v.13, n.2, p.325-337, mai.-ago. 2006

ABORDAGEM DE UM PROBLEMA MÉDICO POR MEIO DO PROCESSO DE *KDD* COM ÊNFASE À ANÁLISE EXPLORATÓRIA DOS DADOS

Maria Teresinha Arns Steiner

Departamento de Matemática, UFPR,
C. P. 19081, CEP 81531-990, Curitiba, PR,
e-mail: tere@mat.ufpr.br

Nei Yoshihiro Soma

Divisão da Ciência da Computação, ITA, Pça. Mal. Eduardo Gomes, 50,
Vila das Acácias, CEP 12228-990, São José dos Campos, SP,
e-mail: nysoma@comp.ita.br

Tamio Shimizu

Gestão da Tecnologia da Informação,
Departamento de Engenharia de Produção, USP,
Av. Prof. Almeida Prado, 531, 2º andar, Butantã, CEP 05508-900, São Paulo, SP,
e-mail: tmshimiz@usp.br

Júlio Cesar Nievola

Programa de Pós-Graduação em Informática Aplicada, PUC-PR,
Av. Imaculada Conceição, 1155, CEP 80215-901, Curitiba, PR,
e-mail: nievola@ppgia.pucpr.br

Pedro José Steiner Neto

Departamento de Administração, UFPR,
C. P. 19081, CEP 81531-990, Curitiba, PR,
e-mail: pedrosteiner@ufpr.br

Recebido em 14/2/2005
Aceito em 11/4/2006

Resumo

A "Descoberta de Conhecimento em Bases de Dados" (Knowledge Discovery in Databases, KDD) é um processo composto de várias etapas, iniciando com a coleta de dados para o problema em pauta e finalizando com a interpretação e avaliação dos resultados obtidos. O presente trabalho objetiva mostrar a influência da análise exploratória dos dados no desempenho das técnicas de Mineração de Dados (Data Mining) quanto à classificação de novos padrões por meio da sua aplicação a um problema médico, além de comparar o desempenho delas entre si, visando obter a técnica com o maior percentual de acertos. Pelos resultados obtidos, pode-se concluir que a referida análise, se conduzida de forma adequada, pode trazer importantes melhorias nos desempenhos de quase todas as técnicas abordadas, tornando-se, assim, uma importante ferramenta para a otimização dos resultados finais. Para o problema em estudo, a técnica que envolve um modelo de Programação Linear e uma outra que envolve Redes Neurais foram as técnicas que apresentaram os menores percentuais de erros para os conjuntos de testes, apresentando capacidades de generalização satisfatórias.

Palavras-chave: mineração de dados, processo KDD, análise exploratória dos dados.

1. Introdução

Abordar técnicas e ferramentas que buscam transformar os dados armazenados, sejam de indústrias, bancos, hospitais, telecomunicações, ecológicos, imobiliários e outros, em conhecimento é o objetivo da área denominada Descoberta de Conhecimentos em Bases de Dados (*Knowledge Discovery in Databases – KDD*).

O processo de *KDD* é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Segundo Fayyad et al. (1996), esse conjunto é composto de cinco etapas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (*Data Mining*); e interpretação e avaliação dos resultados. A interação entre estas diversas etapas pode ser observada na Figura 1, sendo que as três primeiras podem ser interpretadas como a análise exploratória dos dados.

O processo *KDD* refere-se a todo processo de descoberta de conhecimento útil nos dados, enquanto *Data Mining* refere-se à aplicação de algoritmos para extrair modelos dos dados; até 1995, muitos autores consideravam os termos *KDD* e *Data Mining* como sinônimos. Segundo Freitas (2000), o conhecimento a ser descoberto deve satisfazer a três propriedades: deve ser correto (tanto quanto possível); deve ser compreensível por usuários humanos; e deve ser interessante / útil / novo. Ainda, o método de descoberta do conhecimento deve apresentar as seguintes três características: deve ser eficiente (acurado), genérico (aplicável a vários tipos de dados) e flexível (facilmente modificável).

O objetivo do presente trabalho é mostrar, por meio de um problema médico apresentado na seção 3, a importância da análise exploratória dos dados quanto à classificação de padrões (discriminando-os tanto quanto possível para que, dados novos padrões, sua classificação possa ser efetuada com o menor erro possível), no contexto de *KDD*, apresentado nas seções 4 e 5. Na seção 2, é feita uma revisão da literatura sobre o assunto e, na seção 6, são apresentadas as conclusões.

2. Revisão da literatura

Com o amplo uso de tecnologia avançada para bases de dados, desenvolvida durante as últimas décadas, não tem sido difícil armazenar grandes volumes de dados em computadores e resgatá-los quando necessário. Embora os dados armazenados sejam um bem valioso de uma organização, muitas se deparam com o problema de ter "muitos dados, mas pouco conhecimento" sobre eles (*data rich but knowledge poor*) (Lu et al., 1995).

Esta grande quantidade de dados armazenada supera em muito as nossas habilidades de interpretá-los, criando a necessidade de se ter técnicas que permitam a sua automatização e análise de forma inteligente. As técnicas que buscam transformar os dados armazenados em conhecimento são o objetivo da área chamada de *KDD* (Fayyad, 1996). O *KDD* refere-se a todo processo de descoberta de conhecimento útil em bases de dados, enquanto *Data Mining*, principal etapa do *KDD*, refere-se à aplicação de técnicas de forma a extrair modelos dos dados (Figura 1).

Em geral, as técnicas de *Data Mining* desempenham as tarefas de classificação ou agrupamento dos dados ou, ainda, de descoberta de regras de associação entre os dados. Dentre os métodos de *Data Mining*, capazes de fazer o reconhecimento de padrões (classificação), podem-se citar as populares árvores de decisão, as máquinas de suporte de vetores (*Support Vector Machines, SVM*), os métodos estatísticos, as redes neurais, os algoritmos genéticos e as meta-heurísticas de uma forma geral; estas técnicas vêm sendo amplamente exploradas na literatura. Já as três etapas iniciais do *KDD*, referentes à análise exploratória dos dados, que fazem uso de ferramentas estatísticas, não têm recebido a mesma atenção por parte dos pesquisadores; é exatamente este aspecto que se pretende explorar no presente trabalho: a influência da análise dos dados preliminarmente à utilização das técnicas de *Data Mining*.

Dentre os numerosos trabalhos que abordam as técnicas de *Data Mining* para classificação, podem-se citar os relacionados a seguir que se concentram, principalmente, em Redes Neurais e Algoritmos Genéticos, comparando os seus desempenhos com as Árvores de Decisão.

Lu et al. (1995) e Lu et al. (1996) relatam em seus artigos que a abordagem conexionista, baseada em redes neurais, tem sido julgada como não adequada para a mineração de dados, pelo fato do conhecimento gerado não ser explicitamente representado na forma de regras adequadas para verificação ou interpretação por usuários e, por este motivo, os autores apresentam o algoritmo chamado *Neurorule* que faz a extração de regras a partir de uma rede neural treinada, obtendo regras do tipo SE-ENTÃO (*IF-THEN*). Estas regras, descrevem os autores, são similares a, ou mais concisas, que aquelas geradas por métodos simbólicos (árvores de decisão). Para a obtenção destas regras de classificação quanto a mineração de dados usando redes neurais, os autores utilizam uma abordagem que consta de três passos: 1. treinamento da rede neural; 2. poda da rede neural; e 3. extração de regras da rede neural treinada (algoritmo *Neurorule*). O desempenho desta abordagem é verificado, em ambos os artigos, em um problema de crédito bancário, sendo que, para facilitar a referida extração de regras, os valores dos atributos numéricos foram discretizados, dividindo-os em subintervalos. Após a discretização, o esquema de codificação "termômetro" foi empregado para obter representações binárias dos intervalos anteriormente definidos, obtendo-se, assim, as entradas para a rede neural. Os resultados obtidos nos artigos indicam que, usando a abordagem proposta, regras de alta qualidade podem ser descobertas a partir de um conjunto de dados.

Fidelis et al. (1999) apresentam um algoritmo de classificação baseado em Algoritmos Genéticos que descobre regras compreensíveis do tipo *IF-THEN* no contexto de *Data Mining*. O Algoritmo Genético proposto tem

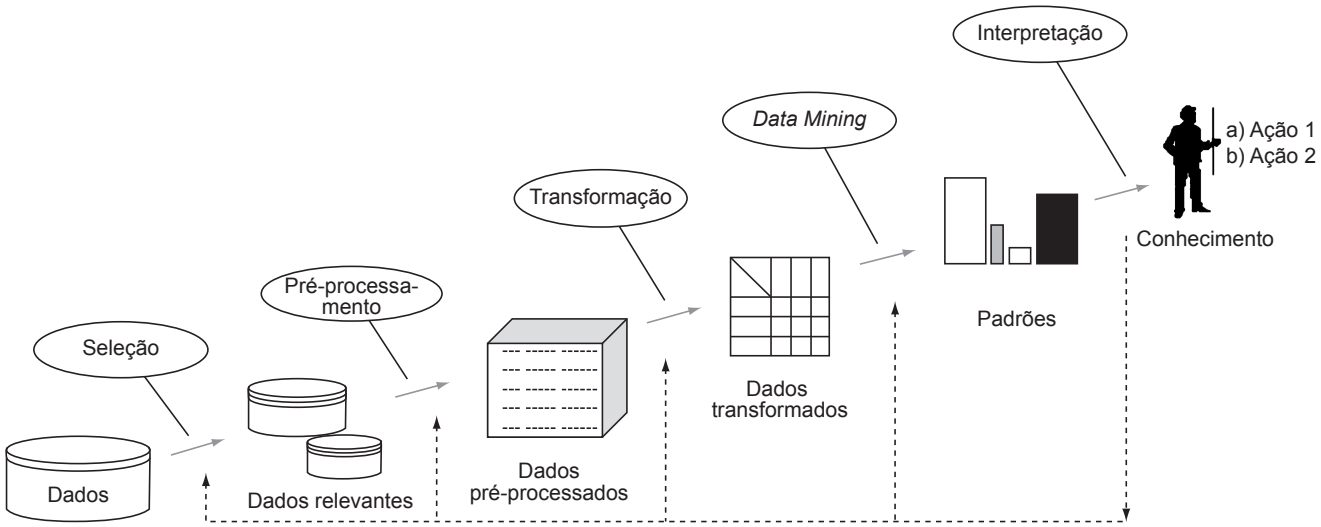

Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

um cromossomo flexível e cada um deles corresponde a uma regra de classificação. No trabalho de codificação, o número de genes (genótipo) é sempre fixo, cada gene corresponde a uma condição da parte *IF* de uma regra, mas o número das condições das regras (fenótipo) é variável. O Algoritmo Genético proposto possui operadores específicos de mutação e foi avaliado em duas bases de dados médicos de domínio público, de dermatologia e de câncer de mama, obtidos do *UCI* (*University of California at Irvine*) – Repositório de Aprendizado de Máquina (*Machine Learning Repository*).

Batista et al. (1999) tratam em seu artigo de um fator que pode levar a um desempenho inadequado na tarefa de classificação (e que tem recebido pouca atenção da comunidade científica): diferença grande na quantidade de padrões pertencentes a cada classe; são consideradas duas classes no estudo. São discutidos métodos de seleção de padrões que buscam diminuir criteriosamente o número de padrões da classe majoritária; tais métodos são ali denominados "métodos de seleção unilateral". Estes métodos detectam padrões rotulados erroneamente (atípicos), redundantes e próximos à borda de decisão da classe majoritária, sem perder a precisão de classificação desta classe.

Santos et al. (2000) ressaltam que um problema comum em *KDD* é a presença de ruído nos dados a serem minerados e, pelo fato das Redes Neurais serem robustas e terem boa tolerância a ruído, faz com que elas sejam adequadas para mineração de dados com muitos ruídos. O método proposto no artigo usa um Algoritmo Genético para definir uma topologia adequada a uma Rede Neural a ser treinada, sendo que, após o treinamento, esta topologia é então passada para um algoritmo de extração de regras e a qualidade das regras extraídas, baseada na acurácia e compreensibilidade delas, é examinada e retroalimentada ao Algoritmo Genético até que um critério de convergência seja alcançado. No Algoritmo Genético, cada candidato à solução é um candidato à topologia de uma Rede Neural e a função *fitness* envolve a acurácia preditiva e a compreensibilidade das regras extraídas dos indivíduos à topologia da rede. O sistema proposto foi avaliado em três conjuntos de dados disponíveis no repositório *UCI*: *Iris, Wine e Monks-2*, sendo que os resultados mostram que a abordagem é válida.

Baesens et al. (2003) apontam em seu trabalho a alta taxa de acurácia preditiva das Redes Neurais, mas também a sua falta de capacidade de explanação e, por este motivo, os autores abordam três métodos para a extração de regras de uma rede neural, comparativamente: *Neurorule*; *Trepan*; e *Nefclass*. Para comparar os desempenhos dos métodos abordados, foram utilizadas três bases de dados reais de crédito: *German Credit* (obtida do repositório *UCI*), *Bene 1* e *Bene 2* (obtidas das duas maiores instituições financeiras da *Benelux*). Os algoritmos mencionados são ainda comparados com os algoritmos *C4.5-árvore*, *C4.5-regras* e regressão logística. Os autores ainda mostram como as regras extraídas podem ser visualizadas como uma tabela de decisão na forma de um gráfico compacto e intuitivo, permitindo uma melhor leitura e interpretação dos resultados ao gerente de crédito.

Olden e Jackson (2002), pesquisadores preocupados com modelagem nas ciências ecológicas, descrevem alguns métodos da literatura para "desvendar" os mecanismos de uma rede neural artificial (Diagrama de Interpretação de uma Rede Neural, Algoritmo de Garson e Análise de Sensibilidade) e, além disso, propõem um método adicional, o qual denominam de abordagem randômica (*randomization approach*) para estatisticamente entender a importância

das conexões (pesos) de uma rede neural, assim como a contribuição das variáveis de entrada.

Na avaliação imobiliária, pode-se citar o trabalho de Nguyen e Cripps (2001) que compara o desempenho preditivo de Redes Neurais Artificiais com a Análise de Regressão Múltipla para a venda de casas de família. Múltiplas comparações foram feitas entre os dois modelos nas quais foram variados: o tamanho da amostra de dados, a especificação funcional e a predição temporal. No trabalho de Bond et al. (2002), os autores examinam o efeito que a vista de um lago (Lago Erie, E.U.A.) tem sobre o valor de uma casa. No estudo, foram levados em consideração os preços baseados na transação das casas (preço de mercado). Os resultados indicam que, além da variável vista, que se apresenta significativamente mais importante do que as demais, também a área construída e o tamanho do lote são importantes.

A seguir, é feita a descrição do problema médico abordado neste trabalho que será tratado, conforme já mencionado, por meio de algumas técnicas de *Data Mining* objetivando a tarefa de classificação, precedidas da análise exploratória dos dados, mostrando a sua influência no desempenho das técnicas.

3. Descrição do problema médico

A icterícia (do grego ikteros = amarelidão) que representa somente um sintoma, traduzido pela cor amarelada da pele e das mucosas e, eventualmente percebida nas secreções, pode ser proveniente de um imenso universo de doenças. É necessário que o médico separe essas inúmeras doenças em dois grandes grupos iniciais:

- Colestase (*chole* = bile, *stásis* = parada): é o caso em que há dificuldade ou impedimento do fluxo dos componentes da bile do fígado para o intestino; e
- outras causas.

Somente o 1º grupo - das colestases - será objeto de estudo (Steiner et al., 2004). Para fazer esta distinção inicial, o clínico se apoia geralmente em exames simples, definindo quais doentes apresentam a síndrome colestática, com segurança razoável. Isto, porém, não é suficiente e a separação em mais dois grupos se impõe:

- obstrução por câncer; e
- obstrução por cálculo.

Este diagnóstico diferencial geralmente é possível com os dados já obtidos, aliados a exames como a ultrassonografia e, eventualmente, tomografia axial computadorizada. Porém, de 16% a 22% dos doentes não são classificados, sendo que os exames complementares mencionados na região do duto biliar principal apresentam erros em torno de 30% a 40%. Exames capazes de estabelecer a real diferença entre câncer e cálculo como causa de obstrução existem e, quando utilizados em conjunto com os anteriores, apresentam precisão muito grande, acima de 95%. Entretanto, são geralmente invasivos e apresentam riscos de complicações graves e até letais.

Tendo em vista o risco do paciente e os altos custos envolvidos para um diagnóstico adequado, tem-se a justificativa para a utilização do processo *KDD* a este tipo de problema, em uma tentativa de otimizar o processo do diagnóstico (minimizando riscos e custos aos pacientes e, por outro lado, maximizando a eficácia nos resultados).

Para tanto, dados históricos dos pacientes enquadrados nos dois casos anteriores se fazem necessários. Foram utilizados dados de 118 pacientes do Hospital das Clínicas (HC) de Curitiba, PR, dos quais, comprovadamente, 35 possuíam câncer e 83 possuíam cálculo no duto biliar. De cada um destes pacientes, foram considerados 14 atributos oriundos de medidas de exames clínicos sugeridos por médico especialista da área, que se encontram no apêndice deste artigo.

De posse destes dados (padrões), já classificados pelo médico especialista da área como ictéricos com câncer ou ictéricos com cálculo, pode-se aplicar neles a análise exploratória, numa tentativa de separá-los ao máximo para, em seguida, utilizando as cinco técnicas de *Data Mining* aqui abordadas, fazer o treinamento delas com os referidos dados, visando à classificação de futuros padrões com a máxima acurácia possível por meio da técnica que, dentre as pesquisadas, tenha apresentado o melhor desempenho.

4. Técnicas utilizadas neste trabalho

As cinco etapas do processo *KDD* enumeradas na seção 1 deste trabalho (Figura 1) foram aqui abordadas, resumidamente, em três etapas distintas: análise exploratória dos dados; *Data Mining* e obtenção e análise dos resultados.

4.1 Análise exploratória dos dados

A complexidade de muitos problemas exige que o pesquisador colete muitas observações (dados, padrões) contendo, cada uma delas, muitas variáveis (atributos; entradas). A análise exploratória é assim denominada, pois objetiva utilizar métodos estatísticos para captar / explorar informações destes dados (Johnson e Wichern, 1998; Duda et al., 2001). Esta ampla área de estudo envolve diversas técnicas estatísticas sendo que, para este trabalho específico, esta análise ficou composta das seguintes: teste T^2 de Hotelling; transformação dos atributos; e descarte dos dados atípicos.

Os dados coletados foram organizados em uma matriz $\boldsymbol{X} = [x_{ij}]$, $i = 1,...,(m + k)$; $j = 1,...,n$, contendo os dados de $\boldsymbol{A}$ (com $m = 35$ dados) e $\boldsymbol{B}$ (com $k = 83$ dados) $\in R^n$, ou seja, $\boldsymbol{A}$ e $\boldsymbol{B}$ contêm n (= 14) atributos. Primeiramente, foi verificado se os pontos de $\boldsymbol{A}$ e $\boldsymbol{B}$ são distintos por meio do teste T^2 de Hotelling. Satisfeita esta condição, um

ajuste das respostas (ictéricos com câncer, representados por "1", e ictéricos com cálculo, representados por "0", respectivamente) e os atributos e coatributos derivados daqueles é executado por um Modelo Logístico Múltiplo, obtido após a análise da função desvio em consecutivos ajustes preliminares. Para obter-se este modelo, muitas modificações (adicionando, removendo, transformando e/ou combinando atributos, gerando os coatributos) são feitas até obter-se um modelo com a menor função desvio. Ao definir-se o modelo, simultaneamente se podem obter os resíduos de Pearson, indicando quais pontos de $\boldsymbol{A}$ e $\boldsymbol{B}$ são atípicos (*outliers*); e, se encontradas justificativas para tal, podem ser descartados.

Os testes comparando os métodos descritos na seção 4.2 deste trabalho foram baseados em duas matrizes: a primeira, $\boldsymbol{M}_1$ (= $\boldsymbol{X}_{m+k}$), contém os dados originais de $\boldsymbol{A}$ e $\boldsymbol{B}$ que não sofreram qualquer alteração; a segunda, $\boldsymbol{M}_2$ (= $\boldsymbol{X}_{m+k-r}$), em que r é a quantidade de dados descartados, contém os dados que foram explorados estatisticamente.

4.1.1 Teste T^2 de Hotelling

Inicialmente foi aplicado o teste T^2 de Hotelling (Johnson e Wichern, 1998) para verificar a igualdade dos vetores médios das duas amostras multivariadas $\boldsymbol{A}$ e $\boldsymbol{B}$. Para a aplicação deste teste, calcula-se:

$$\frac{T^2(m+k-n-1)}{(m+k-2)n} \sim F_{n,\,m+k-n-1} \tag{1}$$

em que:

$$T^2 = (\mathbf{x}_A - \mathbf{x}_B)'[(\frac{1}{m}+\frac{1}{k})\mathbf{S}_p]^{-1}(\mathbf{x}_A - \mathbf{x}_B) \tag{2}$$

O resultado é comparado com a distribuição $F_{n,\,m+k-n-1}(0.95)$ (distribuição *F de Snedecor* a uma probabilidade de 95%). Nesta expressão tem-se que:

$\boldsymbol{S}_p$ = matriz de covariância conjunta de $\boldsymbol{A}$ e $\boldsymbol{B}$;

$\boldsymbol{x}_A$ = vetor (n x 1) médio da amostra $\boldsymbol{A}$;

$\boldsymbol{x}_B$ = vetor (n x 1) médio da amostra $\boldsymbol{B}$;

$\boldsymbol{S}_p^{-1}$ = inversa da matriz covariância amostral conjunta:

$$\mathbf{S}_p = \frac{(m-1)\mathbf{S}_A + (k-1)\mathbf{S}_B}{m+k-2} \tag{3}$$

em que $\boldsymbol{S}_A$ = matriz de covariância da amostra $\boldsymbol{A}$ e $\boldsymbol{S}_B$ = matriz de covariância da amostra $\boldsymbol{B}$.

Neste teste, se:

$$\frac{T^2(m+k-n-1)}{(m+k-2)n} >> F_{n,\,m+k-n-1}(0.95) \tag{4}$$

rejeita-se, fortemente, com uma probabilidade de 95%, a hipótese de que as amostras estejam centradas no mesmo vetor de médias. Vale salientar que o teste T^2 de Hotelling é adequado para amostras de tamanho razoavelmente grande, que é o caso deste trabalho, e que, apesar do atributo "sexo" ser dicotômico (distribuição de Bernoulli), a estrutura de dados necessária à aplicação do teste não exige que todos os atributos do padrão sejam contínuos (Johnson e Wichern, 1998).

4.1.2 Transformação de atributos

Na procura por um Modelo Logístico Múltiplo, detalhado na seção 4.2.3 mais adiante, para ajustar as variáveis resposta ("1" ou "0", padrão pertencente ao conjunto $\boldsymbol{A}$ ou $\boldsymbol{B}$, respectivamente) com os atributos e/ou coatributos, uma medida da adequação do modelo é feita com a função desvio (*deviance*), introduzida por Nelder e Wedderburn (1972). Conforme o valor da função desvio fosse estatisticamente significativo (ou não), o coatributo era incorporado ao modelo (ou não). Neste trabalho, alguns atributos foram transformados na escala, originando vários coatributos, conforme apresentado na seção 5 mais adiante, na tentativa de se captar melhor a sua informação.

O desvio é uma medida da distância dos valores ajustados aos valores observados, ou equivalentemente, do modelo corrente ao modelo saturado; em geral, procuram-se modelos com desvios moderados. A função desvio é definida por (Nelder e Wedderburn, 1972):

$$s_p = -2\,\{L_p - L_{(m+k)}\} \tag{5}$$

em que L_p é o máximo do logaritmo da função de verossimilhança, para o modelo em investigação com p parâmetros, e $L_{(m+k)}$ é o máximo do logaritmo da função de verossimilhança, para o modelo saturado com $(m+k)$ parâmetros, ou seja, para o modelo saturado.

Existem dois modelos que são casos limites no procedimento de ajustamento: o modelo nulo e o modelo saturado. O modelo nulo tem um único parâmetro μ representativo para todos os Y_i's (resposta para cada padrão i) e entende-se que toda a variação nos dados é devida à componente aleatória. Já o modelo saturado possui $(m+k)$ parâmetros, um para cada observação Y_i, ajustando-se exatamente aos dados, isto é, as estimativas das médias são iguais às próprias observações e toda a variação é devida à componente sistemática (Steiner, 1995).

Na realidade, deve-se procurar um modelo com p parâmetros, situado entre esses modelos limites. O modelo saturado é degradado, pois não assume os dados, os repete, porém é útil como limite da discrepância para o modelo em investigação e é dado por (4.5).

Um modelo mal ajustado possui grande desvio e, obviamente, um modelo bem ajustado possui um desvio pequeno (igual a zero no modelo saturado). Os graus de liberdade associados ao desvio são definidos por: $\nu = (m+k) - p$.

O teste da razão de verossimilhança pode ser usado para decidir sobre o modelo mais adequado, sendo que a estatística do teste é dada por:

$$s_p = -2\{L_p - L_{p+1}\} \sim \chi^2_v \quad (6)$$

e, com base no valor-*p* correspondente a s_p,

$$P(\chi^2_v > s_p / \beta_{p+1} = 0) \quad (7)$$

mantém-se ou retira-se o atributo ou coatributo do modelo.

4.1.3 Descarte de dados atípicos

Ao obter-se o Modelo Logístico Múltiplo com o menor desvio possível, é importante analisar os resíduos dos dados em relação ao modelo para identificar pontos atípicos e, também, as suas causas. O procedimento para descarte de pontos atípicos foi feito com base no cálculo dos resíduos de Pearson que é feito para cada um dos dados, pelo seguinte cálculo (Johnson e Wichern, 1998; Duda et al., 2001):

$$e_i = \frac{Y_i - \theta_i}{\sqrt{\theta_i(1-\theta_i)}} \quad (8)$$

em que Y_i é o valor assumido pelo ponto *i* no modelo saturado ("1" ou "0", conforme já comentado) e θ_i é a estimativa deste valor feita pelo modelo. Um valor de $|e_i| \geq 1$ indica que o ponto *i* está sendo classificado erroneamente pelo modelo, ou seja, a observação *i* encontra-se "deslocada" em relação a sua amostra (***A*** ou ***B***).

Sugere-se que, para os casos em que $|e_i| \geq 1{,}5$, sejam procuradas justificativas para tal ocorrência como, por exemplo, medição errada, falha de equipamento e/ou outras. Se encontrada, a observação *i* pode ser descartada da amostra e, por conseguinte, do modelo. Observe-se que, neste caso, as estimativas para o modelo devem ser recalculadas. Não se descartam observações sem justificativa de erro externo ao modelo.

4.2 Técnicas de Data Mining

Desde o trabalho de Fisher, em 1936, numerosos trabalhos têm sido desenvolvidos com o propósito de apresentar técnicas de análise discriminante para a tarefa de classificação de padrões (Johnson e Wichern, 1998; Duda et al., 2001); estas técnicas podem ser enquadradas em *Data Mining* no contexto de *KDD* (Figura 1). Neste presente trabalho, o objetivo se atém à tarefa de classificação pela discriminação de padrões de clientes com câncer ou com cálculo no duto biliar.

Para tanto, fez-se a abordagem de cinco técnicas de *Data Mining* capazes de fazer a classificação (discriminação) entre os dados pertencentes aos conjuntos ***A*** e ***B***: uma técnica que faz uso da Programação Linear; duas técnicas estatísticas; uma que utiliza Redes Neurais; e, a última, Árvores de Decisão. São descritas sucinta e comparativamente a seguir, visando obter a técnica que discrimine os padrões com a máxima acurácia. Desta forma, dados novos padrões, os especialistas da área (médicos) terão a sua disposição um respaldo adicional para as tomadas de decisão quanto aos seus diagnósticos.

4.2.1 Geração de uma superfície que Minimiza erros

Bennett e Mangasarian (1992) propuseram uma formulação de um único programa linear que gera um plano que minimiza a média ponderada da soma das violações dos dados dos conjuntos ***A*** e ***B*** que estão do "lado errado" do plano separador. Quando as coberturas convexas dos dois conjuntos são disjuntas, o plano separa completamente os dois conjuntos. Quando as suas coberturas convexas se interceptam, o programa linear proposto gera um plano que minimiza os erros, que pode ser obtido por meio do seguinte modelo de Programação Linear, em que $\boldsymbol{e}_k$ e $\boldsymbol{e}_m \in R^k$ e R^m, respectivamente; $\boldsymbol{w}$ = vetor "peso"$\in R^n$, normal ao plano separador ótimo e $\gamma \in R$, fornece a localização da superfície separadora ótima $\boldsymbol{wx} = \gamma$.

$$\underset{w,\gamma,y,z}{Min}\ \frac{\boldsymbol{e}_m \boldsymbol{y}}{m} + \frac{\boldsymbol{e}_k \boldsymbol{z}}{k} \quad (9)$$

$$s.a:\ \boldsymbol{A}w - \boldsymbol{e}_m\gamma + \boldsymbol{y} \geq \boldsymbol{e}_m$$

$$-\boldsymbol{B}w + \boldsymbol{e}_k\gamma + \boldsymbol{z} \geq \boldsymbol{e}_k$$

$$\boldsymbol{y} \geq 0,\ \boldsymbol{y} \in R^m$$

$$\boldsymbol{z} \geq 0,\ \boldsymbol{z} \in R^k$$

Trata-se de um método não iterativo, ou seja, o plano separador obtido por este modelo é único para os conjuntos ***A*** e ***B***.

A Figura 2 mostra uma situação na qual o método foi aplicado ao plano R^2 (apenas dois atributos, x_1 e x_2) com 99 dados ($|\boldsymbol{A}| = 53$ and $|\boldsymbol{B}| = 46$) e está mostrado o plano separador ótimo $wx = \gamma$.

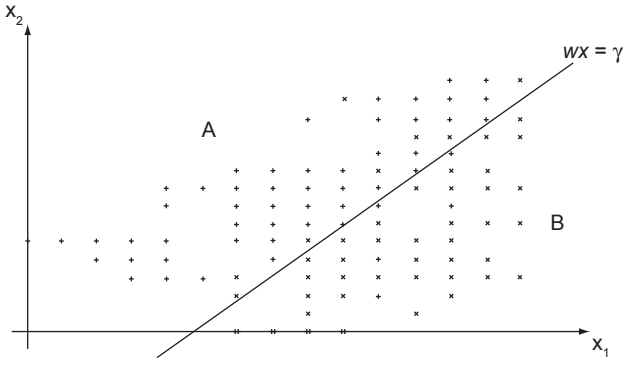

Figura 2. Ilustração gráfica do método "Geração de uma Superfície que Minimiza Erros" na qual está mostrado o plano separador ótimo para os conjuntos *A* e *B*, $wx = \gamma$.

4.2.2 Função discriminante linear (FDL) de Fisher

A terminologia "discriminar" e "classificar" foi introduzida na área de Estatística por Fisher no primeiro tratamento moderno dos problemas de separação de conjuntos (Johnson e Wichern, 1998). Dadas duas amostras $\boldsymbol{A}$ e $\boldsymbol{B}$ de observações multivariadas $\boldsymbol{x} \in R^n$, a idéia de Fisher foi transformar estas observações multivariadas em observações univariadas *Y's*, de tal modo que estejam separadas tanto quanto possível. O método cria os *Y's* como combinações lineares dos $\boldsymbol{x}$'*s*, ou seja, $Y = \boldsymbol{c}' \boldsymbol{x}$, em que $\boldsymbol{c} \in R^n$, pelo fato da combinação linear ser de fácil obtenção matematicamente.

A melhor combinação é obtida da razão entre o quadrado da distância entre as médias dos conjuntos $\boldsymbol{A}$ e $\boldsymbol{B}$ ($\boldsymbol{x}_A$ e $\boldsymbol{x}_B$) e a variância de *Y*. Neste contexto, a FDL de Fisher amostral, é dada a seguir.

$$Y = (\mathbf{x}_A - \mathbf{x}_B)' \, \mathbf{S}_p^{-1} \, \mathbf{x} \qquad (10)$$

em que $\boldsymbol{x}$ = vetor das variáveis aleatórias correspondentes às características amostrais observadas. Deste modo, para a classificação de um novo padrão $\boldsymbol{x}_0 \in R^n$, a regra de decisão, quanto à classificação, é dada a seguir. A Figura 3 faz uma ilustração deste método.

- Se $\boldsymbol{x}_0 \in \boldsymbol{A}$, então:

$$y_0 = (\mathbf{x}_A - \mathbf{x}_B)' \, \mathbf{S}_p^{-1} \mathbf{x}_0 \geq q = \frac{1}{2} (\mathbf{x}_A - \mathbf{x}_B)' \, \mathbf{S}_p^{-1} (\mathbf{x}_A + \mathbf{x}_B)$$

- Se $\boldsymbol{x}_0 \in \boldsymbol{B}$, então:

$$y_0 = (\mathbf{x}_A - \mathbf{x}_B)' \, \mathbf{S}_p^{-1} \mathbf{x}_0 < q = \frac{1}{2} (\mathbf{x}_A - \mathbf{x}_B)' \, \mathbf{S}_p^{-1} (\mathbf{x}_A + \mathbf{x}_B)$$

4.2.3 Modelo de regressão logística

A Regressão Logística (ou Modelo Logístico Múltiplo) consiste em relacionar, por meio de um modelo, a variável resposta (padrões pertencentes ao conjunto $\boldsymbol{A}$ ou $\boldsymbol{B}$) com os atributos que influenciam em sua ocorrência (Hair et al., 1998). Neste estudo, deseja-se quantificar a influência dos atributos como, por exemplo, nível de bilirrubina direta, na ocorrência de pacientes com cálculo ou câncer no duto biliar.

Quando a variável aleatória resposta *Y*, para a qual se deseja um modelo de ajuste é do tipo dicotômico (respostas "1" ou "0") e deseja-se estudar a relação entre *Y* e os diversos atributos (originais) e os diversos co-atributos (atributos derivados dos originais), o que se faz é estimar *Y* usando a função matemática sigmoidal (logística):

$$Y = f(\mathbf{x}) = (1 + e^{-\eta})^{-1}, \, \mathbf{x} \in \mathrm{R}^{\,n} \qquad (11)$$

com $\eta = g(\boldsymbol{x})$ obtido num ajuste linear. A qualidade do ajuste é medida pela função desvio s_p definida anteriormente na seção 4.1.2. Na Figura 4, está ilustrado este método.

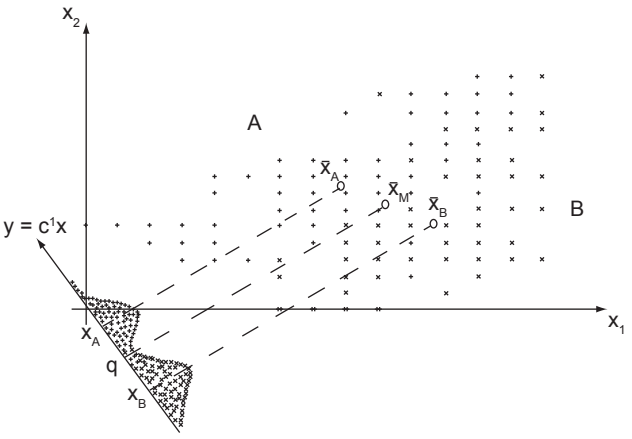

Figura 3. Ilustração gráfica da função discriminante linear de Fisher, na qual q separa os conjuntos *A* e *B*.

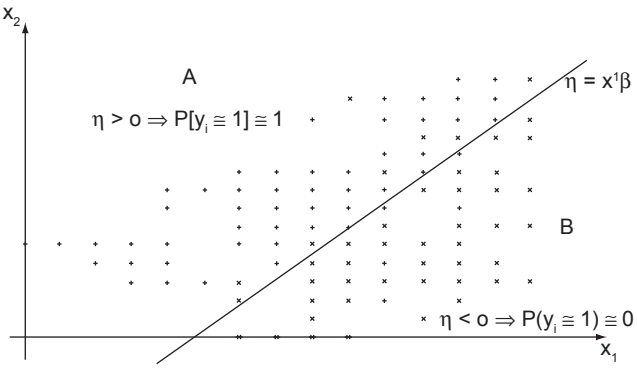

Figura 4. Ilustração gráfica do método "Regressão Logística", com $\eta = x' \beta$ obtido por um ajuste linear e $P(Y_i = 1) = 1 / (1 + e^{-\eta})$.

4.2.4 Redes neurais

As Redes Neurais de Múltiplas Camadas (RNMC) ou, também chamadas, redes do tipo "alimentadas para a frente" (*feed-forward*), utilizadas neste trabalho, constituem o modelo mais utilizado e amplamente divulgado pela comunidade científica e apresenta, de uma forma geral, resultados bastante satisfatórios. O algoritmo de retropropagação (*back-propagation*), utilizado para o seu treinamento, é um algoritmo de aprendizado supervisionado (Kröse e Van Der Smagt, 1993; Fausett, 1995).

Durante o treinamento da rede com o algoritmo *back-propagation*, a rede opera em uma seqüência de duas fases. Numa primeira fase, um padrão (dos $(m + k) = (35 + 83)$ padrões) pertencente ao conjunto $\boldsymbol{A}$ ou $\boldsymbol{B}$ deste trabalho, é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada, até obter-se a resposta na camada de saída. Na segunda fase, a resposta obtida é comparada com a resposta desejada ("1" ou "0") para esse padrão e o erro é calculado. O erro é propagado a partir da camada de saída até a cama-

da de entrada, e os pesos das conexões das unidades das camadas internas vão sendo ajustados conforme o erro é retropropagado (Kröse e Van Der Smagt, 1993). Uma iteração é completada quando todos os $(m + k)$ padrões tiverem sido apresentados à rede. Para os dados apresentados e utilizados neste trabalho, a rede neural precisou de, aproximadamente, 1.000 iterações para convergir em cada uma das situações de teste. Uma interpretação geométrica para este método está na Figura 5.

4.2.5 Árvores de decisão

Quinlan (1993) desenvolveu a técnica que permitiu o uso da representação do conhecimento por meio das Árvores de Decisão. A sua contribuição consistiu na elaboração de um algoritmo chamado *ID3* que, juntamente com suas evoluções *ID4*, *ID6*, *C 4.5*, *See 5*, são ferramentas adaptadas ao uso das árvores de decisão.

As Árvores de Decisão podem ser usadas em conjunto com a tecnologia de Indução de Regras, mas são únicas no sentido de apresentar os resultados num formato com priorização. Nelas, o atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostrados nos nós subseqüentes. As vantagens principais das Árvores de Decisão são que elas induzem a uma escolha no processo de decisão, levando em consideração os atributos que são mais relevantes, além de serem compreensíveis para as pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

Uma Árvore de Decisão utiliza a estratégia chamada *dividir-para-conquistar*. Um problema complexo é decomposto em subproblemas mais simples; recursivamente a mesma estratégia é aplicada a cada subproblema (Witten e Frank, 2000). A capacidade de discriminação de uma Árvore de Decisão advém das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço.

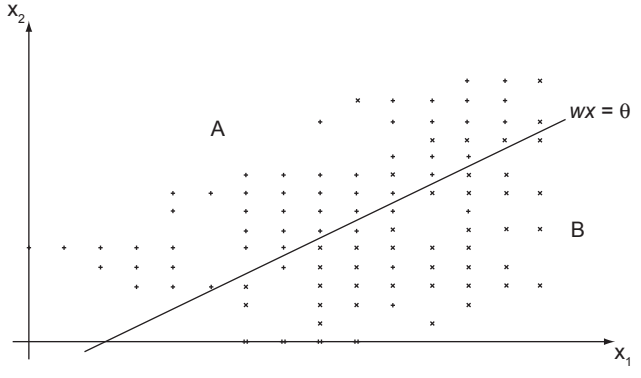

Figura 5. Ilustração gráfica de uma RNMC na qual o plano separador $wx = \theta$ é obtido com o treinamento de uma rede sem camada oculta.

Após a construção de uma Árvore de Decisão, é possível derivar regras. Essa transformação da Árvore de Decisão em regras geralmente é feita com o intuito de facilitar a leitura e a compreensão humana. Assim, as Árvores de Decisão podem ser representadas como conjuntos de regras do tipo *IF-THEN*.

A utilização de uma Árvore de Decisão apresenta, ainda, outras vantagens, dentre as quais podem ser destacadas as seguintes: não assumem nenhuma distribuição particular para os dados; as características ou atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos); podem-se construir modelos para qualquer função, desde que o número de exemplos de treinamento seja suficiente e de elevado grau de compreensão.

5. Implementação das técnicas ao problema médico, obtenção e análise dos resultados

Conforme descrito a seguir, os testes computacionais foram feitos em duas matrizes de dados: em uma primeira matriz, $\boldsymbol{M_1}$, contendo os dados e atributos originais, na qual foi aplicado apenas o teste T^2 de Hotelling, e em uma segunda matriz, $\boldsymbol{M_2}$, na qual os dados tiveram seus atributos transformados e seus pontos atípicos descartados, além de serem analisados pelo teste T^2 de Hotelling; desta forma, $\boldsymbol{M_2}$ contém dados "ajustados" ou "transformados" estatisticamente.

Para a aplicação do teste T^2 de Hotelling (descrito em 4.1.1), foi desenvolvido um programa computacional em *Visual Basic*. Como resultado da aplicação desse programa, foi verificado que as amostras, pacientes com câncer e pacientes com cálculo, são distintas em um nível de confiança de 95%. Mais especificamente, o teste T^2 de Hotelling, com estatística $T^2(m + k - n - 1)/(m + k-2)n$, comparada com $F_{n,\, m+k-n-1}\,(0.95)$ para cada uma das matrizes de dados $\boldsymbol{M_1}$ (dados sem ajuste) e $\boldsymbol{M_2}$ (dados com ajuste), forneceu os seguintes valores:

- $\boldsymbol{M_1}$: $4.84 > 1.78896 = F_{14,103}\,(0.95)$;
- $\boldsymbol{M_2}$: $12.54 > 1.82239 = F_{13,97}(0.95)$.

Conseqüentemente, rejeita-se a igualdade dos dois vetores de média, pois a probabilidade de elas serem iguais é bem menor do que 5%, nos dois casos, especialmente em $\boldsymbol{M_2}$. Assim, a amostra de ictéricos cancerosos é distinta da de ictéricos com cálculo nos atributos e coatributos estudados.

A transformação dos atributos e o descarte de pontos atípicos (descritos em 4.1.2 e 4.1.3), para obter a matriz $\boldsymbol{M_2}$, foram feitos simultaneamente com a utilização do *software* estatístico *GLIM* (*Generalized Linear Interactive Modeling*) (Aitkin et al., 1989), pela análise

dos desvios (valor de s_p), para cada um dos modelos obtidos, e dos resíduos (valor de $|e_i|$), para cada um dos pontos, respectivamente. Neste procedimento, foram descartados sete pontos (do total de 118), ou seja, 6% do total, pontos estes considerados atípicos (apresentaram $|e_i| > 1{,}5$). Esta hipótese foi assumida após discussão com os especialistas da área e apontadas as causas.

Estes procedimentos (transformação dos atributos e descarte de dados atípicos) foram quase que "artesanais", pois para a obtenção dos Modelos Logísticos Múltiplos, contendo os atributos e suas "variações" (adicionando, removendo, dividindo, combinando e outros, os atributos originais, gerando os coatributos), foi feita manualmente, pesquisando algumas das inúmeras possibilidades, uma de cada vez, de se "trabalhar" com os 14 atributos originais. A cada "variação" verificava-se o valor do desvio do modelo, com o objetivo de minimizá-lo, mas não em demasiado, conforme explicado em 4.1.2.

Fazendo este trabalho "artesanal", obteve-se um Modelo Logístico Múltiplo com 13 atributos. Na transformação dos atributos, a matriz $\boldsymbol{M_2}$ ficou definida com os seis atributos originais e sete coatributos (transformados a partir dos originais). Os seis atributos originais são os seguintes (ver apêndice): idade (id); bilirrubina total (bt); bilirrubina direta (bd); amilase (ami); fosfatases alcalinas (fa); e volume globular (vg). Já os sete coatributos são: $(\text{bilirrubina direta})^2$, ou seja, $(\text{bd})^2$; ln(amilase), ou ainda, lnam; a divisão dos atributos sgot por sgpt, ou seja, sgot/sgpt, ou ainda, st; $(\text{sgot/sgpt})^2$, ou seja, st^2; $(\text{fosfatases alcalinas})^2/1000$, ou seja, fa2n; $(\text{volume globular})^2$, ou seja, $(\text{vg})^2$; e $(\text{bilirrubina total})^2$, ou seja, $(\text{bt})^2$. Observe-se que foram descartados os seguintes seis atributos ao se compor $\boldsymbol{M_2}$: sexo (sex); bilirrubina indireta (bi); tempo de atividade da protrombina (tap); albumina (alb); creatinina (cr); e leucócitos (le).

Ao obter-se cada um dos modelos, em especial este último modelo com 13 atributos (seis atributos originais e sete coatributos), o *software GLIM* automaticamente já fornece os valores dos resíduos de Pearson. Levando-se em consideração o procedimento descrito em 4.1.3, foram descartados sete pontos. Vale ressaltar que a matriz $\boldsymbol{M_1}$ contém os dados e atributos originais, sem alterações.

Tendo-se as matrizes $\boldsymbol{M_1}$ (matriz com os dados originais, sem ajuste, da ordem 118 x 14) e $\boldsymbol{M_2}$ (matriz com os dados ajustados, da ordem 111 x 13), fez-se a aplicação dos cinco métodos de *Data Mining* abordados neste trabalho.

Para o 1º método ("Geração de uma Superfície que Minimiza Erros"), utilizou-se o *software* comercial ***LINGO*** (*Language for Interactive General Optimizer*) para a resolução do modelo de Programação Linear; para o 2º (FDL de Fisher) e o 4º (RNMC) métodos, fez-se a implementação computacional em *Visual Basic*; para o 3º método (Regressão Logística), fez-se uso, novamente, do *software* estatístico *GLIM* (*Generalized Linear Interactive Modeling*); e, finalmente, para o 5º e último método (Árvore de Decisão), utilizou-se o *software* livre ***WEKA*** (*Waikato Environment for Knowledge Analysis*, disponível no site *www.cs.waikato.ac.nz/ml/weka*).

Vale salientar que foi utilizada uma rede neural com três camadas: camada de entrada, com o número de neurônios de entrada igual ao número de informações (14 ou 13, para $\boldsymbol{M_1}$ e $\boldsymbol{M_2}$, respectivamente), camada intermediária, contendo um número de neurônios variando de um a 20, conforme metodologia apresentada por Steiner (1995), e camada de saída, com um neurônio (paciente com câncer, "1", ou paciente com cálculo, "0"). Já para a formação da árvore de decisão fez-se uso do algoritmo de classificação em árvore *J4.8* (*C4.5 release 8*).

A metodologia para testes em todos os cinco métodos consistiu em aplicar o método *holdout* estratificado (Witten e Frank, 2000) repetido 10 vezes (10 simulações), para a matriz $\boldsymbol{M_1}$ e 10 vezes para $\boldsymbol{M_2}$, e, como o desvio padrão entre elas foi relativamente pequeno para as duas matrizes, optou-se por considerar apenas as 3 "melhores" simulações (com menor percentual de erros) tanto para $\boldsymbol{M_1}$ como para $\boldsymbol{M_2}$, para o cálculo da média apresentada no Quadro 1.

Para executar as 10 simulações para cada uma das matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$, dividiram-se os conjuntos de pontos $\boldsymbol{A}$ e $\boldsymbol{B}$ aleatoriamente em dois subconjuntos: um dos subconjuntos, denominado "Conjunto para Treinamento", serviu para "treinar" cada um dos cinco métodos e o outro subconjunto, "Conjunto para Teste", serviu para testar os modelos treinados. Este procedimento foi repetido 10 ve-

Quadro 1. Média das percentagens dos acertos (três simulações) dos cinco métodos para o problema médico, com a matriz M_1 (118 x 14), contendo os dados originais, e com a matriz ajustada M_2 (111 x 13), contendo os dados explorados estatisticamente.

Métodos	Progr. linear		F.D.L. Fisher		Regr. log.		Redes neurais		Árv. dec.	
Conjunto de Dados	Treinam. (%)	Teste (%)	Treinam. (%)	Teste (%)	Treinam. (%)	Teste (%)	Treinam. (%)	Teste (%)	Treinam. (%)	Teste (%)
M_1	83,65	75,01	81,13	80,56	83,33	80,56	77,00	72,33	95,91	77,78
M_2	100,00	96,97	93,34	90,91	98,50	95,45	85,67	96,97	97,67	75,75

zes para cada matriz variando-se os dois subconjuntos e, conforme já comentado, foram considerados os resultados das três "melhores" (menor percentual de erros) simulações e a média das percentagens dos erros destas três simulações foi calculada (Quadro 1). O "Conjunto de Treinamento" foi usado para testar os métodos também.

Os valores para os desvios padrões entre as 10 simulações, considerando os Conjuntos de Treinamento e de Testes, para todos os cinco métodos, variaram nos intervalos (1,27%; 3,47%) e (1,78%; 5,12%) para a matriz M_1, respectivamente; já para a matriz M_2, estes valores ficaram nos intervalos (0,17; 2,09) e (0,67; 3,58).

As cinco técnicas de *Data Mining* foram aplicadas nas matrizes M_1 (118 x 14) e M_2 (111 x 13), separadamente, para se poder comparar o desempenho delas sobre os dados originais e sobre os dados ajustados, conforme apresentado no Quadro 1.

Neste quadro 1, tem-se, por exemplo, para o 1º método abordado, "Geração de uma Superfície que Minimiza Erros", que envolve um modelo de Programação Linear (duas primeiras colunas), 83,65% e 100% de acertos para M_1 e M_2, respectivamente, considerando o Conjunto de Treinamento e 75,01% e 96,97% de acertos para M_1 e M_2, respectivamente, considerando o Conjunto de Testes. Da mesma forma, tem-se a interpretação para os demais quatro métodos.

Como se pode observar neste quadro, o 1º método ("Geração de uma Superfície que minimiza Erros" ou, simplesmente, o método que envolve Programação Linear) e o 4º método (envolve Redes Neurais) foram os que apresentaram o maior percentual de acertos para o Conjunto de Testes (96,97%), mostrando que estes dois métodos apresentam uma capacidade de generalização bastante satisfatória. A estes dois métodos seguem-se os métodos de Regressão Logística, Função Discriminante Linear de Fisher e Árvores de Decisão, com 95,45%, 90,91%, e 75,75% de acertos, considerando o Conjunto de Testes.

6. Conclusões

Neste trabalho, foi estudada a importância da análise exploratória dos dados preliminarmente ao uso das técnicas de *Data Mining*, por meio dos resultados obtidos com a aplicação destas a um problema médico real (dados de 118 pacientes, cada um deles com 14 atributos). As cinco técnicas de *Data Mining* estudadas ("Geração de uma Superfície que Minimiza Erros" ou Programação Linear; Função Discriminante Linear de Fisher; Regressão Logística; Redes Neurais e Árvores de Decisão) foram aplicadas a partir dos dados originais (matriz M_1 (118 x 14)) e a partir dos dados explorados estatisticamente (transformação dos atributos e descarte de dados atípicos (matriz M_2 (111 x 13))). Os resultados encontram-se no Quadro 1.

Analisando-se estes resultados, nota-se que todos os métodos, com exceção do Conjunto para Teste da Árvore de Decisão, apresentaram uma melhoria significativa nos seus desempenhos com a adoção da análise estatística exploratória dos dados, preliminarmente à aplicação das técnicas de *Data Mining*, como se pode verificar, comparando-se os resultados obtidos por meio de M_1 e de M_2. Este fato enfatiza a importância de se ter dados confiáveis e consistentes e, por conseguinte, dados multivariados explorados estatisticamente. Dos métodos abordados, os que envolvem Programação Linear e Redes Neurais são os que apresentaram os melhores resultados para o problema médico estudado, ou seja, maiores percentuais de acertos para os Conjuntos de Testes, mostrando com isso, as suas capacidades de generalização satisfatórias: 97% de acerto.

Vale destacar, ainda, que de todas as técnicas de *Data Mining* aqui abordadas, apenas a árvore de decisão deixa claro ao usuário quais são os atributos que estão discriminando os padrões (compreensibilidade) e de que forma ("pontos de corte") está ocorrendo, como se pode observar na Figura 6, que exemplifica um dos testes feitos. Nesta árvore, tem-se que Bd (Bilirrubina direta) é o atributo mais importante ou a "raiz" da árvore. Se o seu valor for menor ou igual a 6,2, então o paciente possui cálc (cálculo); caso contrário, o atributo Ami (Amilase), 2º atributo mais importante, é analisado. Se o seu valor for maior do que 164, então o paciente possui cálc também; caso contrário analisa-se, novamente, o valor do atributo ami: se for maior do que 92, então o paciente possui cân (câncer); caso contrário o atributo fa (fosfatases alcalinas) é analisado e, assim, de forma análoga, tem-se a interpretação do restante desta árvore.

O percentual de erros relativamente alto desta técnica, se comparado com o das demais, é compensado por esta característica altamente desejável, ou seja, deixar claro os atributos "discriminadores" com os seus respectivos pontos de corte, num formato de priorização, conforme comentado em 4.2.5. Vale salientar que a partir de uma árvore de decisão, podem-se gerar regras de classificação; para o caso da Figura 2, tem-se as seguintes regras:

Se Bd ≤ 6,2 ⇒ Class = cálc

Se Bd > 6,2 e Ami > 164 ⇒ Class = cálc

Se Bd > 6,2 e 92 < Ami ≤ 164 ⇒ Class = cân

Se Bd > 6,2 e Ami ≤ 92 e Fa > 337,7 ⇒ Class = cân

Se Bd > 6,2 e Ami ≤ 92 e Fa ≤ 337,7 e Vg > 42,3 ⇒ Class = cálc

Se Bd > 6,2 e Ami ≤ 92 e Fa ≤ 337,7 e Vg ≤ 42,3 e Bi ≤ 5,5 ⇒ Class = cálc

Se Bd > 6,2 e Ami ≤ 92 e Fa ≤ 337,7 e Vg ≤ 42,3 e Bi > 5,5 ⇒ Class = cân

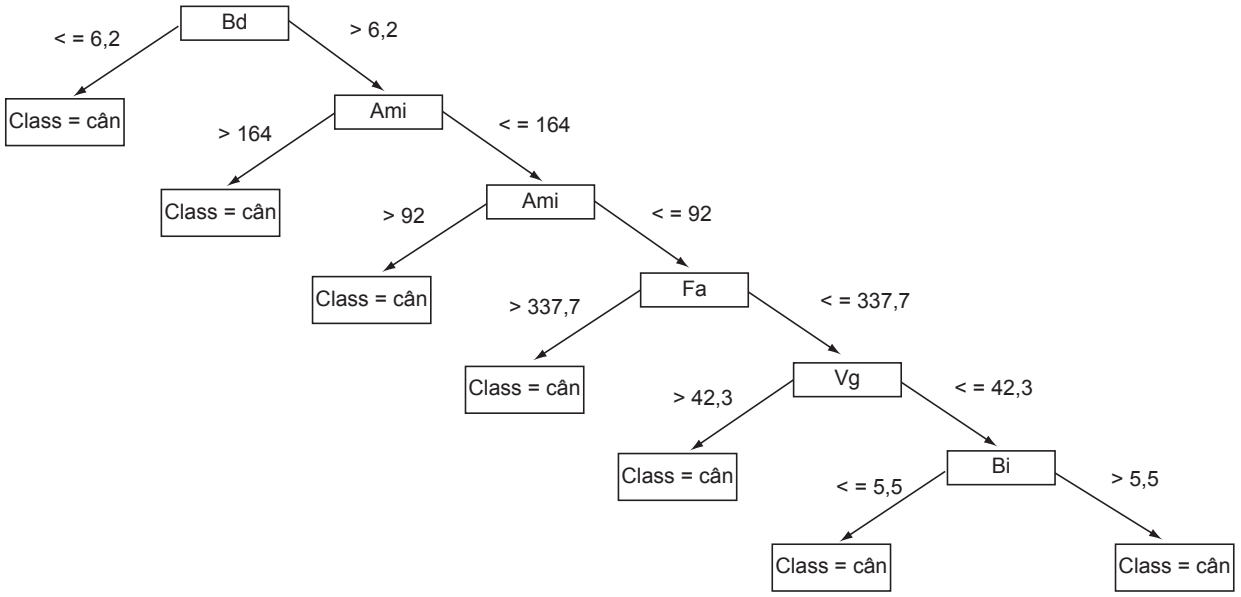

Figura 6. Um exemplo de árvore de decisão para o problema médico abordado, utilizando as abreviaturas dos atributos conforme apresentado no apêndice.

Dentre as técnicas abordadas, a técnica que envolve Redes Neurais, também poderia tornar-se mais compreensível aos usuários, bastando para isto utilizar algum algoritmo de extração de regras a partir da rede neural treinada, conforme apresentado por Lu et al. (1996) e Santos et al. (2000), dentre outros. Da mesma forma, poder-se-ia pensar em construir algoritmos de extração de regras para os demais três métodos também.

Os métodos apresentados podem ser utilizados nos mais diversos problemas reais de classificação, como o caso médico abordado neste trabalho. Deste modo, especialistas das mais diversas áreas poderiam avaliar os resultados fornecidos pelos métodos abordados (e/ou outros métodos adicionais utilizados para classificação) e validar (ou não) a plausibilidade das predições realizadas por eles, tendo-se, pois, uma ferramenta auxiliar para as suas tomadas de decisão. Assim, de posse do método validado com menor percentual de erros, novos padrões (pacientes, neste trabalho) poderiam ser classificados por ele, fornecendo um maior respaldo às tomadas de decisão dos especialistas, ou seja, as suas classificações (cálculo ou câncer de novos pacientes).

Apêndice

Os 14 atributos considerados pelo especialista na área foram os seguintes (com as suas respectivas abreviaturas):

1. Idade (Id);
2. Sexo (Sex);
3. Bilirrubina total (Bt) (do latim bilis = bile + ruber = vermelho): mede a intensidade da cor amarelada. Representa a soma das bilirrubinas indiretas e diretas;
4. Bilirrubina direta (Bd): representa a bilirrubina conjugada com ácido glicurônico;
5. Bilirrubina indireta (Bi): representa alterações produzidas como conseqüências de afecções totalmente distintas como as obstruções causadas por cálculos biliares ou por neoplasmas (como os cânceres);
6. Fosfatases alcalinas (Fa): a elevação desta enzima é o indicador de colestase mais amplamente usado no mundo e provavelmente o mais sensível;
7. SGOT (Sgot): são as transaminases oxalacéticas do soro. Existem em vários tecidos como muscular, esquelético e cardíaco, no rim, no cérebro, no citoplasma e nas mitocôndrias;
8. SGPT (Sgpt): são as transaminases glutâmico-pirúvicas. Lesões pouco acentuadas aumentam inicialmente esta enzima, enquanto os aumentos acentuados indicam lesões intensas das células geralmente com necroses. De uma maneira geral, existe aumento dessas enzimas (sgot e sgpt) em todas as doenças hepáticas. A diminuição dos níveis dessas enzimas após uma elevação inicial, nem sempre é indicativo de melhora, ao contrário, pode ser indicativo de gravidade. Nas obstruções crônicas por cálculos ou por neoplasmas como câncer, os seus níveis permanecem ligeira ou moderadamente elevados;

9. Tempo de atividade da protrombina (Tap): a dosagem dos fatores de coagulação sintetizados no fígado, serve para diagnóstico diferencial entre lesão predominante de hepatócitos e impedimento de absorção da vitamina K por obstrução ao fluxo biliar;
10. Albumina (Alb): é uma proteína sintetizada exclusivamente no fígado. Indica a função global do fígado; normalmente um fígado de adulto sintetiza 12 g dessa proteína por dia, tendo uma reserva de 500 g;
11. Amilase (Ami): é uma enzima hidrolítica que digere amido; essas alfa-amilases são encontradas principalmente na saliva e no pâncreas. A migração de cálculos biliares pelas vias biliares principais pode provocar inflamação aguda do pâncreas;
12. Creatinina (Cr): é um composto orgânico nitrogenado e não-protéico formado a partir da desidratação da creatina, que é sintetizada nos rins, fígado e pâncreas;
13. Leucócitos (Le): ou glóbulos brancos são o segundo tipo de células mais comuns do sangue. Nas doenças infecciosas agudas produzidas por bactérias, o número total de leucócitos pode estar muito elevado, não sendo raros valores de 15.000 a 30.000 por mL de sangue; e
14. Volume Globular (Vg): é a medida do volume globular.

    Classificação (class ou resposta): câncer (cân) ou cálculo (cálc).

Referências Bibliográficas

AITKIN, M.; ANDERSON, D.; FRANCIS, B.; HINDE, J. **Statistical Modelling in GLIM**. 1. ed. Clarendon Press, New York: Oxford Statistical Science Series, 1989. 374 folhas.

BAESENS, B.; SETIONO, R.; MUES, C.; VANTHIENEN, J. Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. **Management Science**, Evanston, IL, INFORMS, v. 49, n. 3, p. 312-329, 2003.

BATISTA, G. E. A. P. A.; CARVALHO, A. C. P. L. F.; MONARD, M. C. Aplicando Seleção Unilateral em Conjuntos de Exemplos Desbalanceados: Resultados Iniciais. In: XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO "EDUCAÇÃO E APRENDIZAGEM NA SOCIEDADE DA INFORMAÇÃO", 20 e 21 de julho de 1999, RJ. **Anais...** EntreLugar, RJ. v. 4, p. 327-340, 1999.

BOND, M. T.; SEILER, V. L.; SEILER, M. J. **Residential Real Estate Prices: a Room with a View**. The Journal of Real Estate Research, Fullerton, CA, American Real Estate Society, v. 23, n. 1, p. 129-137, 2002.

BENNETT, K. P.; MANGASARIAN, O. L. **Robust Linear Programming Discrimination of Two Linearly Inseparable Sets**. Optimization Methods and Software, London, United Kingdom, Taylor and Francis Group, v. 1, p. 23-34, 1992.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 1 ed. New York, John Wiley & Sons, inc., 2001. 654 folhas.

FAUSETT, L. **Fundamentals of Neural Networks - Architectures, Algorithms, and Applications**. 1 ed. Florida Institute of Technology. Prentice Hall, Upper Saddle River, New Jersey, 1995. 461 folhas.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery & Data Mining**. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1996. 611 folhas.

FIDELIS, M. V.; LOPES, H. S.; FREITAS, A. A. Um Algoritmo Genético para Descobrir Regras de Classificação em *Data Mining*. In: XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, "EDUCAÇÃO E APRENDIZAGEM NA SOCIEDADE DA INFORMAÇÃO", 20 e 21 de julho de 1999, RJ. **Anais...** EntreLugar, RJ. v. 5, p. 17-29, 1999.

FREITAS, A. A. **Uma Introdução a *Data Mining***. Informática Brasileira em Análise, Centro de Estudos e Sistemas Avançados do Recife (C.E.S.A.R.), Recife, Pe, ano II, n. 32, 2000.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 4 ed. New Jersey, Prentice-Hall, inc., 1998. 815 folhas.

HAIR JR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados** (traduzido). 5 ed. Bookman, São Paulo, 1998. 593 folhas.

KRÖSE, B. J. A.; VAN DER SMAGT, P. P. **An Introduction to Neural Networks**. 5 ed. University of Amsterdam, Amsterdam, 1993. 130 folhas.

LU, H.; SETIONO, R.; LIU, H. NeuroRule: A Connectionist Approach to *Data Mining*. In: 21st. VERY LARGE DATA BASE CONFERENCE, (CD), Zurich, Switzerland, **Proceedings...** 1995. p. 478-489.

LU, H.; SETIONO, R.; LIU, H. **Effective *Data Mining* using Neural Networks**. IEEE Transactions on Knowledge and Data Engineering, v. 8, n. 6, p. 957-961, 1996.

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models**. J. R. Statistical Society, Series A, Royal Statistical Society, United Kingdom, v. 135, n. 3, p. 370-384, 1972.

NGUYEN, N.; CRIPPS, A. **Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks**. The Journal of Real Estate Research, Fullerton, CA, American Real Estate Society, v. 22, n. 3, p. 313-336, 2001.

OLDEN, J. D.; JACKSON, D. A. **Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks**. Ecological Modelling, Oxford, United Kingdom, Elsevier Science, v. 154, p. 135-150, 2002.

QUINLAN, J. C. **C4.5: Programs for Machine Learning**. 1 ed. Morgan Kaufmann, San Mateo, Califórnia, 1993. 302 folhas.

SANTOS, R. T.; NIEVOLA, J. C.; FREITAS, A. A. Extracting Comprehensible Rules from Neural Networks via Genetic Algorithms. In: SYMPOSIUM ON COMBINATIONS OF EVOLUTIONARY COMPUTATION AND NEURAL NETWORK (ECNN-2000), San Antonio, Texas. **Proceedings...**San Antonio: Texas IEEE Press, 2000. v. 1, p. 130-139.

STEINER, M. T. A. **Uma Metodologia para o Reconhecimento de Padrões Multivariados com Resposta Dicotômica**. 158 folhas. Tese de Doutorado em Engenharia de Produção – UFSC, Florianópolis, SC, 1995.

STEINER, M. T. A.; SOMA, N. Y.; SHIMIZU, T.; NIEVOLA, J. C.; LOPES, F.; SMIDERLE, A. *Data Mining* como Suporte à Tomada de Decisões - uma Aplicação no Diagnóstico Médico. In: XXXVI SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, "O IMPACTO DA PESQUISA OPERACIONAL NAS NOVAS TENDÊNCIAS MULTIDISCIPLINARES", 23 a 26 de Novembro de 2004, São João del-Rei, MG (CD). **Anais...** Rio de Janeiro, RJ, 2004. p. 96-107.

WITTEN, I. H.; FRANK, E. ***Data Mining*: Pratical Machine Learning Tools and Techniques with Java Implementations**. Morgan Kaufmann Publishers. San Francisco, California, 2000. 371 folhas.

STUDY OF A MEDICAL PROBLEM USING KDD, WITH EMPHASIS ON EXPLORATORY DATA ANALYSIS

Abstract

Knowledge Discovery in Databases – KDD – is a process that consists of several steps, beginning with the collection of data for the problem under analysis and ending with the interpretation and evaluation of the final results. This paper discusses the influence of exploratory data analysis on the performance of Data Mining techniques with respect to the classification of new patterns, based on its application to a medical problem, and compares the performance of these techniques in order to identify the one with the highest percentage of successes. The results of this study lead to the conclusion that, providing this analysis is done properly, it can significantly improve the performance of these techniques and serve as an important tool to optimize the end results. For the problem under study, the techniques involving a Linear Programming model and Neural Networks were the ones showing the lowest percentages of errors for the test sets, presenting good generalization capacities.

Keywords: data mining, KDD process, exploratory data analysis.