* - PNNs) utilizando um algoritmo genérico (*Genetic Algorithm* - GA) buscando um subgrupo ótimo de variáveis de entrada que a PNNs utiliza. Segundo Mantzaris *et al*. (2011), o grupo de fatores de diagnóstico considerados deve

se mínimo tal que forneça a informação necessária para que o médico alcance uma decisão médica com maior credibilidade.

F-Score (medida F) é um método que mede a distinção de dois grupos de números reais. A importância de cada variável é medida por *F-Score*. Quanto maior o *F-Score* for, mais discriminativa é a variável (RANKA *et al*., 2009; CHEN; LIN, 2006; CHEN *et al*., 2012; XIE; WANG, 2011). O método F-*Score* é um critério estatístico que faz o ranking de cada variável (CHANG *et al*., 2010).

Com o objetivo de diminuir a taxa de erro no diagnóstico de câncer de tireoide, Chang *et al*. (2010) apresentou um sistema de classificação de nódulos da tireoide utilizando Máquinas de Suporte Vetorial - MSV (*Support Vector Machine*). Para reduzir o tempo computacional necessário e para melhorar a acurácia da classificação, Chang *et al*. (2010) utilizou o método *F-Score* para fazer a seleção de variáveis, reduzindo assim o número de variáveis.

Chen *et al*. (2012) propõe uma técnica de inteligência de enxames (*Swarm Intelligence Technique*) para classificação baseada em MSV para o diagnóstico do câncer de mama. O objetivo foi garantir que os médicos pudessem fazer um diagnóstico preciso do câncer de mama, distinguindo tumores malignos dos benignos. Chen *et al*. (2012) utilizou o método *F-Score* para fazer a seleção de variáveis, com o propósito de identificar as variáveis mais relevantes e eliminar as variáveis menos relevantes a serem utilizadas na construção do modelo de classificação. A eficácia do método proposto foi avaliada utilizando o banco de dados WBCD.

Xie e Wang (2011) desenvolveram um modelo de diagnóstico baseado em MSV para o diagnóstico de dermatoses, a partir do banco de dados da UCI (*University of California at Irvine*), e utilizaram o F-*Score* como método para fazer a seleção de variáveis. Através do *F*-Score, Xie e Wang (2011) mediram a importância de cada variável a fim de encontrar o melhor subgrupo de variáveis (subgrupo que contenha o número mínimo de variáveis além da acurácia de classificação mais alta).

Com o objetivo de classificar bancos de dados médicos, Polat e Güneş (2009) desenvolveram o método *Kernel F-Score Feature Selection* (KFFS) para selecionar as variáveis a serem incluídas no modelo. Primeiramente o espaço das variáveis de entrada foi

mapeado utilizando as Funções de *Kernel* de Base Radial (*Radial Basis Function* - RBF) ou Linear (*Lin*), transformando assim um banco de dados não linearmente separável em um banco de dados linearmente separável. Em seguida os valores de *F-Score* foram calculados. Após esse cálculo, o valor médio dos *F-Score* calculados foram computados. Se o valor do *F-Score* de qualquer variável no banco de dados é maior que esse valor médio, essa variável será selecionada. Caso contrário, essa variável é removida. O banco de dados utilizado foi o banco de dados de doenças do coração da UCI (*University California at Irvine*). Para fazer a classificação Polat e Güneş (2009) utilizaram Máquinas de Suporte Vetorial por Mínimos Quadrados – MSV-MQ (*Least Square Support Vector Machine*) e Redes Neurais Artificiais com Levenberg–Marquardt – RNA-LM (Levenberg–Marquardt *Artificial Neural Network*).

Jaganathan *et al*. (2011) aplicou o método de seleção de variáveis KFFS no diagnóstico de câncer de mama com o objetivo de melhorar a acurácia no diagnóstico médico. As variáveis selecionadas pelo método foram selecionadas para classificação, e utilizadas na classificação de casos benignos e malignos do WBCD utilizando MSV.

Akay (2009) propôs em seu trabalho um método baseado em MSV para melhorar a acurácia de classificação do diagnóstico de câncer de mama em duas classes: benignos e malignos. O cálculo do *F-Score* foi utilizado para selecionar as variáveis a serem incluídas no modelo. A importância de cada variável foi medida pelo cálculo do *F-Score*. Para avaliar o método proposto, foi conduzido experimentos no WBCD.

A tabela 2.2 apresenta os principais métodos de seleção de variáveis para diagnósticos médicos e seus respectivos autores.

Tabela 2.2 - Principais métodos de seleção de variáveis para diagnósticos médicos e seus respectivos autores

	Método		
Autores	Predição	Classificação	Seleção de variáveis
Santos *et al*.(2005)	RL,RNA	-	*Stepwise*
Costa *et al*.(2007)	RML	-	
Weber *et al*.(2004)	RL	-	*Stepwise,Backward*
Looy *et al*.(2007)	RSV,FBR RSV,RML	-	*Stepwise,Backward,Forward*
Erkanli *et al*.(2006)	RL	-	*Bayesian*
Roca-Pardinas *et al*.(2009)	RL	-	*ERB*
Royston e Sauerbrei(2003)	RL	-	*Backward*
Ghosh e Chinnaiyan(2004)	-	MSV,LASSO	LASSO
Boulesteix e Strobl(2009)	-	FR,MSV,VC	LASSO
Díaz-Uriarte e Andrés(2006)	-	FR	FR
Mantzaris *et al*.(2011)	-	PNN	*Network Pruning*
Akay(2009)	-	MSV	*F-Score*
Chang *et al*.(2010)	-		
Chen *et al*.(2012)	-		
Xie e Wang(2011)	-		
Polat e Güneş(2009)	-	MSV-MQ,RNA-LM	*Kernel* F-*Score*
Jaganathan *et al*.(2011)	-	MSV	

Fonte: elaborado pelos autores.

2.3.2.1. Critérios de desempenho de classificação

A discussão a seguir está restrita a problemas de classificação para fins de diagnóstico em duas classes: positivos e negativos. O desempenho dos métodos de classificação é avaliado utilizando a matriz de confusão ou confundimento (do inglês, *confusion matrix*), por medidas como acurácia, sensibilidade e especificidade, e através de gráficos da curva ROC (análise ROC - *Receiver Operating Characteristic)*. Quatro possibilidades de classificação de observações permitem definir formalmente os critérios descritos acima; são elas: Positivos Verdadeiros (PV), Negativos Verdadeiros (NV), Positivos Falsos (PF) e Negativos Falsos (NF) (ANZANELLO; FOGLIATTO, 2011; PRATI *et al*., 2008).

A matriz de confusão relaciona as classificações reais (correspondentes às linhas da matriz) e preditas (correspondentes às colunas da matriz) realizadas por um sistema de classificação. O desempenho desse sistema é avaliado utilizando os dados na matriz. Se todas as amostras são corretamente classificadas, a matriz de confusão correspondente somente terá

elementos não-nulos na diagonal principal (MARCANO-CEDEÑO *et al*., 2011; POLAT; GÜNEŞ, 2007; NETO *et al*., 2006; NETO, 2006). A Tabela 2.3 exemplifica uma matriz de confusão para o caso de duas classes.

Tabela 2.3 - Representação da matriz de confusão

Real	Predito	
	Positivo	Negativo
Positivo	PV	PF
Negativo	NF	NV

Fonte: elaborado pelos autores.

Onde PV é dado pelo total de classificações corretas de observações conformes, PF é dado pelo total de classificações erradas de observações não-conformes, NF é dado pelo total de classificações erradas de observações conformes e NV é dado pelo total de classificações corretas de observações não-conformes. Observações conformes são aquelas que apresentam as características esperadas ou que atendem a especificações pré-definidas. Por exemplo, em um contexto de classificação de observações oriundas de pacientes de câncer, pode-se definir uma observação conforme como correspondente a um paciente que possui a doença. O indicador PV, nesse caso, corresponderia ao número de vezes em que indivíduos com a doença foram corretamente classificados como tal.

A acurácia pode ser definida como a fração de observações corretamente classificadas. A sensibilidade está relacionada à fração de observações *conformes* corretamente classificadas. A especificidade está relacionada à fração de observações *não-conformes* corretamente classificadas (ANZANELLO; FOGLIATTO, 2011).

A acurácia, sensibilidade e especificidade são dadas pelas equações (1), (2) e (3) respectivamente.

$$Acurácia = \frac{PV + NV}{PV + NV + PF + NF}. \quad (1)$$

$$Sensibilidade = \frac{PV}{PV + NF}. \quad (2)$$

$$Especificidade = \frac{NV}{NV + PF}. \quad (3)$$

Os elementos nas equações (1) a (3) seguem as definições apresentadas na matriz de confusão.

A análise ROC (*Receiver Operating Characteristic*) é um método gráfico para avaliação do desempenho da classificação. A curva ROC é uma medida de desempenho em duas dimensões, muito utilizada em pesquisas biomédicas para avaliar o desempenho dos testes diagnósticos. Os gráficos ROC apresentam os percentuais de Positivos Falsos no eixo das abcissas e os percentuais de Positivos Verdadeiros no eixo das coordenadas (MARCANO-CEDEÑO *et al*., 2011; POLAT; GÜNEŞ, 2007).

A Figura 2.1 representa o espaço ROC. Na linha diagonal ascendente (0,0) – (100%, 100%), cada ponto (*p*,*p*) pode ser obtido pela previsão da classe positiva com probabilidade *p* e da classe negativa com probabilidade 100% - *p*. A diagonal descendente (0, 100%) – (100%,0) representa modelos de classificação que desempenham igualmente em ambas as classes. À esquerda dessa linha estão os modelos que apresentam melhor desempenho para a classe negativa em detrimento da positiva e, à direita, os modelos que apresentam melhor desempenho para a classe positiva (PRATI *et al*., 2008).

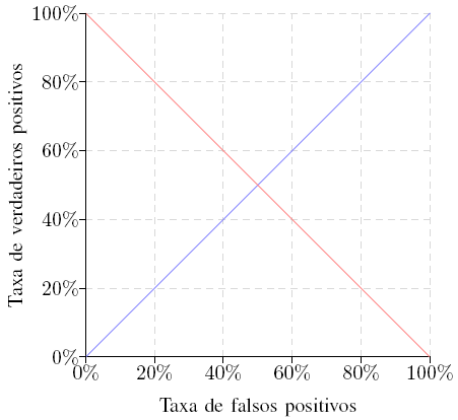


Figura 2.1 - Espaço ROC. Fonte: PRATI *et al*., 2008.

Devido ao fato das coordenadas deste gráfico representarem medidas de probabilidade, estas variam entre 0 e 100%. A área sob a curva ROC (*Area Under the ROC Curve* – AUC) é uma medida do desempenho da acurácia da classificação. Como a AUC é uma fração da área de um quadrado unitário, o seu valor sempre irá satisfazer a seguinte condição (MARCANO-CEDEÑO *et al*., 2011; MARTINEZ *et al*., 2003; PRATI *et al*., 2008):

$$0 \leq AUC \leq 1 \tag{4}$$

Quanto maior a área sob a curva, maior a acurácia do teste (NITRINI, *et al.*,1994). Analiticamente, a AUC pode ser determinada através de métodos de resolução numérica (regra do trapézio), métodos estatísticos (relação com a estatística de Wilcoxon-Mann-Witney) ou através de estimativa de máxima verossimilhança (MARTINEZ *et al*., 2003).

Na Figura 2.2 é apresentado um exemplo de uma Curva ROC, com o eixo das abscissas representando a sensibilidade e o eixo das ordenadas representando (1 – especificidade). A área abaixo da diagonal tracejada representa um classificador sem poder de discriminação de padrões (SOVIERZOSKI *et al*., 2011).

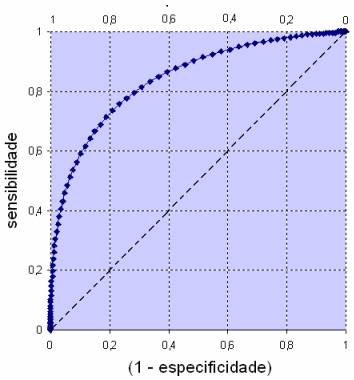


Figura 2.2 - Curva ROC. Fonte: Sovierzoski *et al*., 2011.

2.4 Conclusões

Bancos de dados, como os oriundos de diagnóstico médico, são usualmente constituídos por centenas de variáveis correlacionadas e ruidosas, o que compromete o desempenho de várias técnicas estatísticas. A eliminação de variáveis que não contribuem

para uma conclusão ou dedução é importante, pois variáveis ruidosas tendem a produzir um modelo preditivo errôneo ou incompleto (ANZANELLO, 2009).

Neste artigo foi feita uma revisão sistemática da literatura sobre métodos de seleção de variáveis para predição e classificação de diagnósticos médicos. Nas abordagens de seleção de variáveis para predição foram utilizados testes de significância para os coeficientes e as variáveis eram selecionadas utilizando os métodos *Forward, Backward, Stepwise, Baysesian* e *Recursive Bootstrap Elimination* para determinar o melhor modelo. Nas abordagens de seleção de variáveis para classificação o acerto na classificação foi o critério para inserir a variável no modelo e a seleção de variáveis foi realizada utilizando os métodos *Network Prunning*, *F-Score*, *LASSO* e *Random Forest.*

Na literatura visitada, é possível identificar citações a métodos de seleção de variáveis para posterior classificação e/ou predição para algum diagnóstico, porém na maior parte dos casos, os autores não explicam como essa seleção de variáveis foi feita. Desenvolvimentos futuros incluem uma busca de um método mais robusto para a seleção de variáveis para fins de classificação para o diagnóstico.

2.5 Referências

ABBASS, H. A. An evolutionary artificial neural networks approach for breast cancer diagnosis. **Artificial Intelligence in Medicine**, v. 25, p. 265-281, 2002.

AKAY, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, p. 3240–3247, 2009.

ANZANELLO, M. J. Seleção de Variáveis com vistas à Classificação de Bateladas de Produção em duas Classes. **Gestão da Produção**, São Carlos, v. 16, n. 4, p. 526-533, out.-dez. 2009.

ANZANELLO, M. J.; FOGLIATTO, F. S . A Review of Recent Variable Selection Methods in Chemometrics and Manufacturing Applications. Submetido à Revista Pesquisa Operacional.

ANZANELLO, M. J.; FOGLIATTO, F. S . Identificação de Variáveis Relevantes para Categorização de Bateladas de Produção com Base em Critérios de Sensibilidade E especificidade. In: XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011, Ubatuba. XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011.

ARAÚJO-FILHO, J. L. S.; MELO-JÚNIOR, M. R.; LINS, C. A. B.; LINS, R. A. B.; MACHADO, M. C. F. P.; CARVALHO-JR, L. B.; FILHO N. T. P. Galectina-3 em Tumores de Próstata: Imuno-Histoquímica e Análise Digital de Imagens. **Jornal Brasileiro de Patologia e Medicina Laboratorial**, v. 42, n. 6, p. 469-475, dezembro 2006.

BOULESTEIX, A. L.; STROBL, C. Optimal Classifier Selection and Negative Bias in Error Rate Estimation: An Empirical Study on High-Dimensional Prediction. **BMC Medical Research Methodology**, v. 9, p. 85, 2009.

BREIMAN, L. Random forests. **Machine learning**, v. 45(1), p. 5-32, 2001.

CHANG, C.-Y..; CHEN, S.-J.; TSAI, M.-T. Application of Support-Vector-Machine-Based Method for Feature Selection and Classification of Thyroid Nodules in Ultrasound Images. **Pattern Recognition**, v. 43, p. 3494–3506, 2010.

CHEN, Y.-W.; LIN, C.-J. Combining SVMs with Various Feature Selection Strategies. **Studies in Fuzziness and Soft Computing**, v. 207, p. 315-324, 2006.

CHEN, H.-L.; YANG, B.; WANG, G.; WANG, S.-J.; LIU, J.; LIU, D.-Y. Support Vector Machine Based Diagnostic System for Breast Cancer Using Swarm Intelligence. **Journal of Medical Systems**, v. 36, I. 4, p. 2505-2519, 2012.

COSTA, R.; PACHECO, A.; FIGUEIREDO, B. Prevalência e preditores de sintomatologia depressiva após o parto. **Revista de Psiquiatria Clínica**, v. 34 (4), p. 157-165, 2007.

DÍAZ-URIARTE, R.; ANDRÉS, S. A. Gene Selection and Classification of Microarray Data Using Random Forest. **BMC Bioinformatics**, v. 7, p. 3, 2006.

DHANASEKARAN, S. M.; BARRETTE, T. R.; GHOSH, D.; SHAH, R.; VARAMBALLY, S.; KURACHI, K.; PIENTA, K. J.; RUBIN, M. A.; CHINNAIYAN, A. M. Delineation of prognostic biomarkers in prostate cancer. Nature, v. 412(6849), p. 822–826, 2001.

EFRON, B.; TIBSHIRANI, R.J. **An Introduction to the Bootstrap**. New York: Chapman & Hall, 1993. 436p.

ERKANLI, A.; TAYLOR, Douglas D.; DEAN, D.; EKSIR, F.; EGGER, D.; GEYER, J.; NELSON, B. H.; STONE, B.; FRITSCHE, H. A.; RODEN, R. B. S. Application of Bayesian Modeling of Autologous Antibody Responses against Ovarian Tumor-Associated Antigens to Cancer Detection. **Cancer Research**, v. 66, p. 1792-1798, 2006.

GHOSH, D.; CHINNAIYAN, A. M. Classification and Selection of Biomarkers in Genomic Data Using LASSO. **Journal of Biomedicine and Biotechnology**, v. 2, p. 147–154, 2005.

GOLUB, T. R.; SLONIM, D. K.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J. P.; COLLER, H.; LOH, M. L.; DOWNING, J. R.; CALIGIURI, M. A.; BLOOMÞELD, C. D.; LANDER, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, v. 286(5439), p. 531-537, 1999.

GUO, P.; TANAKA, H. Dual Models for Possibilistic Regression Analysis. **Computational Statistics & Data Analysis**, v. 51, p. 253 – 266, 2006.

JAGANATHAN, P.; RAJKUMAR, N.; NAGALAKSHMI, R. A Kernel Based Feature Selection Method Used in the Diagnosis of Wisconsin Breast Cancer Dataset. **Communications in Computer and Information Science**, v. 190, Part 8, p. 683-690, 2011.

KHAN, J.; WEI, J. S.; RINGNÉR, M.; SAAL, L. H.; LADANYI, M.; WESTERMANN, F.; BERTHOLD, F.; SCHWAB, M.; ANTONESCU, C. R.; PETERSON, C.; MELTZER, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, v. 7(6), p. 673-679, 2001.

LAVORANTI, O. J. Estabilidade e Adaptabilidade Fenotípica Através da Reamostragem “Bootstrap" no Modelo AMMI. Tese de Doutorado, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo. PIRACICABA, São Paulo – Brasil. Julho – 2003.

LOOY, S. V.; VERPLANCKE, T.; BENOIT, D.; HOSTE, E.; MAELE, G. V.; TURCK, F. D.; DECRUYENAERE, J. A Novel Approach for Prediction of Tacrolimus Blood Concentration in Liver Transplantation Patients in the Intensive Care Unit Through Support Vector Regression. **Critical Care**, v. 11, no 4, 2007.

MANTZARIS, D.; ANASTASSOPOULOS, G.; ADAMOPOULOS, A. Genetic Algorithm Pruning of Probabilistic Neural Networks in Medical Disease Estimation. **Neural Networks**, v. 24, p. 831–835, 2011.

MARCANO-CEDEÑO, A; QUINTANILLA-DOMÍNGUEZ, J.; ANDINA, D. WBCD Breast Cancer Database Classification Applying Artificial Metaplasticity Neural Network. **Expert Systems with Applications**, v. 38, p. 9573–9579, 2011.

MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, B. B. A Curva ROC para Testes Diagnósticos. **Cadernos Saúde Coletiva,** Rio de Janeiro, v. 11 (1), p. 7 – 31, 2003 – 7.

NETO, A. R. R. SINPATCO - Sistema Inteligente para Diagnóstico de Patologias da Coluna Vertebral. Fortaleza. Dissertação de Mestrado do Programa de Pós- Graduação em Engenharia em Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, 2006.

NETO, A. R.; BARRETO, G. A.; CORTEZ, P. C.; MOTA, H. Classificação de Patologias da Coluna Vertebral Usando Redes Neurais Artificiais. In: X Congresso Brasileiro de Informática em Saúde, 2006, Florianópolis. X Congresso Brasileiro de Informática em Saúde, 2006.

NITRINI, R.; LEFÈRE, B.; MATHIAS, S. C.; CARAMELLI, P.; CARRILHO, P. E. M.; SAUAIA, N.; MASSAD, E.; TAKIGUTI, C.; SILVA, I. O.; PORTO, C. S.; MAGILA,

M. C.; SCAFF, M. Testes Neuropsicológicos de Aplicação Simples para o Diagnóstico de Demência. **Arquivos de Neuro-Psiquiatria**, v.52(4), p. 457-465, 1994.

OLIVEIRA, M. C.; AZEVEDO-MARQUES, P. M.; FILHO, W. C. C. Grades Computacionais na Otimização da Recuperação de Imagens Médicas Baseada em Conteúdo. **Radiologia Brasileira**, v. 40(4), p. 255–261, 2007.

PAOLUCCI, L. A. Comparação de dois Métodos para Representação da Força de Reação do Solo no Desempenho de Classificação de Padrões da Marcha. Belo Horizonte. Dissertação de Mestrado da Escola de Educação Física, Fisioterapia e Terapia Ocupacional, Universidade Federal de Minas Gerais, 2006.

PARAÍBA, C. C. M. Análise de Sobrevivência de Dados *Microarray*: Seleção de Genes Prognósticos Quando *p* é Maior do que *n*. São Carlos. Dissertação de Mestrado do Programa de Pós-Graduação em Estatística da Universidade Federal de São Carlos, 2009.

PASTOR-BÁRCENAS, O.; SORIA-OLIVAS, E.; MARTÍN-GUERRERO, J. D.; CAMPS-VALLS, G.; CARRASCO-RODRÍGUEZ, J. L.; VALLE-TASCÓN, S. Unbiased Sensitivity Analysis and Pruning Techniques in Neural Networks for Surface Ozone Modelling. **Ecological Modelling**, v. 182, i. 2, p. 149–158, 2005.

PEREIRA, J. M.; MUNIZ, J. A.; SÁFADI, T.; SILVA C. A**.** Comparação entre Modelos para Predição do Nitrogênio Mineralizado: uma Abordagem Bayesiana. **Ciência e Agrotecnologia,** Lavras, v. 33, Edição Especial, p. 1792 -1797, 2009.

POLAT, K.; GÜNEŞ, S. Breast cancer diagnosis using least square support vector machine. **Digital Signal Processing**, v. 17, p. 694-701, 2007.

POLAT, K.; GÜNEŞ, S**.** A New Feature Selection Method on Classification of Medical Datasets: Kernel F-Score Feature Selection. **Expert Systems with Applications**, v. 36, p. 10367–10373, 2009.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Evaluating Classifiers Using ROC Curves**. IEEE Latin America Transactions**, v.6, i.2, June 2008.

RANKA, S.; ALURU, S.; BUYYA, R.; CHUNG, Y.-C.; DUA, S.; GRAMA, A.; GUPTA, S. K. S.; KUMAR, R.; PHOHA, V. V. Enhancing the Performance of LibSVM Classifier by Kernel F-Score Feature Selection. **Contemporary Computing, Communications in Computer and Information Science**, v. 40, pp. 533–543, 2009.

ROCA-PARDINAS, J.; CADARSO-SUÁREZ, C.; TAHOCES, P. G.; LADO, M. J. Selecting Variables in Non-Parametric Regression Models for Binary Response. An Application to the Computerized Detection of Breast Cancer. **Statistics in Medicine**, v. 28, p. 240–259, 2009.

ROYSTON, P.; SAUERBREI, W. Stability of Multivariable Fractional Polynomial Models with Selection of Variables and Transformations: A Bootstrap Investigation. **Statistics in Medicine**, v. 22, p. 639–659, 2003.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A Review of Feature Selection Techniques in Bioinformatics. **Bioinformatics**, v. 23 i. 19, p. 2507–2517, 2007.

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

SETIONO, R. Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. **Artificial Intelligence in Medicine**, v. 18, p. 205-219, 2000.

SOVIERZOSKI, M. A.; ARGOUD, F. I. M.; AZEVEDO, F. M. Treinamento de um Classificador Neural Binário: Estimativa da Distribuição de Padrões. Sociedade Brasileira de Informática em Saúde, 2011. Disponível em: www.sbis.org.br/cbis11/arquivos/665.pdf.

TIBSHIRANI, R. Regression Shrinkage and Selection via the LASSO. **Journal of the Royal Statistical Society**, Series B, v. 58(1), p. 267–288, 1996.

XIE, J.; WANG, C. Expert Using Support Vector Machines with a Novel Hybrid Feature Selection Method for Diagnosis of Erythemato-Squamous Diseases. **Systems with Applications**, v. 38, p. 5809–5815, 2011.

WEBER, G.; VINTERBO, S.; OHNO-MACHADO, L. Multivariate Selection of Genetic Markers in Diagnostic Classification. **Artificial Intelligence in Medicine**, v. 31, p. 155—167, 2004.

3 Segundo Artigo: Método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis

Nicole Holsbach

Flávio Sanson Fogliatto

Michel José Anzanello

Resumo

Na maioria dos países, o câncer de mama (CM) entre as mulheres é predominante. Se diagnosticado precocemente, apresenta alta probabilidade de cura. Diversas abordagens baseadas em Estatística foram desenvolvidas para auxiliar na sua detecção precoce. Este artigo apresenta um método para a seleção de variáveis para classificação dos casos em duas classes de resultado, benigno ou maligno, baseado na análise citopatológica de amostras de células da mama de pacientes. As variáveis são ordenadas de acordo com um novo índice de importância de variáveis que combina os pesos de importância da ACP (Análise de Componentes Principais) e a variância explicada a partir de cada componente retido. Observações da amostra de treino são categorizadas em duas classes através dos métodos KVP (*k*-vizinhos mais próximos) e AD (Análise Discriminante), seguida pela eliminação da variável com o menor índice de importância. Usa-se o subconjunto com a máxima acurácia para classificar as observações na amostra de teste (novas observações a serem classificadas). Aplicando ao WBCD (*Wisconsin Breast Cancer Database*), o método proposto apresentou uma média de 97,77% de acurácia de classificação, retendo uma média de 5,8 variáveis de um total de 9 variáveis.

Palavras-chave: Seleção de variáveis, Diagnóstico de câncer de mama, k-vizinhos mais próximos.

Abstract

In the majority of countries, female breast cancer (BC) is predominant. If diagnosed in early stages, it presents a high probability of cure. Several statistical-based approaches have been developed to assist early BC detection. This paper presents a method for feature selection for the classification of cases into two classes, benign or malignant, based on cytopathologic analysis from patients' breast cell samples. Features are ranked according to a new feature importance index that combines PCA (Principal Component Analysis) weights and the variance explained by each retained component. Observations of a training set are categorized into two classes through the KNN (*k*-nearest neighbor) tool and DA (Discriminant Analysis), followed by elimination of the feature with the smallest importance index. The subset with the maximum accuracy is used to classify observations in the testing set. When applied to the WBCD (Wisconsin Breast Cancer Database), the proposed method led to average 97.77% accurate classifications while retaining an average of 5.8 features.

Keywords: Feature selection, Breast cancer diagnosis, *k*-nearest Neighbor.

3.1 Introdução

Mesmo com os avanços na detecção e tratamento precoce, o câncer está evoluindo para uma condição crônica em muitos países. Nota-se a predominância de três tipos de câncer: o câncer de mama (CM) entre as mulheres, na grande maioria dos países, a nível mundial, o câncer de colo do útero na África e no Sul da Ásia, e o câncer de próstata na América do Norte, Oceania, Norte da Europa e Europa Ocidental (BRAY *et al.*, 2012). O CM é o segundo tipo mais frequente no mundo, atrás somente do câncer de pulmão, e é o mais comum entre as mulheres, respondendo por 22% dos casos novos a cada ano. Se diagnosticado precocemente, identificado em estágios iniciais, quando as lesões são de tamanho inferior a dois centímetros de diâmetro, apresenta uma alta percentagem de cura (IARC, 2002; WHO, 2007; BRASIL, 2004). A doença consiste no crescimento desordenado de células do tecido da mama, formando nódulos que podem ser malignos (tumores) ou benignos. No Brasil, as taxas de mortalidade por câncer de mama continuam elevadas, muito possivelmente porque a doença ainda é diagnosticada em estados avançados. O Brasil gastou R$117.849.636,17 em 2008, R$

129.301.592,94 em 2009 e R$ 148.992.855,26 em 2010 somente com mamografia, representando um crescimento de 15% em 2009 e 16% em 2010 (BRASIL, 2010). Segundo revisões sistemáticas recentes, o impacto do rastreamento mamográfico na redução da mortalidade por câncer de mama pode chegar a 25%. O tratamento varia de acordo com o estadiamento da doença, suas características biológicas e das condições da paciente (idade, status menopausal, comorbidades). As modalidades de tratamento do câncer de mama podem ser divididas em tratamento local (cirurgia e radioterapia) e tratamento sistêmico (quimioterapia, hormonioterapia e terapia biológica) (BRASIL, 2004). O rastreamento do câncer de mama em condições de sucesso (cobertura da população alvo, qualidade dos exames de rastreamento e garantia de acesso ao diagnóstico e tratamento) reduz em 30% a mortalidade do câncer de mama (BRASIL, 2009). Na população mundial, a sobrevida média após cinco anos é de 61%. O CM não é comum antes dos 35 anos, e acima desta faixa etária sua incidência cresce rápida e progressivamente. Estatísticas indicam aumento de sua incidência tanto nos países desenvolvidos quanto nos em desenvolvimento. Segundo a Organização Mundial da Saúde (OMS), nas décadas de 60 e 70 registrou-se um aumento de 10 vezes nas taxas de incidência ajustadas por idade nos Registros de Câncer de Base Populacional de diversos continentes (BRASIL, 2004). O diagnóstico precoce aumenta as taxas de sobrevivência em pacientes com CM, o que tem sido provado ao longo dos anos através de investigação clínica, como nos estudos de Shapiro *et al*. (1982) e Humphrey *et al*. (2002).

O diagnóstico do CM depende da interpretação do médico a partir das informações obtidas dos pacientes através de exames, os quais incluem exame clínico da mama, mamografia e análise de tecido da mama. O exame clínico da mama, apesar de simples, é pouco eficiente na detecção de pequenos tumores (menores que 1 cm) quando comparado a exames de imagem ou laboratoriais (citopatológicos). Baker (1982) demonstrou que, em um grupo de 280.000 mulheres americanas rastreadas clinicamente quanto ao CM, 6% dos pequenos cânceres no grupo de pacientes foram detectados através do exame clínico da mama e 57% através da mamografia. Em termos de análise, o exame clínico da mama fornece dados univariados para interpretação (mais simples), enquanto que exames laboratoriais (citopatológicos) e de imagem produzem dados multivariados, os quais demandam maior processamento de informações.

Abordagens baseadas em métodos de classificação têm sido propostas para auxiliar profissionais de saúde no processamento das informações geradas pelos exames laboratoriais (citopatológico) de CM, como em Street *et al*. (1993) e Fogel *et al*. (1995). Tais abordagens usualmente apoiam-se em dados de exames (geralmente imagens) para chegar a uma conclusão a respeito da observação analisada, seja maligno ou benigno, no caso de nódulos mamários. Dentre os métodos de classificação mais difundidos na literatura, destacam-se redes neurais artificiais e abordagens baseadas em teoria *fuzzy* (ABONYI; SZEIFERT, 2003). As abordagens geradas com base nesses métodos permitem a inserção de observações em classes com base em dados de entrada, levando a avaliações/categorizações mais acuradas.

Neste artigo é apresentado um método para seleção de variáveis oriundas de exames clínicos com vistas à classificação de observações em categorias distintas. A técnica multivariada Análise de Componentes Principais (ACP) é inicialmente aplicada no banco de dados, onde as observações referem-se a pacientes e as variáveis a dados extraídos de exames clínicos. As variáveis são então ordenadas de acordo com um novo índice que combina os pesos gerados pelos componentes principais retidos na ACP com a variância explicada por estes componentes. Na sequência, as observações da porção de treino são categorizadas em duas classes (benigno ou maligno) utilizando dois métodos de classificação: (*i*) a ferramenta de mineração de dados *k*-vizinhos mais próximos (KVP), e (*ii*) análise discriminante (AD). Por fim, calcula-se a acurácia de classificação. Em seguida, a variável com o menor índice de importância é removida e uma nova classificação é realizada utilizando as variáveis remanescentes. Esse processo iterativo de eliminação e classificação é repetido até que reste somente uma variável. Finalmente, o subconjunto de variáveis que leva à máxima acurácia é escolhido e utilizado para classificar as observações do conjunto de teste.

Uma contribuição importante deste trabalho é a integração de uma técnica multivariada (ACP) com dois métodos de classificação: KVP e AD. A ACP é um método conhecido para a redução da dimensionalidade de dados a partir da obtenção de combinações lineares de variáveis altamente correlacionadas (RENCHER, 1995). Outra contribuição do artigo consiste na proposição de um novo índice de importância baseado em parâmetros da ACP, o qual guia a eliminação recursiva de variáveis.

Vários estudos propondo métodos de classificação testam seu desempenho no *Wisconsin Breast Cancer Database* (WBCD), obtido da universidade de Wisconsin e

disponibilizado *online*. Neste banco, nove variáveis foram analisadas em imagens de amostra de células da mama de 699 indivíduos, para os quais o diagnóstico foi elaborado. Estudos relevantes utilizando o WBCD são apresentados na segunda seção.

O restante desse trabalho está organizado como segue. Na segunda seção é apresentado o referencial teórico sobre sistemáticas de classificação aplicadas no WBCD. O método proposto é detalhado na terceira seção. Os resultados obtidos pelo método proposto são apresentados na quarta seção. A conclusão é apresentada na última seção.

3.2 Referencial teórico

Nesta seção é apresentada uma revisão das metodologias propostas para classificação das observações do WBCD. Algumas abordagens incluem sistemáticas de seleção de variáveis, visando aumentar a acurácia dos classificadores. Propostas de sistemas especialistas para o diagnóstico de câncer de mama que não utilizam o WBCD foram revisadas por Eltoukhy *et al.* (2012); abordagens para seleção de variáveis em problemas de classificação foram revisadas por Dash e Liu (1997).

Os classificadores apresentados nesta seção podem ser categorizados conforme o fundamento teórico em que estão baseados: estatística/máquinas de suporte vetorial (E/MSV), árvores de decisão/programação linear (ADD/PL), redes neurais (RN) ou teoria *fuzzy* (TF). As abordagens são apresentadas em ordem cronológica de publicação; os principais resultados de cada abordagem são resumidos na Tabela 3.1, apresentada no final da seção.

Street *et al.* (1993) relatam análises preliminares realizadas no WBCD com o objetivo de organizar o banco de dados. Os autores classificam os casos do WBCD utilizando o método *Multi-surface*, um modelo de programação linear que encontra o melhor grupo em planos separados no espaço das variáveis. Já Fogel *et al.* (1995) propõem um classificador baseado em redes neurais. A seleção das variáveis é realizada nos experimentos de redes, porém nenhum resultado é explicitado. A acurácia média em uma divisão de 60% das observações em porção de treino e 40% em porção de teste é de 98,05%. Com propósitos semelhantes, Quinlan (1996) sugere um classificador baseado em árvore de decisão, que melhora o desempenho do classificador C4.5 em Quinlan (1993) de duas maneiras: o novo classificador elimina o viés que favorecia variáveis contínuas e que podia levar a testes de

decisão baseados em variáveis irrelevantes; na sequência, os testes de decisão são avaliados utilizando o critério de razão de ganho (ganho de informação / informação da divisão). A seleção de variáveis é realizada através da análise das árvores de decisão. A acurácia de classificação aplicando o método proposto no WBCD é de 94,74%, utilizando 90% das observações na porção de treino.

Também baseado em RN, Setiono (1996) apresenta uma abordagem cujo foco está na geração de regras de classificação no treinamento da rede. Para isso, as saídas da rede são avaliadas utilizando a função de entropia, sendo definido um termo de penalização para medir a perda de acurácia devida a eliminação de variáveis. O erro máximo da classificação é definido pelo usuário, e a melhor rede é encontrada minimizando o termo de penalização. Testes no WBCD apresentaram uma média de acurácia de 96,58%. O mesmo algoritmo foi expandido em Setiono (2000) para incluir um estágio de pré-processamento do classificador da rede. O estágio adicional é realizado em dois passos. No primeiro passo, os casos com valores desconhecidos são removidos do banco de dados. No segundo passo, a rede neural com apenas uma unidade oculta é treinada para uma melhor acurácia na porção de treino, indicando o menor grupo de variáveis a ser usado no classificador. A maior acurácia (96,7%) é obtida quando a rede é treinada para 98% de acurácia na porção de treino, utilizando 50% das observações na porção de treino.

Peña-Reyes e Sipper (1999) combinaram sistemas *fuzzy* e algoritmos evolucionários em uma ferramenta de diagnóstico. O método é dividido em dois passos. Primeiramente, um sistema *fuzzy* pontua casos no WBCD conforme a sua malignidade, baseado nos valores das variáveis. Em seguida, um sistema limítrofe interpreta as saídas do sistema *fuzzy* para a classificação dos casos em benignos e malignos. O método proposto obteve uma acurácia de classificação de 97,8%, utilizando uma divisão de 75% /25% no banco de dados. Nauck e Kruse (1999) propõem um classificador *neuro-fuzzy* utilizando técnicas de aprendizado da teoria de redes neurais. Cinco técnicas de treino são propostas para aumentar o grupo de regras *fuzzy* utilizadas na classificação. Uma delas é baseada na determinação da correlação das variáveis de uma observação com a classe em que está inserida e exclusão das variáveis com valores menores do que valores limítrofes especificados. Tal sistemática obteve uma acurácia de 95,06% em bancos com 90% de observações na porção de treino, além de excluir as variáveis 1 e 9 do WBCD. Lee *et al*. (2001) também propõem um classificador *fuzzy* com

seleção de variáveis: o classificador gera regiões de decisão *fuzzy* que não se sobrepõem, reduzindo o esforço computacional e a complexidade da classificação. Para a seleção de variáveis, eles propõem uma medida de entropia *fuzzy* baseada na entropia de Shannon (1948). O classificador alcança uma acurácia de 94,67% quando todas as variáveis são incluídas, e 95,14% quando apenas 6 variáveis são retidas, valendo-se de uma divisão 50% /50%.

Albrecht *et al*. (2002) propõem uma sistemática de classificação baseada no algoritmo *Perceptron*. A fim de encontrar uma função linear limítrofe que garanta um bom desempenho de classificação, o método *Simulated Annealing* é utilizado na otimização. Um procedimento de seleção de variáveis baseado no ordenamento das variáveis de acordo com o valor do coeficiente gerado pelo algoritmo *Perceptron* também é proposto, apesar de não ser testado no WBCD. A acurácia de classificação no WBCD é de 98,80%.

Abbass (2002) apresenta um classificador baseado na rede neural artificial *Memetic Pareto* com vistas à redução do esforço computacional imposto pelo treinamento das redes neurais. A proposta foi testada no WBCD utilizando 400 indivíduos como porção de treino: os autores obtiveram acurácia média de 98,1% em 120 rodadas.

Verikas e Bacauskiene (2002) propõem um classificador baseado em redes neurais no qual uma função de custo do erro de entropia cruzada é adicionada de um termo que restringe as derivadas das funções de transferência das saídas da rede e dos nodos ocultos. A seleção de variáveis é realizada monitorando o erro de classificação em bases de dados de validação cruzada (VC), a medida que variáveis são removidas; o objetivo é encontrar a melhor solução de compromisso entre erro e número de variáveis retidas. Os melhores resultados na classificação são obtidos usando uma divisão 50% /50% do WBCD: 95,77% de acurácia usando duas variáveis. Retendo as 9 variáveis, a acurácia aumenta para 96,44%. O objetivo é alcançar 100% de acerto na classificação.

Abonyi e Szeifert (2003) apresentam um classificador baseado na regra *fuzzy* com as seguintes características: a regra pode representar mais de uma classe, ao contrário dos classificadores tradicionais *fuzzy*, e um novo protótipo de *cluster* (e algoritmo de clusterização associado) é apresentado, permitindo a identificação direta supervisionada dos classificadores *fuzzy*. Para a seleção de variáveis, uma modificação da função de separação de *Fisher* é

apresentada, na qual a importância das variáveis é estimada com base em sua matriz de covariâncias. A acurácia média encontrada foi de 95,57%, em uma divisão 50% /50% do WBCD.

Polat e Günes (2007) apresentam um classificador de máquina de suporte vetorial no qual um grupo de equações lineares é utilizado para treino. Nenhuma seleção de variáveis é realizada. A maior acurácia de classificação encontrada foi de 98,53% em uma divisão 50% /50% do WBCD. Akay (2009) também propõe um classificador baseado em máquina de suporte vetorial. A seleção de variáveis é o primeiro passo na metodologia proposta, realizada através do *F-score* de Chen e Lin (2006), um índice que mede a discriminação entre dois grupos de números. Todos os índices derivados da classificação da matriz de confusão (*confusion matrix*) são utilizados para avaliar o desempenho do classificador, além das curvas ROC. Os melhores resultados são obtidos utilizando uma divisão 80% /20% do WBCD, com uma acurácia de 99,51%, utilizando 5 das 9 variáveis do banco de dados.

Por fim, Marcano-Cedeño *et al*. (2011) propõem um classificador baseado em redes neurais, que simula a propriedade biológica de metaplasticidade em um algoritmo *perceptron* de múltiplas camadas com propagação reversa. A metaplasticidade pode ser definida como a indução de mudanças sinápticas também dependentes de atividade sináptica prévia. 60% das observações do WBCD foram usadas na porção de treino e 100 experimentos, com diferentes parâmetros de rede, foram rodados, com 100 repetições cada. A melhor acurácia de classificação encontrada na literatura foi de 99,26%.

Tabela 3.1 - Acurácia de classificação obtida no WBCD em diferentes métodos disponíveis na literatura

Fonte	Método	Acurácia (%)
Street *et al*. (1993)	ADD/PL	97,30
Fogel *et al*. (1995)	RN	98,05
Quinlan (1996)	ADD/PL	94,74
Setiono (1996)	RN	96,58
Setiono (2000)	RN	96,70
Peña-Reyes e Sipper (1999)	TF	97,80
Nauck e Kruse (1999)	RN e TF	95,06
Lee *et al*. (2001)	TF	95,14
Albrecht *et al.* (2002)	ADD/PL	98,80
Abbass (2002)	RN	98,10
Verikas e Bacauskiene (2002)	RN	96,44
Abonyi e Szeifert (2003)	TF	95,57
Polat e Günes (2007)	E/MSV	98,53
Akay (2009)	E/MSV	99,51
Marcano-Cedeño *et al*. (2011)	RN	99,26

Fonte: elaborado pelos autores.

3.3 Método

O método de seleção de variáveis para categorização das observações do WBCD em duas classes baseia-se em 4 passos operacionais: (*i*) dividir o banco de dados original em porções de treino e de teste, e aplicar a ACP na porção de treino; (*ii*) gerar índices de importância das variáveis baseados nos pesos da ACP e na percentagem da variância explicada pelos componentes retidos; (*iii*) classificar o banco dos dados utilizando KVP e AD separadamente. Em seguida eliminar a variável com o menor índice de importância, classificar o banco de dados novamente, e calcular a acurácia de classificação. Continuar tal processo iterativo até restar uma variável; e (*iv*) selecionar o subgrupo de variáveis que apresenta a máxima acurácia de classificação e classificar a porção de teste baseado nessas variáveis. Esses passos operacionais estão detalhados na sequência.

3.3.1 Passo 1: Dividir o banco de dados original em porções de treino e teste, e aplicar a ACP na porção de treino

Dividir aleatoriamente o banco de dados em uma porção de treino com N_{tr} observações e uma porção de teste com N_{ts} observações, tal que $N_{tr} + N_{ts} = N$. A porção de

treino é utilizada para selecionar as variáveis mais importantes e a porção de teste representa as novas observações a serem classificadas. Diferentes proporções de N_{tr} e N_{ts} serão testadas no método apresentado, conforme descrito no Passo 4.

Em seguida, caracterizar a relação entre variáveis na porção de treino utilizando a técnica multivariada ACP. Os parâmetros gerados pela ACP fornecem informações relevantes sobre como as variáveis e componentes principais (combinações lineares das variáveis) explicam a variância nos dados. Tais informações são utilizadas para avaliar a importância das variáveis no método proposto. Os parâmetros de interesse incluem os pesos (ou cargas) dos componentes (p_{jr}) e o percentual da variância explicado pelo componente retido r $(r = 1, \dots, R)$, λ_r. O número de componentes a serem retidos pode ser definido com base na variância acumulada, conforme sugerido em Montgomery *et al*. (2001).

3.3.2 Passo 2: Gerar índices de importância das variáveis utilizando os parâmetros da ACP

O índice de importância das variáveis permite guiar a remoção das variáveis menos relevantes. O índice associado à variável j é denotado por v_j, $j = 1, \dots, J$. Quanto maior o valor de v_j, mais importante é a variável j na categorização das observações em classes.

O índice v_j é gerado baseado nos pesos da ACP (p_{jr}) e no percentual de variância explicado por cada componente retido (λ_r); ver equação (5). As variáveis com o maior p_{jr} nos componentes com maior valor de λ_r serão as preferidas, uma vez que apresentam elevada variabilidade e permitem uma melhor discriminação das observações em classes (DUDA *et al*., 2001). Um índice similar é proposto por Anzanello *et al*. (2011), mas não leva em consideração o percentual da variância explicada por cada componente retido.

$$v_j = \sum_{r=1}^{R} |p_{jr}| \lambda_r , j = 1, \dots, J \tag{5}$$

3.3.3 Passo 3: Classificar a porção de treino utilizando os métodos de classificação KVP e AD, e eliminar as variáveis menos relevantes

Classificar as observações de treino em duas classes considerando todas as J variáveis utilizando KVP e AD, separadamente. O método de classificação KVP insere observações em categorias binárias, 0 ou 1, baseada na distância euclidiana da observação aos *k*-vizinhos mais próximos. Cada um dos *k*-vizinhos tem sua classe conhecida *a priori*; a nova observação é alocada na classe 0 se a maioria dos *k*-vizinhos mais próximos estiver em 0. O valor de k é selecionado de forma a maximizar a acurácia de classificação na porção de treino, onde a classe de cada observação é previamente conhecida.

Por sua vez, a AD é um método de classificação e discriminação de amostras (classifica as observações em classes distintas), que permite alocar novas observações a grupos pré-determinados. A AD permite a classificação de novas observações nos grupos já existentes sem a necessidade de rearranjar os grupos. Um grupo de observações onde os membros já estão identificados é utilizado para estimar pesos (ou cargas) de uma função discriminante conforme alguns critérios. O propósito do método é, basicamente, estimar a relação entre uma variável dependente e um conjunto de variáveis independentes. Essa relação é expressa através de uma função discriminante consistindo em uma combinação linear das variáveis independentes (ANZANELLO *et al*., 2011).

Concluída a primeira classificação, calcular a acurácia de classificação, definida como a proporção de classificações corretas relativamente ao total de classificações realizadas. Em seguida, identificar e remover a variável com o menor valor de v_j. Realizar uma nova classificação considerando as $J - 1$ variáveis remanescentes e recalcular a acurácia de classificação. Esse procedimento é repetido removendo a próxima variável com menor valor de v_j e aplicando KVP e AD nas variáveis remanescentes, até restar uma única variável.

3.3.4 Passo 4: Selecionar o melhor subgrupo de variáveis e classificar a porção de teste utilizando as variáveis selecionadas

Selecionar o subgrupo de variáveis que apresenta a máxima acurácia gerada pelos classificadores KVP e AD. No caso de haver subgrupos alternativos com valores de acurácia idênticos, escolher aquele com o menor número de variáveis retidas. Na sequência, classificar a porção de teste utilizando as variáveis selecionadas e calcular a acurácia.

A fim de avaliar a consistência do método proposto, repetir os passos 1 a 4 em diferentes proporções de N_{tr} e N_{ts}, de forma a garantir a consistência do método frente a diferentes partições do banco de dados original. Para cada proporção N_{tr}/N_{ts} repetir o método proposto em amostras contendo um número elevado de dados, gerados misturando e dividindo as observações do WBCD aleatoriamente, certificando-se de que todas as observações apareçam pelo menos uma vez na porção de teste. Em seguida calcular a média da acurácia de classificação e o número de variáveis retidas para cada proporção, e identificar as variáveis que aparecem com mais frequência nos subgrupos selecionados.

Medidas alternativas de desempenho de classificação podem ser calculadas para a porção de teste, incluindo sensibilidade e especificidade. Tais medidas são assim definidas. Considere duas classes: positivo, representando um caso de nódulo mamário maligno (tumor/câncer), e negativo, representando um caso de nódulo mamário benigno. Em seguida, considere quatro subgrupos possíveis de classificações: 1) positivos verdadeiros (PV), representando classificações corretas de casos positivos; 2) negativos verdadeiros (NV), representando classificações corretas de casos negativos; 3) positivos falsos (PF), representando classificações erradas de casos negativos; e 4) negativos falsos (NF), representando classificações erradas de casos positivos. A sensibilidade, dada pela equação (6), corresponde à fração de casos positivos corretamente classificados; a especificidade, dada pela equação (7), corresponde à fração de casos negativos corretamente classificados.

$$Sensibilidade = \frac{PV}{PV + NF}. \tag{6}$$

$$Especificidade = \frac{NV}{NV + PF}. \quad (7)$$

3.4 Resultados

O WBCD é composto por 699 observações (16 delas incompletas) obtidas a partir da punção aspirativa por agulha fina (PAAF) de células da mama. A PAAF é um procedimento médico direcionado à investigação de pacientes com massas, que permite a investigação da malignidade em nódulos mamários (ALBRECHT *et al*., 2002). A técnica consiste na retirada de pequena porção de tecido por aspiração através de uma agulha fina e posterior coloração e análise microscópica. Nove variáveis foram analisadas em cada amostra de células da mama, utilizando uma escala de valores inteiros de 10 pontos; as variáveis estão listadas na Tabela 3.2. A classe (benigna ou maligna) a que cada observação pertence é conhecida. Na amostra de 683 valores completos utilizada nesta análise, há 239 casos malignos e 444 casos benignos. O banco de dados está disponível em http://www.ics.uci.edu/~mlearn/MLRepository. Este banco de dados foi selecionado para testar o desempenho do método proposto porque a literatura apresenta um grande número de pesquisas sobre técnicas de classificação utilizando o WBCD.

Tabela 3.2 - Código e descrição das variáveis no banco de dados WBCD

Código	Descrição
F_1	Aglomeração de células
F_2	Uniformidade do tamanho celular
F_3	Forma celular uniforme
F_4	Adesão marginal
F_5	Tamanho da célula epitelial sozinha (ou de uma célula)
F_6	Núcleo desencapado
F_7	Cromatina frouxa (ou não condensada)
F_8	Nucléolo normal
F_9	Mitose

Fonte: elaborado pelos autores.

Uma visualização do WBCD é apresentada na Figura 3.1.

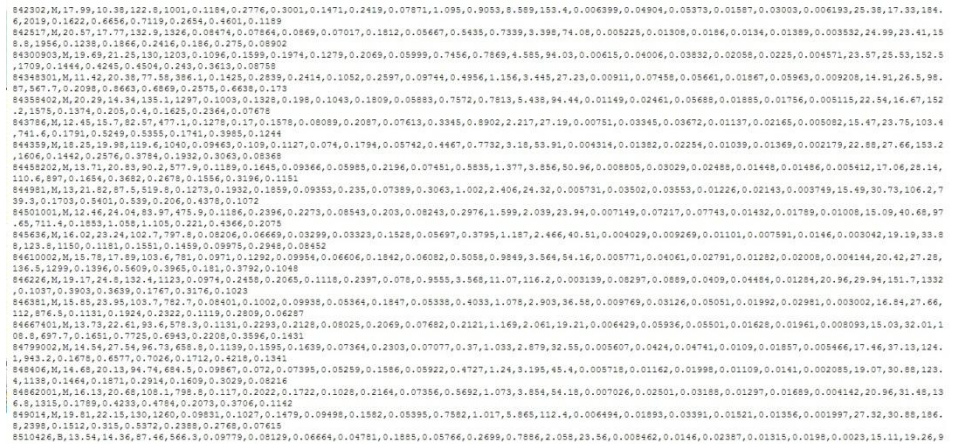


Figura 3.1 - Visualização do WBCD.

Aplicando a técnica proposta, para cada proporção 1000 repetições foram executadas em grupos de treino e teste obtidos amostrando aleatoriamente as observações do WBCD. A Tabela 3.3 apresenta a média da acurácia de classificação para as porções de treino e teste, nas diferentes proporções N_{tr}/N_{ts} e a média da sensibilidade e especificidade utilizando as variáveis selecionadas em cada repetição, utilizando o método KVP e AD. Utilizando KVP, o método proposto atinge a maior acurácia média de classificação de 97,77%, ao reter 5,87 variáveis, em média e o melhor desempenho para a sensibilidade, com uma média de 0,9790. Utilizando a AD, o método proposto atinge a maior acurácia média de classificação de 97,07%, ao reter 5,95 variáveis, em média, e o melhor desempenho para especificidade, com uma média de 0,9856. O método KVP apresenta maior acurácia retendo um menor número de variáveis, em comparação ao método AD. O método KVP apresentou melhor acurácia que o método AD para todas as proporções testadas. O método KVP apresentou melhor desempenho para sensibilidade e pior desempenho para especificidade em relação ao método AD para todas as proporções testadas. Essas medidas de classificação parecem aumentar conforme a proporção N_{tr}/N_{ts} aumenta, sugerindo que quanto maior a porção de treino, mais informação é oferecida para a construção do modelo de classificação. A acurácia média é uma medida de desempenho de classificação mais confiável que a acurácia estimada, obtida do banco de dados através das porções de treino e teste. Executando uma única repetição do método classificatório em uma porção favorável do banco de dados pode levar a resultados não confiáveis.

Tabela 3.3 - Média da acurácia de classificação da porção de teste, das variáveis retidas e das medidas de classificação para as proporções testadas utilizando os métodos KVP e AD

Média das medidas	Observações (%porção de treino/%porção de teste)							
	50%/50%		70%/30%		80%/20%		90%/10%	
	KVP	AD	KVP	AD	KVP	AD	**KVP**	**AD**
Acurácia	0,9702	0,9642	0,9716	0,9654	0,9739	0,9670	**0,9777**	0,9707
N° variáveis retidas	7,15	7,18	6,96	6,97	6,61	6,50	**5,87**	5,95
Sensibilidade	0,9593	0,9317	0,9668	0,9350	0,9733	0,9384	**0,9790**	0,9463
Especificidade	0,9766	0,9821	0,9748	0,9826	0,9752	0,9835	0,9785	**0,9856**

Fonte: elaborado pelos autores.

Na Tabela 3.4 é apresentada a frequência de inclusão das variáveis nas repetições das amostragens realizadas nas diferentes proporções N_{tr} / N_{ts} do banco de dados utilizando o KPV, o qual obteve a melhor acurácia. Há uma pequena variação no número de variáveis responsável pela máxima acurácia (a máxima acurácia foi obtida retendo 5 ou 6 variáveis). As variáveis 9, 7 e 6 foram retidas com maior frequência, independente da proporção N_{tr} / N_{ts}. As variáveis 5, 3 e 1, retidas em mais de 59,7% dos subgrupos selecionados, são omitidas em alguns subgrupos selecionados em virtude da variabilidade nas observações da porção de treino. Essa variabilidade gera diferentes pesos da ACP e pequenas mudanças na ordem da eliminação recursiva das variáveis.

Tabela 3.4 - Inclusão das variáveis nos subgrupos retidos para as proporções testadas

Porções do Banco de Dados (% treino / % teste)							
50% / 50%		70% / 30%		80% / 20%		90% / 10%	
variável	inclusão no subgrupo retido (%)	variável	inclusão no subgrupo retido (%)	variável	inclusão no subgrupo retido (%)	variável	inclusão no subgrupo retido (%)
9	100	9	100	9	100	9	100
6	100	6	100	7	100	7	100
7	97,0	7	100	6	99,5	6	99,0
1	86,0	3	88,5	5	89,0	5	83,3
3	84,5	5	88,5	3	81,0	3	69,7
5	82,0	1	84,0	1	76,5	1	59,7
2	64,0	2	59,5	2	56,0	2	40,3
4	52,5	4	41,0	4	34,0	4	23,0
8	49,0	8	35,0	8	25,5	8	14,0

Fonte: elaborado pelos autores.

Para uma melhor visualização dos resultados de classificação, uma matriz de confusão é apresentada na Tabela 3.5. O pequeno número de erros de classificação, particularmente na porção 90% /10% do banco de dados, representa um desempenho satisfatório do método.

Tabela 3.5 - Matriz de confusão para as proporções testadas

Real	Predito		Porções do banco de dados (% treino / % teste)
	Benigno	Maligno	
Benigno	212,1	9,2	50% / 50%
Maligno	4,7	115,1	
Benigno	130,4	3,4	70% / 30%
Maligno	2,4	68,7	
Benigno	86,9	2,2	80% / 20%
Maligno	1,3	46,6	
Benigno	43,6	0,95	90% / 10%
Maligno	0,51	23,9	

Fonte: elaborado pelos autores.

3.5 Conclusões

O diagnóstico precoce aumenta as taxas de sobrevivência em pacientes com CM, justificando o grande número de abordagens com o objetivo de classificar corretamente os nódulos mamários em duas classes, benigno e maligno, baseado em amostras de célula da mama. O acerto do diagnóstico tem um impacto direto nos recursos destinados à saúde pública, já que o câncer está evoluindo para uma condição crônica em muitos países. O método proposto pode aumentar o acerto do diagnóstico.

O método proposto seleciona as variáveis mais relevantes para fins de classificação de forma a maximizar a sua acurácia, além de propor o teste de dois métodos de classificação na análise de um banco de dados. As proposições são testadas no banco de dados WBCD. Primeiramente as variáveis são ordenadas utilizando um novo índice de importância baseado nos pesos da ACP e na variância explicada por cada componente retido. Em seguida, o método proposto classifica iterativamente os registros dos pacientes em duas classes, benigno e maligno, através de dois métodos de mineração de dados, KVP e AD; a variável menos importante é removida e a classificação realizada nas variáveis restantes até restar uma única variável. O método proposto, para uma proporção de 90% /10%%, classificou corretamente os

dados do WBCD em 97,77% dos casos, em média, utilizando uma média de 5,8 variáveis na classificação utilizando o método KVP. O melhor desempenho para a sensibilidade foi de 0,9790 utilizando o método KVP, e o melhor desempenho para especificidade foi de 0,9856 utilizando o método AD. É importante ressaltar que, para o rastreamento de câncer de mama, o método deve ser o mais sensível possível para que se consiga detectar o maior número de casos possível da doença. Desenvolvimentos futuros incluem testes com técnicas multivariadas mais robustas para identificar as variáveis mais relevantes, e sua integração com métodos alternativos de mineração de dados para fins de classificação. Também pretende-se transformar os dados originais utilizando técnicas de *Kernel*, com o objetivo melhorar o desempenho de classificação dos métodos de mineração de dados.

3.6 Referências

ABBASS, H. A. An evolutionary artificial neural networks approach for breast cancer diagnosis. **Artificial Intelligence in Medicine**, v. 25, p. 265-281, 2002.

ABONYI, J.; SZEIFERT, F. Supervised fuzzy clustering for the identification of fuzzy classifiers. **Pattern Recognition Letters**, v. 14, p. 2195-2207, 2003.

AKAY, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, p. 3240-3247, 2009.

ALBRECHT, A. A.; LAPPAS, G.; VINTERBO, S. A.; WONG, C.K.; OHNO-MACHADO, L. Two applications of the LSA macrine. In: **Proceedings of the 9th International Conference on Neural Information Processing**, Nov 18-22, Singapore, p.184-189, 2002.

ANZANELLO, M. J.; FOGLIATTO, F. S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preference**, v. 22, p. 139-148, 2011.

BAKER, L. H. Breast cancer detection demonstration Project: five-year summary report. **CA – A Cancer Journal for Clinicians**, v. 32, p. 194-225, 1982.

BRASIL. Ministério da Saúde. Departamento de Informática do SUS. Informações de saúde. Indicadores de saúde. [site da Internet]. [acessado em 2011 maio 03]. Disponível em: http://tabnet.datasus.gov.br/cgi/dh.exe?pacto/2010/cnv/pactbr.def

BRASIL. Ministério da Saúde. Instituto Nacional de Câncer (INCA). Controle do Câncer de Mama. Documento de Consenso. ***INCA*** [site na Internet]. 2004 Abr [acessado 2012 jul 23]; [cerca de 39 p.]. Disponível em: http://www1.inca.gov.br/publicacoes/Consensointegra.pdf

BRASIL. Ministério Da Saúde. Instituto Nacional de Câncer (INCA). Coordenação Geral de Ações Estratégicas. Encontro internacional sobre rastreamento do câncer de mama: resumo das apresentações. – Rio de Janeiro: INCA, 2009. 393 p. : il. color.

BRAY, F.; REN, J. S.; MASUYER, E.; FERLAY, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. **International Journal of Cancer** [serial on the Internet] 2012 Jul [cited 2012 set 23]; [about 13 p.]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/ijc.27711/pdf

CHEN, Y. W.; LIN, C. J. Combining SVMs with various feature selection strategies. **Studies in Fuzziness and Soft Computing** [serial on the internet] 2006 [cited 2012 Jan 27]; 207: [about 9 p.]. Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/" features.pdf.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent Data Analysis**, v. 1, p. 131-156, 1997.

DUDA, R.; HART, P.; STORK, D. **Pattern Recognition**. 2° ed. New York: Willey, 2001.

ELTOUKHY, M. M.; FAYE, I.; SAMIR, B. B. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. **Computers in Biology and Medicine**, v. 42, p. 123-128, 2012.

FOGEL, D. B.; WASSON III, E. C.; BOUGHTON, E. M. Evolving neural networks for detecting breast cancer. **Cancer Letters**, v. 96, p. 49-53, 1995.

HUMPHREY, L. L.; HELFAND, M.; CHAN, B. K. S.; WOOLF, S. H. Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. **Annals of Internal Medicine**, v. 137, p. 347-360, 2002.

International Agency for Research on Cancer (IARC). IARC Handbooks of Cancer Prevention Volume 7: Breast Cancer Screening. Lyon: ***IARC*** [serial on the Internet] 2002 [cited 2012 jul 23]; [about 243 p.]. Available from: http://www.iarc.fr/en/ publications/pdfs-online/prev/handbook7/index.php

LEE, H.-M.; CHEN, C.-M.; CHEN, J.-M.; JOU, Y.-L. An efficient fuzzy classifier with feature selection based on fuzzy entropy. **IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics**; v. 31(3), p. 426-432, 2001.

MARCANO-CEDEÑO, A.; QUINTANILLA-DOMÍNGUEZ, J.; ANDINA, A. WBCD breast cancer database classification applying artificial metaplasticity neural network. **Expert Systems with Applications**, v. 38, p. 9573-9579, 2011.

MONTGOMERY, D.; PECK, E.; VINING, G. **Introduction to Linear Regression Analysis**. 3° ed. New York: Willey; 2001.

NAUCK, D.; KRUSE, R. Obtaining interpretable fuzzy classification rules from medical data. **Artificial Intelligence in Medicine**, v. 16, p. 149-169, 1999.

PEÑA-REYES, C. A; SIPPER, M. A fuzzy-genetic approach to breast cancer diagnosis. **Artificial Intelligence in Medicine**, v. 17, p. 131-155, 1999.

POLAT, K.; GÜNES, S. Breast cancer diagnosis using a least square support vector machine. **Digital Signal Processing**, v. 17, p. 694-701, 2007.

QUINLAN, J. R. Improved use of continuous attributes in C4.5. **Journal of Artificial Intelligence Research**, v. 4, p. 77-90, 1996.

QUINLAN, J. R. **C4.5: Programs for machine learning**. 5° ed. San Mateo: Morgan Kaufmann; 1993.

RENCHER, R. **Methods of multivariate Analysis**. 1° ed. New York: Wiley; 1995.

SETIONO, R. Extracting rules from pruned neural networks for breast cancer diagnosis. **Artificial Intelligence in Medicine**, v. 8, p. 37-51, 1996.

SETIONO, R. Generating concise and accurate classification rules for breast cancer diagnosis. **Artificial Intelligence in Medicine**, v. 18, p. 205-217, 2000.

SHANNON, C.E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, p. 379-423, 1948.

SHAPIRO, S.; VENET, W.; STRAX, P.; VENET, L.; ROESER, R. Ten- to fourteen-year effect of screening on breast cancer mortality. **Journal of the National Cancer Institute**, v. 69, p. 349-355, 1982.

STREET, W. N.; WOLBERG, W. H.; MANGASARIAN, O. L. Nuclear feature extraction for breast tumor diagnosis. **ISandT/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology**, San Jose, California, v. 1905, p. 861-870, 1993.

VERIKAS, A.; BACAUSKIENE, M. Feature selection with neural networks. **Pattern Recognition Letters**, v. 23, p. 1323-1335, 2002.

World Health Organization (WHO). Cancer control: knowledge into action: WHO guide for effective programmes: early detection. ***WHO 2007*** [serial on the Internet] 2007 [cited 2012 Aug 09]; [about 50 p.]. Available from: http://www.who.int/cancer/ publications/cancer_control_detection/en/

4 Terceiro Artigo: Uso das técnicas de *kernel* para melhorar o desempenho do método de mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis

Nicole Holsbach

Michel Jose Anzanello

Flávio Sanson Fogliatto

Resumo

O câncer de mama (CM) é o mais incidente na população feminina mundial e brasileira. Quando diagnosticado precocemente, apresenta altas taxas de sobrevivência. Diversas abordagens estatísticas vêm sido apresentadas para auxiliar na sua detecção precoce. Este artigo apresenta um método para melhorar o desempenho da seleção de variáveis para classificação de casos de câncer de mama em duas classes de resultado, benigno ou maligno, baseado na análise citopatológica de amostras de célula da mama de pacientes. Para tanto, quatro diferentes tipos de *kernels* polinomiais são utilizados para adensar o banco de dados original. As variáveis no banco de dados transformado são ordenadas de acordo com um novo índice de importância de variáveis, que combina os pesos de importância da ACP (Análise de Componentes Principais) e a variância explicada a partir de cada componente retido. Assim, as observações da porção de treino são categorizadas em duas classes através dos métodos KPV (*k*-vizinhos mais próximos) e AD (Análise Discriminante), e a variável com o menor índice de importância é eliminada. Para classificar as observações na porção de teste, o subconjunto de variáveis com a máxima acurácia é utilizado. Aplicando ao WBCD (*Wisconsin Breast Cancer Database*), a maior acurácia de classificação obtida pelo método proposto apresentou uma média de 98,09%, retendo uma média de 17,24 variáveis de um total de 54 variáveis.

Palavras-chave: Seleção de variáveis, Diagnóstico de câncer de mama, *Kernel*.

Abstract

Female breast cancer (BC) is predominant in most countries, and also among Brazilian women. If diagnosed in early stages, it presents a high survival rate. Several statistical-based approaches have been developed to assist its early detection. This paper presents a method for improve the performance in feature selection for the classification of cases into two classes, benign or malignant, based on cytopathologic analysis from patients' breast cell samples. Four different types of polynomial *kernels* are used to enlarge the original database. Features are ranked according to a new feature importance index that combines PCA (Principal Component Analysis) weights and the variance explained by each retained component. Observations of a training set are categorized into two classes through the KNN (k-nearest neighbor) tool and AD (Discriminant Analysis), followed by elimination of the feature with the smallest importance index. The subset with the maximum accuracy is used to classify observations in the testing set. When applied to the WBCD (Wisconsin Breast Cancer Database), it was possible to attain an average accuracy of 98.09% using the proposed method, while retaining an average of 17.24 features from a total of 54.

Keywords: Feature selection, Breast cancer diagnosis, *Kernel*.

4.1 Introdução

A epidemiologia referente ao câncer de mama (CM) no Brasil, com seus elevados índices de incidência e mortalidade, justificam a implementação de ações nacionais voltadas para a prevenção e o controle do câncer (promoção, prevenção, diagnóstico, tratamento, reabilitação e cuidados paliativos). Ações de detecção precoce têm impacto na mortalidade decorrente dessa neoplasia, onde a adoção de estratégias, tais como a padronização de procedimentos e de condutas que garantam a qualidade dos processos técnicos e operacionais para o controle do câncer, são fundamentais para o controle da doença. O câncer de mama é o mais incidente na população feminina mundial e brasileira, excluindo os casos de câncer de pele (não melanoma). Políticas públicas nessa área vêm sendo desenvolvidas no Brasil desde os anos 80 e foram incentivadas pela implementação do Programa Viva Mulher, em 1998, quando tiveram início ações para diretrizes e estruturação da rede assistencial na detecção precoce do CM (BRASIL, 2011).

As ações de combate e controle do câncer no Brasil são definidas pelo Ministério da Saúde em conjunto com diferentes esferas de governo – Estadual, Municipal e do Distrito Federal – apoiadas pelo Instituto Nacional do Câncer (INCA). Dentre elas, destaca-se a Política Nacional de Atenção Oncológica (Portaria Gabinete do Ministro nº 2439/05), que definiu as diretrizes para a atenção integral à população frente a este agravo, incluindo ações de promoção e prevenção do câncer, rastreabilidade, diagnóstico precoce, tratamento e cuidados paliativos. A importância da detecção precoce foi reafirmada no Pacto pela Saúde em 2006. O Sistema de Informação do Câncer de Mama – SISMAMA – foi desenvolvido pelo INCA, em parceria com o Departamento de Informática do SUS (DATASUS), como ferramenta para gerenciar as ações de detecção precoce do câncer de mama, que consistem no seguinte fluxograma de ações: (*i*) requisição da mamografia, (*ii*) resultado da mamografia, (*iii*) requisição de exame citopatológico (Punção Aspirativa por Agulha Fina – PAAF), e (*iv*) requisição de exame histopatológico. A organização das ações de controle foram propostas em junho de 2009, quando o sistema entrou em vigor, com o aumento da oferta de mamografias pelo Ministério da Saúde (Programa Mais Saúde 2008-2011) e com a publicação de documentos (parâmetros técnicos para o rastreamento do câncer de mama) e as recomendações para a redução da mortalidade do câncer de mama no Brasil, lançadas pelo INCA em 2010. O controle do câncer de mama foi reafirmado como prioridade no Plano de Fortalecimento da Rede de Prevenção, Diagnóstico e Tratamento do Câncer, lançado pela presidente da República, em 2011 (BRASIL, 2011; BRASIL, 2004).

Quando diagnosticado precocemente, as taxas de sobrevivência em pacientes com CM aumentam, e a morbidade associada ao tratamento diminui. A estratégia de diagnóstico precoce contribui para a redução do estágio de apresentação do câncer (BRASIL, 2011; IARC, 2002; WHO, 2007; BRASIL, 2004). A doença consiste no crescimento desordenado de células do tecido da mama, formando nódulos que podem ser malignos (tumores) ou benignos. No Brasil, as taxas de mortalidade por câncer de mama continuam elevadas, provavelmente porque a doença ainda é diagnosticada em estados avançados. O diagnóstico do CM depende da avaliação do médico a partir das informações obtidas dos pacientes através de exames. Esses exames incluem exame clínico da mama (autopalpação), mamografia e análise de tecido da mama. Na autopalpação das mamas, valoriza-se a descoberta casual de pequenas alterações mamárias (IARC, 2002; WHO, 2007; BRASIL, 2011; BRASIL, 2004). O

exame clínico da mama fornece dados univariados para interpretação, enquanto que exames de imagem e laboratoriais (citopatológicos) produzem dados multivariados, que demandam maior processamento de informações.

Neste artigo é apresentado um método para melhorar o desempenho de classificação dos métodos de mineração de dados na seleção de variáveis oriundas de exames clínicos, com vistas à classificação de observações em categorias distintas. Com o objetivo melhorar o desempenho de tais técnicas, os dados originais são transformados utilizando técnicas de *kernel*. Quatro tipos de *kernel* polinomial (multiplicativo, quadrático, cúbico e de raiz cúbica) são utilizados para transformar o banco de dados original. A utilização de *kernels* adensa o banco de dados, já que são acrescidas a ele novas variáveis que são função daquelas originalmente no banco. Em seguida, a técnica multivariada Análise de Componentes Principais (ACP) é aplicada ao banco de dados adensado, onde as observações referem-se a pacientes e as variáveis a dados extraídos do exame e suas transformações através da aplicação de *kernels*. As variáveis são ordenadas de acordo com um novo índice que combina os pesos gerados pelos componentes principais retidos na ACP com a variância explicada por estes componentes. Em seguida, as observações da porção de treino são categorizadas em duas classes (benigno ou maligno), utilizando dois métodos de classificação: (*i*) a ferramenta de mineração de dados *k*-vizinhos mais próximos (KVP), e (*ii*) análise discriminante (AD). Calcula-se então a acurácia de classificação. Posteriormente, remove-se a variável com o menor índice de importância e, utilizando as variáveis remanescentes, uma nova classificação é realizada. Esse processo iterativo de eliminação e classificação é repetido até que reste somente uma variável. Por fim, o subconjunto de variáveis que leva à máxima acurácia é escolhido e utilizado para classificar as observações da porção de teste.

Uma contribuição significativa deste trabalho consiste na utilização de *kernels* no adensamento de bancos de dados, com vistas à melhoria do desempenho de classificação dos métodos de mineração de dados. As transformações *kernel*, de natureza não-linear, permitem que a ACP capture direções não-lineares de variabilidade no banco de dados. Dessa forma, a estrutura de correlação linear e não-linear das variáveis é considerada na seleção de variáveis, o que potencialmente promove uma melhoria na classificação dos casos presentes no banco de dados. Trata-se de uma proposição original na literatura sobre seleção de variáveis, conforme atesta a recente revisão de literatura sobre o assunto, em Anzanello e Fogliatto (2012).

Vários estudos sugerindo métodos de classificação testam seu desempenho no *Wisconsin Breast Cancer Database* (WBCD), obtido da universidade de Wisconsin e disponibilizado *online*. Neste banco de dados, nove variáveis foram analisadas de amostra de células da mama de 699 indivíduos, para os quais o diagnóstico foi elaborado.

O artigo está organizado como segue. Na segunda seção é apresentado o referencial teórico sobre as técnicas de *kernel* e exemplos da aplicação do *kernel* polinomial. O método proposto é detalhado na terceira seção. Os resultados obtidos pela aplicação do método proposto são apresentados na quarta seção. A conclusão é apresentada na última seção.

4.2 Referencial teórico

Nesta seção é apresentada uma descrição da teoria sobre *kernels*, suas definições e teoremas, bem como exemplos do *kernel* polinomial.

O nome *kernel* é derivado da teoria do operador integral. As funções de *kernel* podem ser utilizadas para definir relações não lineares entre as suas entradas. Além das funções lineares de *kernel*, é possível definir funções de *kernel* quadráticas ou exponenciais (WU *et al*., 2008). O *kernel* é a função responsável pelo mapeamento do espaço de dados para o espaço de características. A utilização dos *kernels* permite a projeção dos dados em um espaço de maior dimensão em relação ao espaço original (espaço das variáveis). Essa análise é realizada sem acessar diretamente o espaço das variáveis, através do uso de funções de *kernel* sobre os dados de entrada. Dados não linearmente separáveis podem ser mapeados em um espaço de maior dimensão, onde podem ser separados linearmente através da função *kernel* (MARCONDES, 2009; GUYON *et al*., 1993; DAMIAN, 2011).

O conceito de *kernels* define que são funções que recebem dois pontos x_i e x_j do espaço de entradas e calculam o produto escalar $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ no espaço de características. A função *kernel* é menos complexa que a do mapeamento ϕ, portanto é possível defini-la sem conhecer-se explicitamente o mapeamento ϕ (NALDI; CARVALHO, 2005).

Existem diversos tipos de *kernel* $K(x_i, x_j)$, os quais devem obedecer às condições do teorema de Mercer (SCHOLKOPF; SMOLA, 2002). Alguns exemplos como *kernel* linear,

polinomial, Gaussiano, Função de Base Radial (FBR) e sigmoidal estão listados na Tabela 4.1 (FERREIRA *et al*., 2010; SEMOLINI, 2002).

Tabela 4.1 - Tipos de kernel

Kernel	**Expressão**	**Parâmetros de Ajuste**
Linear	$K(x_i, x_j) = x_i^T x_j$	Não possui
Polinomial	$K(x_i, x_j) = (x_j^T x + 1)^p$	$p = grau\ do\ polinômio$
Gaussiano	$K(x_i, x_j) = e^{\left[-\sum_{l=1}^{N} \sigma_l^2 (x_{il} - x_{jl})^2\right]}$	$\sigma_l^2, l = 1,2,\dots,N$
Função de Base Radial (FBR)	$K(x_i, x_j) = exp\left(\frac{-\|x_j - x_i\|^2}{2\sigma^2}\right)$	$\sigma^2 = variância\ da\ RBF$
Sigmoidal	$K(x_i, x_j) = tanh(\gamma x_i x_j + r)$	γ, r

Fonte: elaborado pelos autores.

Considere M vetores de observações, tal que x_i e x_j sejam dois vetores pertencentes a esse conjunto. Considere a matriz quadrada:

$$k_{ij} = k(x_i, x_j), \tag{8}$$

Designada Matriz *Kernel*, de ordem $(M \times M)$, cujos elementos representam produtos internos entre as observações, definidos pelo *kernel* em uso. A matriz na equação (8) é, assim, simétrica, tal que: $k(x_i, x_j) = k(x_j, x_i)$.

O *kernel* que gera um produto interno modificado pode ser usado como medida de similaridade. A matriz simétrica $k(x_i, x_j) \in IR$ é definida como positiva se possui todos os autovalores não negativos. A função $k(x_i, x_j)$ que gera uma matriz *kernel* positiva, é também positiva.

Segundo o Teorema de Mercer (BOSER *et al*., 1992; SEMOLINI, 2002), se k é um *kernel* contínuo de um operador integral positivo, um mapa ϕ em um espaço F onde k é o produto interno pode ser construído (SEMOLINI, 2002). Se k é positivo definido, existe um mapa ϕ, tal que:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{9}$$

Considere o exame citopatológico PAAF, isto é, o problema de classificação em questão neste artigo. As realizações de um processo organizadas em M rodadas de exames geram resultados classificados como benignos ou malignos. Considere que cada rodada do processo seja monitorada por informações contidas no vetor linha $\mathbf{x}_i (i = 1, \dots, M)$, considerando $\mathbf{x}_i$ de dimensão $(1 \times N)$, que representa uma realização de cada uma das N variáveis contínuas do processo, $\mathbf{x}_i \in IR^N$ onde IR^N é o conjunto dos reais no espaço de dimensão N. Para classificar uma nova rodada do processo, é necessário comparar a informação $\mathbf{x}$ a ela associada com os dados de referência $\mathbf{x}_i$. Define-se então uma medida de similaridade entre esses dois conjuntos de dados, a saber:

$$k: IR^N \times IR^N \rightarrow IR$$

$$(\mathbf{x}_i, \mathbf{x}) \rightarrow k(\mathbf{x}_i, \mathbf{x}) \tag{10}$$

Na equação (10), k é uma função que gera um número real para representar a similaridade entre os vetores $\mathbf{x}_i$ e $\mathbf{x}$, com $k(\mathbf{x}_i, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}_i)$. A função k é denominada *kernel*.

As medidas de similaridade utilizadas em análise multivariada de dados utilizam o produto interno canônico, a seguir:

$$\langle x_i, x \rangle = \sum_{w=1}^{N} [x_i]_w [x]_w, \tag{11}$$

Onde $[x_i]_w$ é a *w*-ésima coluna dos vetores linha $\mathbf{x}_i$ e $\mathbf{x}$ respectivamente, e $[x]_w$ é a *w*-ésima coluna dos vetores linha $\mathbf{x}_i$ e $\mathbf{x}$ respectivamente. O produto interno na equação (11) é a medida do cosseno do ângulo entre os vetores $\mathbf{x}_i$ e $\mathbf{x}$.

O comprimento de um vetor é obtido em função do produto interno canônico:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{12}$$

A distância euclidiana entre dois vetores é obtida em função do produto interno canônico:

$$\|\mathbf{x}_i - \mathbf{x}\|^2 = \langle \mathbf{x}_i, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}_i, \mathbf{x} \rangle \quad (13)$$

Medidas de similaridade também podem ser obtidas a partir da construção de um mapa não linear ϕ:

$$: IR^N \to F$$

$$x \to \phi(x), \quad (14)$$

Onde F é o espaço dos produtos internos ou espaço das variáveis, sendo $(N_F \geq IR^N)$ e $\phi(x)$ é a representação do vetor $\boldsymbol{x}$ no espaço dos atributos F.

A aplicação dos dados de entrada em F, através de ϕ, permite construções geométricas baseadas em produtos internos modificados no espaço de entrada, através da aplicação de uma função k (*kernel*):

$$k(\mathbf{x}_i, \mathbf{x}) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle. \quad (15)$$

A escolha do mapa ϕ permite construir diversas medidas de similaridade no espaço de entrada IR^N, via função k.

No *kernel* tipo polinomial, o único parâmetro a ser determinado pelo usuário é o grau do polinômio a ser utilizado (SILVA *et al*., 2010). O *Kernel* polinomial considera, por exemplo, um vetor $\boldsymbol{x}$ de observações, de dimensão (1×2), com 2 valores e uma função $\phi(x)$:

$$\phi: IR^2 \to F = IR^3$$

$$\mathbf{x} = ([x]_1, [x]_2) \to \phi(\mathbf{x}) = ([x]_1^2, [x]_2^2, [x]_1[x]_2) \quad (16)$$

As informações do vetor $\boldsymbol{x}$ passam a ser analisadas no espaço dos produtos de segunda ordem de seus elementos. No espaço F, utilizar o produto interno canônico entre vetores $\phi(\mathbf{x}_i)$ e $\phi(\mathbf{x})$ equivale a multiplicar monômios (termo que contém apenas o produto de constantes e variáveis) de segunda ordem nos dados de entrada. Porém, também é possível calcular esses produtos sem utilizar os vetores $\phi(\mathbf{x}_i)$ e $\phi(\mathbf{x})$, tal que $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle = k(\mathbf{x}_i, \mathbf{x})$.

Considere o mapa:

$$\phi(\mathbf{x}) = ([x]_1^2, [x]_2^2, [x]_1[x]_2, [x]_2[x]_1) \tag{17}$$

Onde $([x]_1[x]_2)$ *e* $([x]_2[x]_1)$ são distintos e o produto interno no espaço das variáveis F entre os vetores $\phi(\mathbf{x}_i)$ e $\phi(\mathbf{x})$ assim definido:

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle = \langle \mathbf{x}_i, \mathbf{x} \rangle^2 = k(\mathbf{x}_i, \mathbf{x}) \tag{18}$$

É possível obter os produtos internos entre monômios de segunda ordem no espaço das variáveis sem utilizar ϕ, utilizando o *kernel* que calcula o quadrado do produto interno canônico entre as observações originais. Através do *kernel* polinomial de segunda ordem, as estruturas não lineares quadráticas podem ser analisadas (correlações de segunda ordem entre as variáveis; MARCONDES, 2009; SANTOS, 2002; DAMIAN, 2011).

No exemplo citado por Marcondes (2009), a equação (18) pode ser generalizada para os vetores $\mathbf{x}_i$ e $\mathbf{x} \in IR^N$, e para o espaço F de ordem p. Assim, o *kernel* que calcula o produto interno entre os vetores $\phi(\mathbf{x}_i)$ e $\phi(\mathbf{x})$ em F é dado por:

$$k(\mathbf{x}_i, \mathbf{x}) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle = \langle \mathbf{x}_i, \mathbf{x} \rangle^p \tag{19}$$

Para determinada escolha de N e p, o espaço F tem dimensão muito superior ao espaço de entrada N, conforme segue:

$$N_F = \frac{(p + N - 1)!}{p!\,(N - 1)!} \tag{20}$$

Considerando, por exemplo, um processo industrial com observações coletadas acerca de 20 variáveis ($\mathbf{x} \in IR^{20}$) e $p = 5$, F trabalharia com vetores $\phi(\boldsymbol{x})$ de dimensão (1×24504). Porém, esse mapa não é necessariamente acessado e as não-linearidades entre as variáveis podem ser analisadas no espaço dos dados $\boldsymbol{x}$ de entrada, de dimensão (1×20), utilizando o *kernel* da equação (19).

4.3 Método

O método de seleção de variáveis para categorização das observações do WBCD em duas classes é operacionalizado em 4 passos: (*i*) aplicar o *kernel* polinomial no banco de dados original. Posteriormente, dividir os dados transformados em porções de treino e de teste, e aplicar a ACP na porção de treino; (*ii*) gerar índices de importância das variáveis baseados nos pesos da ACP e na percentagem da variância explicada pelos componentes retidos; (*iii*) classificar a porção de treino utilizando KVP e AD individualmente. A seguir eliminar a variável com o menor índice de importância, classificar o banco de dados novamente, e calcular a acurácia de classificação. Continuar tal processo iterativo até restar uma variável; e (*iv*) selecionar o subgrupo de variáveis que apresenta a máxima acurácia de classificação e classificar a porção de teste baseado nessas variáveis. Esses passos operacionais estão detalhados na sequência.

4.3.1 Passo 1: Aplicar a função kernel no banco de dados original

Utilizar o *kernel* polinomial para transformar o banco de dados original, determinando o grau do polinômio a ser utilizado. Considerando que cada *kernel* utilizado leva a um adensamento distinto do banco de dados, a melhor escolha do polinômio é função da acurácia de classificação resultante. Trata-se, assim, de um procedimento iterativo no qual três parâmetros são testados: o grau do polinômio da função *kernel*, a partição de treino e teste do banco de dados e o conjunto de variáveis de classificação a ser utilizado.

Na sequência, divide-se aleatoriamente o banco de dados transformado em uma porção de treino, com N_{tr} observações, e uma porção de teste, com N_{ts} observações, tal que $N_{tr} + N_{ts} = N$. A porção de treino é utilizada para selecionar as variáveis mais relevantes; a porção de teste representa as novas observações a serem classificadas. Diferentes proporções de N_{tr}e N_{ts} serão testadas no método, conforme descrito no Passo 4.

Posteriormente, caracterizar a relação entre variáveis na porção de treino utilizando a técnica multivariada ACP. Os parâmetros gerados pela ACP fornecem informações relevantes sobre como as variáveis e componentes principais (combinações lineares das variáveis) explicam a variância nos dados. Tais informações são utilizadas para avaliar a importância das variáveis no método proposto. Os parâmetros de interesse incluem os pesos (ou cargas)

dos componentes (p_{jr}) e o percentual da variância explicado pelo componente retido r $(r = 1, \dots, R)$, λ_r. O número de componentes a serem retidos pode ser definido com base na variância acumulada, conforme sugerido em Montgomery *et al*. (2001).

4.3.2 Passo 2: Gerar índices de importância das variáveis utilizando os parâmetros da ACP

O índice de importância das variáveis permite guiar a remoção daquelas menos relevantes. O índice associado à variável j é designado por v_j, $j = 1, \dots, J$. Quanto maior o valor de v_j, mais importante é a variável j na categorização das observações em classes.

O índice v_j é gerado baseado nos pesos da ACP (p_{jr}) e no percentual de variância explicado por cada componente retido (λ_r); ver equação (21). As variáveis com maior valor de p_{jr} nos componentes com maior valor de λ_r serão as eleitas, já que apresentam elevada variabilidade e permitem uma melhor discriminação das observações em classes (DUDA *et al*., 2001). Um índice similar é proposto por Anzanello *et al*. (2011), mas considera o percentual da variância explicada por cada componente retido.

$$v_j = \sum_{r=1}^{R} |p_{jr}| \lambda_r \, , j = 1, \dots, J \tag{21}$$

4.3.3 Passo 3: Classificar a porção de treino utilizando os métodos de classificação KVP e AD, e eliminar as variáveis menos relevantes

Fazer a classificação das observações de treino em duas classes considerando todas as j variáveis utilizando KVP e AD, individualmente. O método de classificação KVP introduz observações em categorias binárias, 0 ou 1, baseada na distância euclidiana da observação aos *k*-vizinhos mais próximos. Cada um dos *k*-vizinhos tem sua classe conhecida *a priori*; a nova observação é alocada na classe 0 se a maioria dos *k*-vizinhos mais próximos estiver em 0. O valor de k é selecionado de forma a maximizar a acurácia de classificação na porção de treino, onde a classe de cada observação é previamente conhecida.

Já a AD é um método de classificação e discriminação de amostras (classifica as observações em classes distintas), que permite alocar novas observações a grupos pré-determinados. A AD permite a classificação de novas observações nos grupos já existentes sem a necessidade de rearranjar os grupos. Um grupo de observações onde os membros já estão identificados é utilizado para estimar pesos (ou cargas) de uma função discriminante, conforme alguns critérios. O propósito do método é estimar a relação entre uma variável dependente e um conjunto de variáveis independentes. Essa relação é expressa através de uma função discriminante consistindo em uma combinação linear das variáveis independentes (ANZANELLO *et al*., 2011).

Após a conclusão da primeira classificação é necessário calcular a sua acurácia, definida como a proporção de classificações corretas relativamente ao total de classificações realizadas. A seguir, identificar e remover a variável com o menor valor de v_j. Fazer uma nova classificação considerando as $J-1$ variáveis remanescentes e recalcular a acurácia de classificação. Esse método é repetido removendo a próxima variável com menor valor de v_j e aplicando KVP e AD nas variáveis remanescentes, até restar uma única variável.

4.3.4 Passo 4: Selecionar o melhor subgrupo de variáveis e classificar a porção de teste utilizando as variáveis selecionadas

Eleger o subgrupo de variáveis que apresenta a máxima acurácia gerada pelos classificadores KVP e AD. Se houver subgrupos alternativos com valores de acurácia idênticos, selecionar aquele com o menor número de variáveis retidas. Em seguida, classificar a porção de teste utilizando as variáveis selecionadas e calcular a acurácia.

Com o propósito de avaliar a consistência do método escolhido, repetir os passos 1 a 4 em diferentes proporções de N_{tr} e $N_{ts}\cdot$, de forma que a consistência do método seja garantida frente a diferentes partições do banco de dados. Para cada proporção N_{tr}/N_{ts}, repetir o método proposto em amostras contendo um número elevado de dados, gerados misturando e dividindo as observações do WBCD aleatoriamente, certificando-se de que todas as observações apareçam pelo menos uma vez na porção de teste. Posteriormente, calcular o valor médio da acurácia de classificação e o número de variáveis retidas para cada proporção, e identificar as variáveis que aparecem com mais frequência nos subgrupos selecionados.

Algumas medidas de desempenho de classificação podem ser calculadas para a porção de teste, tais como sensibilidade, especificidade e valor predito positivo e negativo. Pode-se definir tais medidas. Considere duas classes: positivo, representando um caso de nódulo mamário maligno (tumor/câncer), e negativo, representando um caso de nódulo mamário benigno. Em seguida, considere quatro subgrupos possíveis de classificações: 1) positivos verdadeiros (PV), representando classificações corretas de casos positivos; 2) negativos verdadeiros (NV), representando classificações corretas de casos negativos; 3) positivos falsos (PF), representando classificações erradas de casos negativos; e 4) negativos falsos (NF), representando classificações erradas de casos positivos. A sensibilidade, dada pela equação (22), corresponde à fração de casos positivos corretamente classificados; a especificidade, dada pela equação (23), corresponde à fração de casos negativos corretamente classificados. Os valores preditos positivo e negativo são dados pelas equações (24) e (25), respectivamente.

$$Sensibilidade = \frac{PV}{PV + NF} \tag{22}$$

$$Especificidade = \frac{NV}{NV + PF} \tag{23}$$

$$Valor\ Predito\ Positivo = \frac{PV}{VP + PF} \tag{24}$$

$$Valor\ Predito\ Negativo = \frac{NF}{NF + NV} \tag{25}$$

Os passos 1 a 4 acima descritos devem ser repetidos para diferentes escolhas de função polinomial *kernel*.

4.4 Resultados

O WBCD é composto por 699 observações (16 delas incompletas) obtidas a partir da punção aspirativa por agulha fina (PAAF) de células da mama. A PAAF é um procedimento médico direcionado à investigação de pacientes com massas, que permite a investigação da malignidade em nódulos mamários (ALBRECHT *et al*., 2002). A técnica consiste na retirada de pequena porção de tecido por aspiração através de uma agulha fina e posterior coloração e análise microscópica. Nove variáveis foram analisadas em cada amostra de células da mama, utilizando uma escala de valores inteiros de 10 pontos; as variáveis estão listadas na Tabela 4.2. A classe (benigna ou maligna) a que cada observação pertence é conhecida. Na amostra de 683 valores completos utilizada nesta análise, há 239 casos malignos e 444 casos benignos. O banco de dados está disponível em http://www.ics.uci.edu/~mlearn/MLRepository. Este banco de dados foi selecionado para testar o desempenho do método proposto porque a literatura apresenta um grande número de pesquisas sobre técnicas de classificação utilizando o WBCD.

Tabela 4.2 - Código e descrição das variáveis no banco de dados WBCD

Código	**Descrição**
F_1	Aglomeração de células
F_2	Uniformidade do tamanho celular
F_3	Forma celular uniforme
F_4	Adesão marginal
F_5	Tamanho da célula epitelial sozinha (ou de uma célula)
F_6	Núcleo desencapado
F_7	Cromatina frouxa (ou não condensada)
F_8	Nucléolo normal
F_9	Mitose

Fonte: elaborado pelos autores.

Uma visualização do WBCD é apresentada na Figura 4.1.

Figura 4.1 - Visualização do WBCD.

Para cada proporção, 200 repetições foram realizadas em grupos de treino e teste obtidos amostrando aleatoriamente as observações do WBCD. A Tabela 4.3 apresenta a média da acurácia de classificação para as porções de treino e teste, nas diferentes proporções N_{tr}/N_{ts}, para diferentes escolhas de polinômio da função *kernel*. Foram testadas quatro funções: quadrática (x^2), cúbica (x^3), de raiz quadrada ($\sqrt{x}$) e multiplicativa ($x_i x_j$). No banco de dados adensado, foram mantidas as variáveis originais somente quando da utilização do *kernel* multiplicativo; nas demais funções, as variáveis originais foram removidas, pois a acurácia de classificação resultante era maior.

Na Tabela 4.3 também são apresentadas a média da sensibilidade, especificidade, valor predito positivo e valor predito negativo utilizando as variáveis selecionadas em cada repetição, utilizando o método KVP e AD individualmente para cada tipo de *kernel* utilizado.

Utilizando KVP, o método proposto atinge a maior acurácia média de classificação de 98,09%, ao reter 17,24 variáveis, em média, e o melhor desempenho para a sensibilidade, com uma acurácia média de 0,9818. Em ambos os casos, foi utilizado um adensamento da matriz de dados através do *kernel* multiplicativo. A partição de melhor desempenho do banco de dados utilizou 90% dos dados na porção de treino e 10% dos dados na porção de teste.

Utilizando a AD, o método proposto atinge a maior acurácia média de classificação de 97,65%, ao reter 23,72 variáveis, em média (para tanto, foi utilizado o adensamento da matriz de dados através do *kernel* multiplicativo e uma partição 90%/10% do banco de dados), e o

melhor desempenho para especificidade, com uma média de 0,9898 (para tanto, foi utilizado o adensamento da matriz de dados através do *kernel* quadrático e uma partição 90%/10% do banco de dados).

O método KVP apresenta maior acurácia retendo um menor número de variáveis, em comparação ao método AD. O método KVP apresentou melhor acurácia que o método AD para todas as proporções testadas. O método KVP apresentou melhor desempenho para sensibilidade e para valor predito negativo e pior desempenho para especificidade e para valor predito positivo em relação ao método AD para todas as proporções testadas. Essas medidas de classificação tendem a aumentar conforme a proporção N_{tr}/N_{ts} aumenta, pois quanto maior a porção de treino, mais informação tem-se disponível para a construção do modelo de classificação. A acurácia média é uma medida de desempenho de classificação mais confiável que a acurácia estimada, obtida do banco de dados através das porções de treino e teste. Executando uma única repetição do método classificatório em uma porção favorável do banco de dados pode levar a resultados não confiáveis.

Tabela 4.3 - Média da acurácia de classificação da porção de teste, das variáveis retidas e das medidas de classificação de acordo com o *kernel* utilizado para as proporções testadas utilizando os métodos KVP e AD

		Observações (%porção de treino/%porção de teste)							
Média das medidas	*Kernel*	50%/50%		70%/30%		80%/20%		90%/10%	
		KVP	AD	KVP	AD	KVP	AD	KVP	AD
Acurácia	x^2	0,9639	0,9585	0,9647	0,9607	0,9674	0,9611	0,9712	0,9614
	x^3	0,9630	0,9579	0,9663	0,9605	0,9676	0,9606	0,9704	0,9670
	$x^{1/2}$	0,9698	0,9682	0,9725	0,9695	0,9742	0,9721	0,9792	0,9738
	$x_i x_j$	0,9710	0,9697	0,9749	0,9721	0,9778	0,9739	**0,9809**	0,9765
N° variáveis retidas	x^2	7,8050	7,4200	7,0800	7,2750	6,5550	6,9200	5,5600	6,0100
	x^3	7,6000	7,6550	7,3000	7,1700	6,7350	6,8700	5,8650	6,1550
	$x^{1/2}$	7,2450	7,2250	6,8300	6,6350	6,6400	6,1200	5,6300	5,3200
	$x_i x_j^*$	27,980	32,585	24,695	30,470	22,020	27,795	**17,240**	23,725
Sensibilidade	x^2	0,9465	0,9055	0,9505	0,9120	0,9555	0,9150	0,9674	0,9129
	x^3	0,9437	0,9053	0,9525	0,9124	0,9545	0,9128	0,9636	0,9273
	$x^{1/2}$	0,9626	0,9499	0,9686	0,9541	0,9738	0,9600	0,9791	0,9636
	$x_i x_j$	0,9614	0,9609	0,9717	0,9648	0,9757	0,9690	**0,9818**	0,9669
Especificidade	x^2	0,9738	0,9872	0,9728	0,9875	0,9747	0,9864	0,9745	**0,9898**
	x^3	0,9739	0,9865	0,9747	0,9872	0,9763	0,9874	0,9760	0,9901
	$x^{1/2}$	0,9740	0,9786	0,9753	0,9786	0,9751	0,9800	0,9805	0,9817
	$x_i x_j$	0,9767	0,9750	0,9774	0,9768	0,9801	0,9773	0,9818	0,9825
Valor Predito Positivo	x^2	0,9508	0,9742	0,9486	0,9746	0,9524	0,9727	0,9511	0,9785
	x^3	0,9509	0,9725	0,9528	0,9744	0,9540	0,9749	0,9546	**0,9797**
	$x^{1/2}$	0,9522	0,9597	0,9548	0,9601	0,9540	0,9624	0,9628	0,9637
	$x_i x_j$	0,9566	0,9536	0,9580	0,9574	0,9625	0,9571	0,9659	0,9673
Valor Predito Negativo	x^2	0,9710	0,9508	0,9731	0,9536	0,9751	0,9551	0,9815	0,9520
	x^3	0,9694	0,9507	0,9735	0,9535	0,9740	0,9535	0,9781	0,9599
	$x^{1/2}$	0,9793	0,9726	0,9820	0,9743	0,9849	0,9768	0,9874	0,9778
	$x_i x_j$	0,9787	0,9784	0,9840	0,9798	0,9856	0,9828	**0,9882**	0,9813

Fonte: elaborado pelos autores.

**Neste Kernel, todas as variáveis foram cruzadas duas a duas gerando 36 variáveis. Além disso, foram inseridas todas as variáveis originais e todas as variáveis ao quadrado totalizando 54 variáveis.*

Nas Tabelas 4.4 e 4.5 é apresentado o percentual de incidência de cada variável nos conjuntos selecionados nas repetições das amostragens realizadas nas diferentes proporções N_{tr}/N_{ts} do banco de dados utilizando o *kernel* polinomial tipo $x_i x_j$, o qual apresentou o melhor resultado para KVP e AD, respectivamente. As variáveis 9, 18, 39 e 48, para KVP, e

9, 18, 26, 33, 39, 44, 48 e 51, para AD, foram retidas com maior frequência, independente da proporção N_{tr}/N_{ts}.

Tabela 4.4 - Percentual de incidência de cada variável nos conjuntos selecionados para as proporções testadas utilizando KVP

Porções do Banco de Dados (% treino / % teste)							
50% / 50%		70% / 30%		80% / 20%		90% / 10%	
variável	Incidência (%)	variável	Incidência (%)	variável	Incidência (%)	variável	Incidência (%)
9	100	9	100	9	100	9	100
18	100	18	100	18	100	18	100
39	100	39	100	39	100	48	100
48	100	48	100	48	100	33	99,5
26	99,5	54	100	54	100	54	97,0
33	99,5	33	99,5	33	99,5	39	96,0
54	98,5	26	98,0	26	98,0	26	92,0
44	96,0	44	95,0	44	95,0	44	80,5
53	93,5	53	95,0	53	95,0	53	80,5
51	93,0	51	91,5	51	91,5	51	74,5
49	80,0	49	83,0	49	83,0	49	63,5
6	70,5	6	69,5	6	69,5	6	48,5
7	67,0	36	63,5	36	63,5	7	44,0
15	67,0	15	63,0	15	63,0	15	40,5
23	66,5	7	62,5	7	62,5	24	40,0
16	66,0	37	62,5	37	62,5	23	37,0
24	66,0	24	61,5	24	61,5	36	35,5
37	66,0	23	61,0	23	61,0	30	35,0
30	65,0	30	60,5	30	60,5	37	35,0
36	65,0	16	59,0	16	59,0	16	34,5
31	62,5	31	55,0	31	55,0	31	29,5
3	61,0	3	54,0	3	54,0	3	26,0
2	55,5	2	47,5	2	47,5	41	22,0
41	54,5	41	46,5	41	46,5	42	22,0
42	53,0	42	43,0	42	43,0	2	21,5
19	51,5	19	40,5	19	40,5	19	19,5
20	50,5	20	40,5	20	40,5	20	19,5
27	47,5	45	37,5	45	37,5	27	16,0
12	45,0	12	36,0	12	36,0	12	15,5
50	44,5	27	35,5	27	35,5	50	15,5
45	42,0	50	32,5	50	32,5	45	14,0
1	38,0	1	31,5	1	31,5	4	12,5
52	38,0	11	30,0	11	30,0	52	12,5
4	35,0	46	28,5	46	28,5	1	12,0
11	35,0	21	28,0	21	28,0	47	12,0
34	35,0	4	27,0	4	27,0	11	11,0
21	34,5	10	27,0	10	27,0	21	11,0
46	33,0	52	27,0	52	27,0	46	10,5
47	33,0	47	26,5	47	26,5	10	10,0
10	31,5	34	22,5	34	22,5	34	9,00

28	30,0	28	20,5	28	20,5	28	8,50
29	23,5	14	13,0	14	13,0	14	8,00
14	22,5	40	11,5	40	11,5	13	7,00
35	20,5	35	9,50	35	9,50	8	6,50
40	19,5	22	9,00	22	9,00	29	6,00
32	19,0	25	9,00	25	9,00	38	5,50
22	18,0	29	9,00	29	9,00	25	5,50
13	17,5	38	8,50	38	8,50	35	5,50
25	17,5	5	7,50	5	7,50	40	5,50
38	17,5	13	7,50	13	7,50	5	3,50
5	15,5	8	7,00	8	7,00	43	3,50
8	15,0	43	7,00	43	7,00	32	3,00
43	15,0	32	6,00	32	6,00	22	2,00
17	7,50	17	3,00	17	3,00	17	0,50

Fonte: elaborado pelos autores.

Tabela 4.5 - Percentual de incidência de cada variável nos conjuntos selecionados para as proporções testadas utilizando AD

Porções do Banco de Dados (% treino / % teste)							
50% / 50%		70% / 30%		80% / 20%		90% / 10%	
variável	Incidência (%)	variável	incidência (%)	variável	incidência (%)	variável	incidência (%)
9	100	9	100	9	100	9	100
18	100	18	100	18	100	18	100
26	100	26	100	26	100	33	100
33	100	33	100	33	100	39	100
39	100	39	100	39	100	48	100
44	100	44	100	48	100	54	100
48	100	48	100	53	100	26	99,5
51	100	51	100	54	100	44	98,5
53	100	53	100	44	99,5	53	98,5
54	100	54	100	51	99,0	51	98,0
49	99,5	49	99,5	49	98,5	49	94,5
6	97,0	6	98,0	6	95,5	6	86,0
7	95,5	7	97,5	7	92,5	7	79,0
24	95,5	15	95,0	15	88,5	15	78,5
30	94,5	24	95,0	24	86,5	24	72,5
36	93,5	30	93,0	23	84,0	23	65,5
37	92,5	16	92,5	36	84,0	36	63,5
15	92,0	37	92,5	30	82,0	30	63,0
3	90,0	36	92,0	37	82,0	37	59,5
16	90,0	23	88,5	16	81,5	16	59,0
23	88,0	31	83,5	31	73,0	3	51,5
31	86,5	3	82,0	3	68,0	31	51,0
2	73,0	41	70,0	41	60,0	41	42,5
41	68,5	2	68,5	2	57,5	2	38,0

42	66,0	42	61,5	42	53,5	42	35,5
19	60,0	19	53,0	20	46,5	19	29,5
20	58,5	20	53,0	19	45,5	20	28,0
27	54,5	27	49,0	27	37,0	45	27,5
52	48,0	12	45,0	45	37,0	27	27,0
45	47,5	45	44,5	4	34,0	12	24,0
12	46,5	50	42,5	12	34,0	4	23,0
50	45,5	52	42,0	21	31,0	50	23,0
4	43,0	21	36,5	50	31,0	21	22,5
21	42,0	4	36,0	1	27,5	47	22,5
11	41,0	11	34,0	52	27,5	52	21,5
10	40,5	1	32,5	47	26,0	11	20,0
1	39,5	47	31,0	11	25,0	1	19,5
34	37,0	10	28,0	10	23,5	34	17,5
47	36,5	46	26,0	34	22,5	46	17,5
46	33,0	34	24,5	46	21,0	28	16,5
28	32,0	28	20,0	28	18,5	10	16,0
14	23,0	8	18,0	14	13,0	14	13,5
40	23,0	14	16,5	8	12,5	8	12,5
8	22,5	25	15,0	40	12,5	25	9,50
29	21,0	40	15,0	29	11,0	40	9,50
25	20,5	38	11,0	25	9,00	38	8,50
38	18,0	29	10,5	35	8,50	29	7,00
43	18,0	22	10,0	13	8,00	13	4,50
22	17,0	13	9,50	43	7,50	22	4,50
35	16,5	35	9,00	5	7,00	35	4,50
32	16	32	8,50	38	7,00	32	3,00
5	13,5	5	6,50	22	4,50	5	2,50
13	12,5	43	6,50	32	4,00	43	2,50
17	10,0	17	4,50	17	2,00	17	1,00

Fonte: elaborado pelos autores.

4.5 Conclusões

O diagnóstico precoce aumenta as taxas de sobrevivência em pacientes com CM, justificando o grande número de propostas com o objetivo de classificar corretamente os nódulos mamários em duas classes, benigno e maligno.

O método proposto neste artigo seleciona as variáveis mais relevantes para fins de classificação de forma a maximizar a sua acurácia, além de propor o teste de dois métodos de classificação na análise de um banco de dados. Para melhorar o desempenho de classificação,

quatro diferentes tipos de função *kernel* (multiplicativo, quadrático, cúbico e de raiz quadrada) são utilizados para adensar o banco de dados original.

As proposições são testadas no banco de dados WBCD. Inicialmente as variáveis são ordenadas utilizando um novo índice de importância baseado nos pesos da ACP e na variância explicada por cada componente retido. Posteriormente, o método proposto classifica iterativamente os registros dos pacientes em duas classes, benigno e maligno, através de dois métodos de mineração de dados, KVP e AD; a variável menos importante é removida e a classificação realizada nas variáveis restantes até restar uma única variável.

O método proposto, para uma proporção de 90% /10%% do banco de dados utilizando o *kernel* polinomial tipo $x_i x_j$, o qual apresentou o melhor resultado para KVP e AD, respectivamente, classificou corretamente os dados do WBCD em 98,09% dos casos, em média, utilizando uma média de 17,24 variáveis (em um total de 54 variáveis) na classificação utilizando o método KVP. O melhor desempenho para a sensibilidade foi de 0,9818 utilizando o método KVP, e o melhor desempenho para especificidade foi de 0,9898 utilizando o método AD. É importante observar que, para o rastreamento de câncer de mama, o método deve ser o mais sensível possível para que se consiga detectar o maior número de casos possível da doença.

Desenvolvimentos futuros incluem testes de outros *kernels* e de outros métodos de mineração de dados, como Máquina de Suporte Vetorial.

4.6 Referências

ALBRECHT, A. A.; Lappas, G.; Vinterbo, S. A.; Wong, C. K.; Ohno-Machado, L. Two applications of the LSA macrine. In: **Proceedings of the 9th International Conference on Neural Information Processing**, p. 184-189, 2002.

ANZANELLO, M. J.; FOGLIATTO, F. S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preference**, v. 22, p. 139-148, 2011.

ANZANELLO, M. J.; FOGLIATTO, F. S. A review of recent variable selection methods in industrial and chemometrics applications. **European Journal of Industrial Engineering, no prelo**, 2012.

BOSER, B.; GUYON, I. M.; VAPNK, V. A training algorithm for optimal margin classifiers. In: **Proceedings of the 15th Annual Workshop on Computational learning Theory**. ACM. Pittsburg, 1992.

BRASIL. Ministério da Saúde. Programa Nacional de Controle do Câncer de Mama. Versão revista e ampliada do Programa Viva Mulher, desmembrado em Programa Nacional de Controle do Câncer do Colo do Útero e Programa Nacional de Controle do Câncer de Mama (INCA, 2010), elaborado pela Divisão de Apoio à Rede de Atenção Oncológica em abril de 2011. Brasília: Ministério da Saúde, 2011. 15p http://www2.inca.gov.br/wps/wcm/connect/521d4900470039c08bd8fb741a182d6f/pncc_mama.pdf?MOD=AJPERES&CACHEID=521d4900470039c08bd8fb741a182d6f. Acesso em 16/12/2012.

BRASIL. Ministério da Saúde. Instituto Nacional de Câncer (INCA). Controle do Câncer de Mama. Documento de Consenso. ***INCA*** [site na Internet]. 2004 Abr [acessado 2012 jul 23]; [cerca de 39 p.]. Disponível em: http://www1.inca.gov.br/publicacoes/Consensointegra.pdf

DAMIAN, E. A. Duas metodologias aplicadas à classificação de precipitação convectiva e estratiforme com radar meteorológico: SVM e K-means. Dissertação de Mestrado em Ciências. Programa de Pós-Graduação em Métodos Numéricos em Engenharia. Área de Concentração em Programação Matemática. Setores de Tecnologia e Ciências Exatas da Universidade Federal do Paraná. Curitiba. Setembro 2011.

DUDA, R.; HART, P.; STORK, D. **Pattern Recognition**. 2° ed. New York: Willey, 2001.

FERREIRA, V. H.; LAZZARETTI, A.; NETO, H. V.; RIELLA, R.; OMORI, J. Classificação de eventos em redes de distribuição de energia utilizando transformada Wavelet

e modelos neurais autônomos. **Learning and Nonlinear Models (L&NLM) Journal of the Brazilian Neural Networks Society**, vol. 8, iss. 2, pp. 93-99, 2010.

GUYON, I.; BOSER, B.; VAPNIK, V. Automatic capacity tuning of very large VC-dimension classifiers. In Hanson, S. J., Cowan, J.D., Lee Giles, C., editors, In: **Proceedings of the Advances in Neural Information processing Systems**, San Mateo, CA, v.5, p. 147-155, 1993.

International Agency for Research on Cancer (IARC). IARC Handbooks of Cancer Prevention Volume 7: Breast Cancer Screening. Lyon: ***IARC*** [serial on the Internet] 2002 [cited 2012 jul 23]; [about 243 p.]. Available from: http://www.iarc.fr/en/ publications/pdfs-online/prev/handbook7/index.php

MARCONDES, D. F. Cartas de controle multivariadas baseadas no método *Kernel-Statis* para monitoramento de processos em bateladas. Universidade Federal do Rio Grande do Sul. Escola de Engenharia. Programa de Pós-Graduação em Engenharia de Produção. Tese de Doutorado. Porto Alegre. 2009.

MONTGOMERY D.; PECK E.; VINING G. Introduction to Linear Regression Analysis. 3° ed. New York: Willey; 2001.

NALDI, M. C.; CARVALHO, A. C. P. L. F. Utilização de Algoritmos de Aprendizado de Máquina Evolutivos para Análise de Nível Expressão Gênica. **XXV Congresso da Sociedade Brasileira de Computação**. 22 a 29 de julho 2005 – São Leopoldo/RS.

SANTOS, E. M. Teoria e Aplicação de *Support Vector Machines* à Aprendizagem e Reconhecimento de Objetos Baseados na Aparência. Curso de Pós-Graduação em Informática da Universidade Federal da Paraíba. Dissertação de Mestrado. Universidade Federal da Paraíba, 2002.

SCHOLKOPF, B.; SMOLA, J. Learning with *kernels*. England: The MIT Press, 2002.

SEMOLINI, R. Support Vector Machines, Inferência Transdutiva e o Problema de Classificação. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. Departamento de Engenharia de Computação e Automação Industrial. Tese de Mestrado. Campinas - SP – Brasil. Dezembro de 2002.

SILVA, B. C.; FONSECA, E. S.; OLESKOVICZ, M.; ROMERO, R. A. F. Aplicação de SVMS para o processo de classificação de distúrbios relacionados a qualidade da energia elétrica. XVIII Congresso Brasileiro de Automática. 12 a 16-setembro-2010, Bonito - MS.

World Health Organization (WHO). Cancer control: knowledge into action: WHO guide for effective programmes: early detection. ***WHO 2007*** [serial on the Internet] 2007 [cited 2012 Aug 09]; [about 50 p.]. Available from: http://www.who.int/cancer/publications/cancer_control_detection/en/

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.-H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems,** v. 14, p. 1–37, 2008.

5 Considerações Finais

Este capítulo apresenta as conclusões da dissertação e sugestões para trabalhos futuros.

5.1 Conclusões

A presente dissertação teve como principal objetivo propor métodos para mineração de dados para diagnóstico de câncer de mama baseado na seleção de variáveis.

O primeiro artigo teve como objetivo apresentar a revisão sistemática dos principais métodos de seleção de variáveis para fins de classificação de pacientes a partir de diagnósticos médicos. Neste artigo, verificou-se que as principais abordagens de seleção de variáveis para predição utilizam testes de significância para os coeficientes. As variáveis são selecionadas utilizando os métodos *Forward, Backward, Stepwise, Baysesian* e *Recursive Bootstrap Elimination* para determinar o melhor modelo. Nas abordagens de seleção de variáveis para classificação, índices de acerto na classificação são utilizados como critério para inserção de uma variável no modelo, e a seleção de variáveis é realizada utilizando os métodos *Network Prunning*, *F-Score*, *LASSO* e *Random Forest.*

No segundo artigo, o método proposto selecionou as variáveis mais relevantes para fins de classificação de forma a maximizar a sua acurácia, além de propor o teste de dois métodos de classificação na análise do banco de dados WBCD. O método proposto classificou corretamente 97,77% dos casos, em média, utilizando uma média de 5,8 variáveis através da técnica KVP. O melhor desempenho para a sensibilidade foi de 97,90% utilizando o método KVP, e o melhor desempenho para especificidade foi de 98,56% utilizando o método AD. É importante ressaltar que, para o rastreamento de câncer de mama, o método deve ser o mais sensível possível para que se consiga detectar o maior número de casos da doença.

O terceiro e último artigo consolida o trabalho desenvolvido nos dois primeiros artigos ao sugerir uma variação do método proposto no segundo artigo. O objetivo dessa variação do método foi melhorar o desempenho das técnicas de classificação. O método proposto utilizou o *kernel* polinomial tipo $x_i x_j$, o qual apresentou o melhor resultado para KVP e AD, respectivamente. O método classificou corretamente os dados do WBCD em 98,09% dos casos, em média, utilizando uma média de 17,24 variáveis (de um total de 54

variáveis) e método KVP. O melhor desempenho para a sensibilidade foi de 98,18% utilizando o método KVP, e o melhor desempenho para especificidade foi de 98,98% aplicando-se a AD. Sendo assim, os resultados permitem inferir que o método proposto no terceiro artigo é o mais recomendado por gerar a maior acurácia de classificação com base no menor número de variáveis retidas.

Portanto, as considerações acima permitem afirmar que todos objetivos específicos e o objetivo principal desta pesquisa foram atingidos.

5.2 Sugestões para trabalhos futuros

Sugerem-se as seguintes pesquisas futuras:

a) Ampliação da base de dados para diferentes populações;

b) Ampliação da base de dados em número de observações e para outros tipos de patologia;

c) Desenvolvimento de novos índices de importância das variáveis e sua integração a outros métodos de mineração de dados;

d) Aplicar o método a outro banco de dados.