DOI: 10.1590/1413-81232014194.01722013

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

A data mining method for breast cancer identification based on a selection of variables

Nicole Holsbach [1]
Flávio Sanson Fogliatto [1]
Michel Jose Anzanello [1]

Abstract *In the majority of countries, breast cancer among women is highly prevalent. If diagnosed in the early stages, there is a high probability of a cure. Several statistical-based approaches have been developed to assist in early breast cancer detection. This paper presents a method for selection of variables for the classification of cases into two classes, benign or malignant, based on cytopathological analysis of breast cell samples of patients. The variables are ranked according to a new index of importance of variables that combines the weighting importance of Principal Component Analysis and the explained variance based on each retained component. Observations from the test sample are categorized into two classes using the k-Nearest Neighbor algorithm and Discriminant Analysis, followed by elimination of the variable with the index of lowest importance. The subset with the highest accuracy is used to classify observations in the test sample. When applied to the Wisconsin Breast Cancer Database, the proposed method led to average of 97.77% in classification accuracy while retaining an average of 5.8 variables.*

Key words *Selection of variables, Breast cancer identification, K-nearest neighbor algorithm (KNN), Discriminant analysis*

Resumo *Na maioria dos países, o câncer de mama 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 célula 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 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 das ferramentas k-vizinhos mais próximos e 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. Aplicando ao* Wisconsin Breast Cancer Database, *o método proposto apresentou uma média de 97,77% de acurácia de classificação, retendo uma média de 5,8 variáveis.*

Palavras-chave *Seleção de variáveis, Identificação de câncer de mama, k-vizinhos mais próximos, Análise Discriminante*

[1] Escola de Engenharia, Universidade Federal do Rio Grande do Sul. Av. Osvaldo Aranha 99/5º, Centro. 90.035-190 Porto Alegre RS Brasil. nicole.holsbach@bol.com.br

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 de mama (CM) na grande maioria dos países; o de colo do útero na África e no Sul da Ásia; e o de próstata na América do Norte, Oceania, Norte da Europa e Europa Ocidental[1]. O CM é o segundo tipo mais frequente no mundo e é o mais comum entre as mulheres, respondendo por 22% dos casos novos a cada ano. Se diagnosticado precocemente, apresenta uma alta percentagem de cura[2,3]. 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[4]. 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[5]. A identificação 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.[6] e Humphrey et al.[7].

A identificação 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). Em seu trabalho, Baker[8] 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.[9] e Fogel et al.[10]. 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*[11]. 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

Ciência & Saúde Coletiva, 19(4):1295-1304, 2014

de combinações lineares de variáveis altamente correlacionadas[12]. 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 deste 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.

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 a identificação de câncer de mama que não utilizam o WBCD foram revisadas por Eltoukhy et al.[13]; abordagens para seleção de variáveis em problemas de classificação foram revisadas por Dash e Liu[14].

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 no Quadro 1, apresentado no final da seção. No quadro são listados os trabalhos que utilizaram o WBCD para fins similares ao do presente artigo, a acurácia percentual obtida e o desvio-padrão da acurácia reportada.

No seu trabalho, Street et al.[9] relatam análises preliminares realizadas no WBCD com o objetivo de organizar o banco de dados. Os autores classificam com uma amostra de 569 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. A acurácia obtida foi de 97,30% em um procedimento de validação cruzada do tipo *ten-fold*, no qual a amostra é dividida em 10 porções iguais, cada uma delas usada como porção de teste enquanto as 9 restantes são usadas como porções de treino. Os autores não informam se o valor de acurácia diz respeito à média ou ao valor máximo obtidos nas classificações; o desvio-padrão da acurácia também não é informado. Por usarem uma fração do WBCD no procedimento de classificação, o resultado em Street et al.[9] não é diretamente comparável aos demais apresentados nesta seção, não sendo incluído no Quadro 1.

Já Fogel et al.[10] 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%. O valor reportado de acurácia, entretanto, é tendencioso, já que a porção de treino utilizada nos experimentos com redes sempre consistiu das primeiras 400 observações do WBCD. Com propósitos semelhantes, Quinlan[15,16] sugere um classificador baseado em árvore de decisão, que melhora o desempenho do classificador C4.5 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[17] 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 à 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 acurácia média máxima de classificação, na porção de teste, de 93,87% sobre todas as redes testadas. O mesmo algoritmo foi expandido para incluir um estágio de pré-processamento do classificador da rede[18]. O estágio adicional é realizado em dois passos. No primeiro, os ca-

sos com valores desconhecidos são removidos do banco de dados. No segundo, 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 média (96,71%) é obtida quando a rede é treinada para 98% de acurácia na porção de treino, utilizando 50% das observações na porção de treino.

Em sua pesquisa, Peña-Reyes e Sipper[19] combinaram sistemas *fuzzy* e algoritmos evolucionários em uma ferramenta de identificação. 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. Em seu trabalho, Nauck e Kruse[20] 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. Da mesma forma, Lee et al.[21] 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 de Shannon[22]. 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%.

É proposto por Albrecht et al.[23] 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 destas 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%.

É apresentado por Abbass[24] 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.

Em seu trabalho, Verikas e Bacauskiene[25] 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, à medida que elas são removidas; o objetivo é encontrar a melhor solução de compromisso entre erro e número das 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%.

Já em Abonyi e Szeifert[11] é apresentado 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 delas é estimada com base em sua matriz de covariâncias. A acurácia média encontrada foi de 95,57%, em uma divisão 50%/50% do WBCD.

Em sua pesquisa, Polat e Günes[26] 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. Também em Akay[27] é proposto 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[28], 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.

Ciência & Saúde Coletiva, 19(4):1295-1304, 2014

Quadro 1. Informações de acurácia de classificação obtida no WBCD em diferentes métodos disponíveis na literatura.

Fonte	Método	Acurácia (%)	Desvio-padrão
Fogel et al.[10]	RN	98,05	0,465
Quinlan [15]	ADD/PL	94,74	0,285
Setiono [17]	RN	93,87	1,160
Setiono [18]	RN	96,71	0,570
Peña-Reyes e Sipper [19]	TF	97,80	NI
Nauck e Kruse [20]	RN e TF	95,06	2,535
Lee et al. [21]	TF	95,14	NI
Albrecht et al. [23]	ADD/PL	98,80	NI
Abbass [24]	RN	98,10	0,005
Verikas e Bacauskiene [25]	RN	96,44	0,310
Abonyi e Szeifert [11]	TF	95,57	2,143
Polat e Günes [26]	E/MSV	98,53	NI
Akay [27]	E/MSV	99,51	NI
Marcano-Cedeño et al. [29]	RN	99,26	NI

NI = Não informado

Por fim, Marcano-Cedeño et al.[29] 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. Das observações do WBCD, 60% 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 foi de 99,26%.

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 treino baseado nessas variáveis. Esses passos operacionais estão detalhados na sequência.

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 ($r = 1, ..., R$), λ^r. O número

de componentes a serem retidos pode ser definido com base na variância acumulada, conforme sugerido em Montgomery et al.[30].

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 denotado por v_j, $j = 1, ..., J$. Quanto maior o valor de v_j, mais importante é a variável *j* na categorização das observações em classes.

O índice é gerado baseado nos pesos da ACP (p_{jr}) e no percentual de variância explicado por cada componente retido (λ^r); ver equação (1). 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[31]. Um índice similar é proposto por Anzanello et al.[32], mas não leva em consideração o percentual da variância explicada por cada componente retido.

$$v_j = \Sigma_{r=1}^{R} \mid p_{jr} \mid \lambda_r, j = 1, ..., J \qquad (1)$$

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[32].

Concluída a primeira das classificações, calcular a acurácia delas, definidas como a proporção das corretas relativamente ao total das 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. Esse procedimento é repetido removendo a próxima variável com menor valor de v_j e aplicando KVP e AD nas remanescentes, até restar uma única.

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 definidas a seguir. 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 classifica-

Ciência & Saúde Coletiva, 19(4):1295-1304, 2014

ções 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 (2), corresponde à fração de casos positivos corretamente classificados; a especificidade, dada pela equação (3), corresponde à fração de casos negativos corretamente classificados.

$$\text{Sensibilidade} = \frac{PV}{PV + NF} \qquad (2)$$

$$\text{Especificidade} = \frac{NV}{NV + PF} \qquad (3)$$

Resultados

O WBCD é composto por 699 observações (16 delas incompletas) obtidas a partir da aspiração com agulha fina de células da mama. A aspiração com agulha fina permite a investigação da malignidade em nódulos mamários[23]. Nove variáveis foram analisadas em cada amostra de células da mama, utilizando uma escala de valores inteiros de 10 pontos (Quadro 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[33].

Para cada proporção, 1000 repetições foram executadas em grupos de treino e teste obtidos amostrando aleatoriamente as observações do WBCD. O Quadro 3 apresenta a média e o desvio-padrão médio da acurácia, sensibilidade e especificidade de classificação para diferentes proporções N^{tr}/N^{ts} utilizando as variáveis selecionadas em cada repetição e valendo-se das técnicas de classificação KVP e AD. O método proposto utilizando KVP atinge a maior acurácia média de classificação, 97,77%, ao reter 5,87 variáveis, em média, e o melhor desempenho para a sensibilidade, com uma média de 97,90%. 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 98,56%. O método KVP apresenta maior acurácia retendo um menor número de variáveis em comparação ao método AD para todas as proporções testadas. Além disso, o método KVP apresentou o melhor desempenho para sensibilidade e o pior 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. Percebe-se um aumento na variabilidade das medidas de precisão com a redução do número de observações na porção de teste. A acurácia média, extraída de diversas repetições com distintas formações nas porções de treino e teste, é uma medida de desempenho de classificação mais confiável que a acurácia estimada sobre uma única partição treino/teste. Executando uma única repetição do método classificatório em uma porção favorável do banco de dados pode levar a resultados não confiáveis.

No Quadro 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. Há uma pequena variação no número de variáveis responsável pela máxima acurácia (esta 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.

Para uma melhor visualização dos resultados de classificação, uma matriz de confusão é apresentada no Quadro 5. O pequeno número de erros de classificação, particularmente na pro-

Quadro 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

Quadro 3. Medidas de desempenho e desvio-padrão das medidas de desempenho de classificação e variáveis retidas médias para porção de teste utilizando os métodos KVP e AD.

Medidas	Número de observações na porção de treino/número de observações porção de teste (%porção de treino/%porção de teste)							
	342/341 (50%/50%)		478/205 (70%/30%)		546/137 (80%/20%)		615/68 (90%/10%)	
	KVP	AD	KVP	AD	KVP	AD	KVP	AD
Média da Acurácia	0,9702	0,9642	0,9702	0,9642	0,9702	0,9642	0,9702	0,9642
Desvio-padrão	0,0069	0,0096	0,0069	0,0096	0,0069	0,0096	0,0069	0,0096
Média do N° variáveis retidas	7,15	7,18	7,15	7,18	7,15	7,18	7,15	7,18
Desvio-padrão	1,41	1,73	1,41	1,73	1,41	1,73	1,41	1,73
Média da Sensibilidade	0,9593	0,9317	0,9593	0,9317	0,9593	0,9317	0,9593	0,9317
Desvio-padrão	0,0182	0,0233	0,0182	0,0233	0,0182	0,0233	0,0182	0,0233
Média da Especificidade	0,9766	0,9821	0,9766	0,9821	0,9766	0,9821	0,9766	0,9821
Desvio-padrão	0,0094	0,0062	0,0094	0,0062	0,0094	0,0062	0,0094	0,0062

Quadro 4. Inclusão das variáveis nos subgrupos retidos para as proporções testadas.

Número de observações (treino/teste)							
342/341		478/205		546/137		615/68	
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

porção de 615 observações na porção de treino e 68 na de teste, corrobora o desempenho satisfatório do método.

Conclusão

Neste artigo, propõe-se um método para seleção de variáveis oriundas de exames clínicos com vistas à classificação de observações em categorias distintas. O método congrega três técnicas de

Ciência & Saúde Coletiva, 19(4):1295-1304, 2014

Quadro 5. Matriz de confusão para as proporções testadas.

Predito	Real		Número de observações do banco de dados (treino / teste)
	Benigno	Maligno	
Benigno Maligno	212,1 9,2	4,7 115,1	342 / 341
Benigno Maligno	130,4 3,4	2,4 68,7	478/205
Benigno Maligno	86,9 2,2	1,3 46,6	546/137
Benigno Maligno	43,6 0,95	0,51 23,9	615/68

análise estatística multivariada: a Análise de Componentes Principais (ACP), aplicada na obtenção de um índice de importância para as variáveis, e as análises de *clusters* e discriminante, usadas na classificação das observações contidas nos bancos de dados abordados pelo método.

O método proposto seleciona as variáveis mais relevantes para fins de classificação de forma a maximizar a sua acurácia, além de propor o teste dos dois métodos de classificação citados anteriormente 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 duas técnicas 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 possível de casos 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.

Colaboradores

N Holsbach, FS Fogliatto e MJ Anzanello participaram igualmente de todas as etapas de elaboração do artigo.

Referências

1. Bray F, Ren JS, Masuyer E, Ferlay J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer* 2013; 132(5):1133-1145. International Agency for Research on Cancer (IARC).
2. IARC Handbooks of Cancer Prevention. Vol. 7. *Breast Cancer Screening*. Lyon: *IARC*; 2002.
3. World Health Organization (WHO). *Cancer control: knowledge into action: WHO guide for effective programmes: early detection*. WHO 2007 [Internet] 2007 [acessado 2012 Ago 9]; [cerca de 50 p.]. Disponível em: http://www.who.int/cancer/publications/cancer_control_detection/en/
4. Brasil. Ministério da Saúde (MS). Departamento de Informática do SUS (Datasus). *Informações de saúde. Indicadores de saúde*. [site da Internet]. [acessado 2011 maio 3]. Disponível em: http://tabnet.datasus.gov.br/cgi/dh.exe?pacto/2010/cnv/pactbr.def
5. Brasil. Ministério da Saúde (MS). 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
6. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten-to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982; 69(2):349-355.
7. Humphrey LL, Helfand M, Chan BKS, Woolf SH. Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med* 2002; 137(5 Part 1):347-360.
8. Baker LH. Breast cancer detection demonstration Project: five-year summary report. *CA Cancer J Clin* 1982; 32(4):194-225.
9. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: *ISandT/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology* 1993; 1905:861-870, San Jose, California.
10. Fogel DB, Wasson III EC, Boughton EM. Evolving neural networks for detecting breast cancer. *Cancer Letters* 1995; 96:49-53.
11. Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters* 2003; 14:2195-2207.
12. Rencher R. *Methods of multivariate Analysis*. 1° ed. New York: Wiley; 1995.
13. Eltoukhy MM, Faye I, Samir BB. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. *Computers in Biology and Medicine* 2012; 42:123-128.
14. Dash M, Liu H. Feature selection for classification. *Intelligent Data Analysis* 1997; 1:131-156.
15. Quinlan JR. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* 1996; 4:77-90.
16. Quinlan JR. *C4.5: Programs for machine learning*. 5° ed. San Mateo: Morgan Kaufmann; 1993.
17. Setiono R. Extracting rules from pruned neural networks for breast cancer diagnosis. *Artificial Intelligence in Medicine* 1996; 8:37-51.
18. Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine* 2000; 18:205-217.
19. Peña-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine* 1999; 17:131-155.
20. Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. *Artificial Intelligence in Medicine* 1999; 16:149-169.
21. 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* 2001; 31(3):426-432.
22. Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal* 1948; 27:379-423.
23. Albrecht AA, Lappas G, Vinterbo SA, Wong CK, Ohno-Machado L. Two applications of the LSA macrine. In: *Proceedings of the 9th International Conference on Neural Information Processing*; 2002 Nov 18-22; Singapore. p. 184-189.
24. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine* 2002; 25:265-281.
25. Verikas A, Bacauskiene M. Feature selection with neural networks. *Pattern Recognition Letters* 2002; 23:1323-1335.
26. Polat K, Günes S. Breast cancer diagnosis using a least square support vector machine. *Digital Signal Processing* 2007; 17:694-701.
27. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications* 2009; 36:3240-3247.
28. Chen YW, Lin CJ. Combining SVMs with various feature selection strategies. *Studies in Fuzziness and Soft Computing* [série da Internet] 2006 [acessado 27 Jan 2012]; 207: [cerca de 9 p.]. Disponível em: http://www.csie.ntu.edu.tw/~cjlin/papers/"features.pdf
29. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina A. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications* 2011; 38:9573-9579.
30. Montgomery D, Peck E, Vining G. *Introduction to Linear Regression Analysis*. 3rd Edition. New York: Willey; 2001
31. Duda R, Hart P, Stork D. *Pattern Recognition*. 2nd Edition. New York: Willey; 2001.
32. Anzanello MJ, Fogliatto FS, Rossini K. Data mining-based method for identifying discriminant attributes in sensory profiling. *Food Quality and Preference* 2011; 22(1):139-148.
33. UC Irvine Machine Learning Repository. Center for Machine Learning and Intelligent Systems. [banco de dados na Internet]. [acessado 2014 mar 10]. Disponível em: http://www.ics.uci.edu/~mlearn/MLRepository

Artigo apresentado em 21/03/2013
Aprovado em 29/04/2013
Versão final apresentada em 15/05/2013