A Influência da Análise Multivariada no Processo de Descoberta de Conhecimento em Bases de Dados
- uma Aplicação ao Diagnóstico Médico -

Maria Teresinha Arns Steiner
UFPR – Departamento de Matemática, CP: 19081-CEP: 81531-990, Curitiba, PR; tere@mat.ufpr.br

Nei Yoshihiro Soma
ITA – Divisão da Ciência da Computação, Pça Mal. Eduardo Gomes, 50, Vl. das Acácias
CEP: 12228-990, São José dos Campos, SP; nysoma@comp.ita.br

Tamio Shimizu
USP – Departamento de Engenharia de Produção, São Paulo, SP; tmshimiz@usp.br

Júlio Cesar Nievola
PUC-PR – Programa de Pós-Graduação em Informática Aplicada, Av. Imaculada Conceição, 1155, CEP 80215-901, Curitiba, PR; nievola@ppgia.pucpr.br

Pedro José Steiner Neto
UFPR – Departamento de Administração, CP: 19081-CEP: 81531-990, Curitiba, PR; pedrosteiner@ufpr.br

Resumo
"Descoberta de Conhecimento em Bases de Dados" ("*Knowledge Discovery in Databases*", *KDD*) é um processo composto de várias fases iniciando com a coleta de dados para o problema e finalizando com a avaliação dos resultados finais. Este trabalho objetiva mostrar a influência da análise multivariada na performance das técnicas de Mineração de Dados ("*Data Mining*") para a classificação de novos padrões. Como indicado neste trabalho, podemos concluir que a análise mencionada traz importantes melhorias tornando-se, desta forma, uma importante ferramenta para a otimização dos resultados finais.
Palavras-Chaves: Mineração de Dados, Processo *KDD*, Análise Multivariada.

Abstract
Knowledge Discovery in Databases – KDD – process is composed of several phases starting with the data collection for the problem and ending with the evaluation of the final results. This work aims in showing the data multivariate analysis influence in the Data Mining techniques performance for classification of new patterns. As indicated in this work, we may conclude that the mentioned analysis brings important improvements becoming, by this way, an important tool for the optimization of the final results.
Keywords: Data Mining, KDD Process, Multivariate Analysis.

1. Introdução

Abordar técnicas e ferramentas que buscam transformar os dados armazenados, sejam de empresas, hospitais, telecomunicações 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 5 etapas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (*"Data Mining"*); interpretação e avaliação dos resultados. As 3 primeiras etapas podem ser vistas como a análise multivariada dos dados.

KDD refere-se a todo processo de descoberta de conhecimento útil de dados, enquanto *"Data Mining"* refere-se a 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 três propriedades: deve ser correto (tanto quanto possível); deve ser compreensível por usuários humanos; deve ser interessante / útil / novo (surpreendente). Ainda, segundo Freitas, 2000, o método de descoberta do conhecimento deve apresentar as seguintes características: deve ser eficiente, genérico (ou seja, aplicável a vários tipos de dados) e flexível (facilmente modificável).

O objetivo do presente trabalho é mostrar, através de um problema médico real, apresentado na seção 2, a importância da fase de análise multivariada dos dados quanto a classificação no contexto de *KDD*, nas seções 3 e 4 e, finalmente na seção 5, são apresentadas as conclusões.

2. Descrição do Problema Médico

A icterícia (do grego ikteros = amarelidão) 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:

a. 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; b. outras causas.

Somente o 1° grupo - das colestases - será objeto de estudo. 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: a1. obstrução por câncer; a2. 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, cerca 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 cálculos 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, Brasil, dos quais 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. Estes 14 atributos foram os seguintes (os quais serão apenas listados): 1.

Idade (id); 2. Sexo (sex); 3. Bilirrubina Total (bt); 4. Bilirrubina Direta (bd); 5. Bilirrubina Indireta (bi); 6. Fosfatases alcalinas (fa); 7. SGOT (sgot); 8. SGPT (sgpt); 9. Tempo de atividade da protrombina (tap); 10. Albumina (alb); 11. Amilase (ami); 12. Creatinina (cr); 13. Leucócitos (le); 14. Vg (vg). Para um melhor detalhamento, ver Steiner, 1995.

3. Técnicas utilizadas neste Trabalho

As cinco etapas do processo *KDD* enumeradas na seção 1 deste trabalho, e ilustrada na Figura 1 a seguir, foram aqui abordadas, resumidamente, em 3 etapas distintas: análise multivariada dos dados; *Data Mining* e obtenção e análise dos resultados.

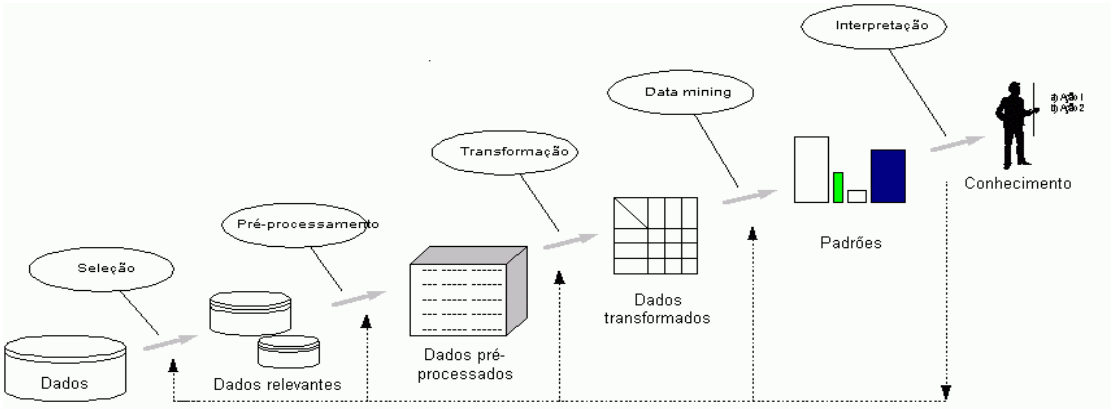

Figura 1. Etapas do Processo *KDD* (FAYYAD et al., 1996)

3.1 Análise Multivariada dos Dados

A complexidade de muitos problemas exige que o pesquisador colete muitas observações contendo, cada uma delas, muitos variáveis (atributos). A análise multivariada objetiva utilizar métodos estatísticos para captar informações destes dados e, pelo fato dos dados incluírem medidas simultâneas sobre muitas variáveis, esta metodologia é chamada de análise multivariada (Johnson & Wichern, 1988). Esta ampla área de estudo envolve inúmeras técnicas estatísticas, sendo que para este trabalho específico, esta análise ficou composta das seguintes técnicas, que foram aplicadas seqüencialmente: teste T^2 de Hotelling; transformação dos atributos e descarte dos dados atípicos.

Nestas técnicas, descritas a seguir, tem-se que A e B são os conjuntos de dados dos dois conjuntos a serem classificados (colestáticos com câncer e colestáticos com cálculo, respectivamente), m é a quantidade de dados do conjunto A, k é a quantidade de dados do conjunto B, n é o número de atributos e S_p é a matriz de covariância conjunta de A e B.

Inicialmente foi aplicado o teste ***T^2* de Hotelling** (Johnson & Wichern, 1988) para verificar a igualdade dos vetores médios das duas populações multivariadas A e B. Para a aplicação deste teste, calcula-se:

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

onde:

$$T^2 = (\underline{x_A} - \underline{x_B})'[(\frac{1}{m} + \frac{1}{k})S_p]^{-1}(\underline{x_A} - \underline{x_B})$$

O resultado deve ser comparado com a distribuição *F*, *Fn,m+k-n-1 (0.95)*. Nesta expressão tem-se que:

$\underline{x_A}$ = *vetor* (nx1) *médio amostral da população* 1:

$$\underline{x_A} = \frac{1}{m}\sum_{j=1}^{m} x_{Aj}$$

$\underline{x_B}$ = *vetor* (nx1) *médio amostral da população* 2:

$$\underline{x_B} = \frac{1}{k}\sum_{j=1}^{k} x_{Bj}$$

S_p^{-1} = *inversa da matriz de* covariância *amostral conjunta*:

$$S_p = \frac{(m-1)S_A + (k-1)S_B}{m+k-2}$$

S_A = *matriz de covariância amostral de 1*:

$$S_A = \frac{1}{m-1}\sum_{j=1}^{k}(\underline{x_{Aj}} - \underline{x_A})(\underline{x_{Aj}} - \underline{x_A})'$$

S_B = *matriz de covariância amostral de 2*:

$$S_B = \frac{1}{k-1}\sum_{j=1}^{k}(\underline{x_{Bj}} - \underline{x_B})(\underline{x_{Bj}} - \underline{x_B})'$$

Neste teste se:

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

rejeita-se, fortemente, com uma probabilidade de 95%, a hipótese de que as populações estão centradas no mesmo vetor de médias.

Na **transformação dos atributos**, como o que se desejava era o prognóstico (*1* ou *0*, padrões pertencentes ao conjunto *A* ou *B*) com o menor erro possível, analisou-se o relacionamento da variável resposta com os 14 atributos originais e outros coatributos derivados destes com base na função desvio. Conforme o valor da função desvio fosse estatisticamente significativo ou não, o coatributo era incorporado ao modelo. Alguns coatributos foram transformados, na escala, 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, procura-se modelos com desvios moderados.

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, da seguinte forma:

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

onde Yi é o valor assumido pelo atributo no modelo saturado e θ_i é a estimativa deste valor feita pelo modelo. Um valor de $|e_i| \geq 1$ indica que o dado *i* está sendo classificado erroneamente pelo modelo, ou seja, a observação i encontra-se "deslocada" em relação ao seu conjunto (*A* ou *B*), o que a caracteriza como atípica, sendo que nestes casos, dependendo da justificativa para tal ocorrência, o referido dado *i* poderá ser descartado.

3.2 Técnicas de *Data Mining*

A escolha das técnicas de *Data Mining* depende fundamentalmente do objetivo do processo de *KDD* (Witten & Frank, 2000), que pode ser tanto classificação, quanto agrupamento ou associação de exemplos. Neste trabalho o objetivo se atém a classificação, ou seja, distinguir exemplos (padrões, dados) de clientes com câncer ou com cálculo no duto biliar. Para tanto, fez-se a abordagem de 4 técnicas de *Data Mining* capazes de fazer a classificação (discriminação) entre os conjuntos *A* e *B*: uma técnica que faz uso da Programação Linear; uma técnica estatística; uma que utiliza Redes Neurais e a última que utiliza Árvores de Decisão, descritas sucintamente a seguir.

3.2.1 Geração de uma Superfície que Minimiza Erros

Bennett e Mangasarian propuseram em 1992 uma formulação de um único programa linear que gera um plano que minimiza a média ponderada da soma das violações dos pontos dos conjuntos *A* e *B* que estão do lado errado do plano separador (de *A* e de *B*). 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 através do seguinte modelo de Programação Linear, onde $e = (1, 1, ..., 1) \in R^n$; w = vetor "peso" $\in R^n$, normal ao plano separador ótimo; γ = número real, fornece a localização do plano separador $wx = \gamma$.

$$\underset{w,\gamma,y,z}{Min}\ \frac{ey}{m}+\frac{ez}{k}$$

$$\begin{aligned} s.a:\ & Aw - e\gamma + y \geq e \\ & -Bw + e\gamma + z \geq e \\ & y \geq 0 \\ & z \geq 0,\ y \in R^m,\ z \in R^k \end{aligned}$$

Observe-se que se trata de um método não iterativo, ou seja, o plano separador obtido através deste modelo é único para os dois conjuntos dados *A* e *B*. Para maior detalhamento consultar Bennett & Mangasarian, 1992 ou Steiner, 1995.

3.2.2 Função Discriminante Linear (FDL) de Fisher

A terminologia "discriminar" e "classificar" foi introduzida na Estatística por Ronald A. Fisher no primeiro tratamento moderno dos problemas de separação de conjuntos na década de 30 (Johnson & Wichern, 1988). Dadas duas amostras de duas populações *1* e *2* de observações multivariadas *X's* (conjuntos *A* e *B*, no caso do presente trabalho), de dimensão *n*, a idéia de Fisher foi transformar estas observações multivariadas em observações univariadas *Y's* de tal modo que estejam separadas tanto quanto possível. Fisher propôs o uso de combinação linear das *n* variáveis aleatórias componentes de $\underline{X}$ para obtenção dos *Y's*, pelo fato da combinação linear ser de fácil obtenção matematicamente.

A FDL de Fisher amostral, considerando-se as amostras *A* de tamanho *m* e *B* de tamanho k, respectivamente das populações *1* e *2*, é obtida por:

$$Y = (\underline{x_A} - \underline{x_B})' S_p^{-1} \underline{x}$$

onde $\underline{x}$ = vetor das variáveis aleatórias correspondentes as características populacionais observadas. Para classificação de um novo vetor $\underline{x_0}$ (novo padrão), a regra de decisão é a seguinte:

se $\underline{x_0} \in A$ então:

$$y_0 = (\underline{x_A} - \underline{x_B})' S_p^{-1} \underline{x_0} \geq q = \frac{1}{2}(\underline{x_A} - \underline{x_B})' S_p^{-1} (\underline{x_A} + \underline{x_B})$$

se $\underline{x_0} \in B$ então:

$y_0 < q$.

Para um melhor detalhamento ver em Johnson & Wichern, 1988.

3.2.3 Redes Neurais

As Redes Neurias de Múltiplas Camadas (RNMC) utilizadas neste trabalho ou, também chamadas, Redes do tipo "Alimentadas para a Frente" ("*Feed-Forward*") constituem o modelo mais utilizado pela comunidade científica e apresenta, de uma forma geral, resultados bastante satisfatórios. O algoritmo de Retro-Propagação (*"Back-Propagation"*), utilizado para o seu treinamento, é um algoritmo de aprendizado supervisionado (Kröse & Van der Smagt, 1993), (Fausett, 1995).

Durante o treinamento 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 *A* ou *B* 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 para esse padrão e o erro é calculado. O erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo ajustados conforme o erro é retro-propagado (Kröse & Van der Smagt, 1993). Uma iteração é completada quando todos os *(m + k)* padrões tiverem sido apresentados à rede. Para os dados apresentados, a Rede Neural precisou de, aproximadamente, 1.000 iterações para convergir em cada uma das situações de teste apresentadas na seção 4 mais adiante.

Os dados fornecidos pelo HC serviram para treinar uma Rede Neural com 3 camadas (camada de entrada, com o número de neurônios de entrada igual ao número de informações (14 ou 13, matriz sem ajuste e ajustada, respectivamente), camada intermediária contendo um número de neurônios variando de 1 a 20 conforme metodologia apresentada por Steiner, 1995, e camada de saída, com 1 neurônio (paciente com câncer ou paciente com cálculo).

3.2.4 Árvores de Decisão

Ross Quinlan (Quinlan, 1993), da Universidade de Sydney, Austrália, desenvolveu a técnica que permitiu o uso da representação do conhecimento através das Árvores de Decisão. A sua contribuição consistiu na elaboração de um algoritmo chamado *ID3*. Este, 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 mostradas 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 (Gama, 2002). 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.

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 SE-ENTÃO *(IF-THEN).*

A utilização dessa abordagem apresenta as seguintes vantagens: não assumem nenhuma distribuição particular para os dados; as características ou atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos); pode-se construir modelos para qualquer função desde que o número de exemplos de treinamento seja suficiente; elevado grau de compreensão.

Neste trabalho, os dados do HC serviram para formar uma árvore de decisão utilizando o algoritmo de classificação em árvore *J48 (C4.5 release 8*), conforme resultados apresentados na seção 4.

4. Implementação das Técnicas ao Problema Médico

Conforme descrito a seguir, os testes computacionais foram feitos em duas matrizes de dados: em uma primeira matriz, M1, contendo os dados originais, na qual foi aplicado apenas o teste T^2 de Hotelling (ou seja, em M1 não houve ajuste nos dados) e, em uma segunda matriz, M2, na qual os dados, além de serem analisados pelo teste T^2 de Hotelling, tiveram seus atributos transformados e seus pontos atípicos descartados (desta forma, M2 contém dados ajustados estatisticamente).

Para a aplicação do teste T^2 de Hotelling foi desenvolvido um programa computacional em *Visual Basic*. Como resultado da aplicação desse programa, foi verificado que as populações, pacientes com câncer e pacientes com cálculo são distintas a um nível de significância de 5%, ou seja, com uma probabilidade de 95% de acerto pode-se dizer que as populações anteriormente referidas são distintas. A aplicação deste teste se justifica pela necessidade de se mostrar quantitativamente as evidências da distinção entre as populações.

Para a transformação dos atributos utilizou-se o pacote estatístico **GLIM** (*Generalized Linear Interactive Modeling*) (Mardía et al., 1979), tendo como parâmetros, os atributos originais. Este *software* também foi usado para o possível descarte de pontos, analisando os resíduos de cada um dos pontos, sendo que neste procedimento foram descartados 7 pontos, 6% do total, pontos estes considerados atípicos. Esta hipótese foi assumida após discussão com os especialistas da área e apontadas as causas. Todos estes procedimentos se seguiram conforme seção 3.1. Os pontos atípicos tiveram as suas origens investigadas, e quando a área técnica ligada ao processo assumia que a natureza do ponto era externa ao processo, ele era eliminado. Após estas etapas a matriz M2 ficou definida com os seguintes **13 atributos**: id, bt, bd2 (= bd.bd), bd, ami, lnam (= loge ami), st2 (= st.st), st (=sgot/sgpt), fa, fa2n (=fa.fa/1000), vg, vg2 (= vg.vg), bt2 (= bt.bt). Observe-se que foram decartados os atributos: sex, bi, tap, alb, cr, le.

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 M1 (dados sem ajuste) e M2 (dados com ajuste), forneceu os seguintes valores: M1: $4.84 > 1.78896 = F_{14,103}\ (0.95)$; M2: $12.54 > 1.82239 = F_{13,97}(0.95)$. Consequentemente, rejeita-se fortemente, nos dois casos, especialmente em M2, a hipótese de que as populações estejam centradas no mesmo vetor de médias. Assim, a população de ictéricos cancerosos é distinta da de ictéricos com cálculo nos atributos estudados.

Tendo-se as matrizes M1 (matriz de dados sem ajuste, da ordem 118x14) e M2 (matriz com os dados ajustados, da ordem 111x13), fez-se a aplicação dos quatro métodos de *Data Mining*. Para o 1° método, utilizou-se o *software* comercial ***LINGO*** (*Language for Interactive General Optimizer*) para a resolução do modelo de Programação Linear; para o 2° (Fisher) e 3° (RNeurias) métodos, fez-se a implementação computacional em *Visual Basic*; para o 4° e último método, utilizou-se o *software* livre

WEKA (*Waikato Environment for Knowledge Analysis*, disponível no site *www.cs.waikato.ac.nz/ml/weka*).

A metodologia para testes em todos os 4 métodos consistiu em aplicar o método *holdout* estratificado (Witten & Frank, 2000) repetido 3 vezes. Para isso, dividiu-se os conjuntos de pontos *A* e *B* aleatoriamente em 2 subconjuntos: um dos subconjuntos, denominado "Conjunto para Treinamento" serviu para "treinar" o programa, e o outro subconjunto, "Conjunto para Teste" serviu para testar o "modelo treinado". Este procedimento foi repetido 3 vezes, variando-se os dois subconjuntos e a média das percentagens dos erros foi calculada e é apresentada no quadro 1. Observe-se que o "Conjunto de Treinamento" foi usado para testar os métodos também.

As 4 técnicas de *Data Mining* foram aplicadas nas matrizes M1 (118x14) e M2 (111x13) separadamente para poder-se comparar a performance das mesmas sobre os dados originais e sobre os dados ajustados.

Métodos	**Progr. Linear**		**F.D.L. Fisher**		**Redes Neurais**		**Árv. Dec.**	
Dados	Conjunto para Treinam.	Conjunto para Teste	Conjunto para Treinam.	Conjunto para Teste	Conjunto para Treinam.	Conjunto para Teste	Conjunto para Treinam.	Conjunto para Teste
M1	16.35	24.99	18.87	19.44	23.0	27.67	4.09	22.22
M2	0	3.03	6.66	9.09	14.33	3.03	2.33	24.25

Quadro 1. Média das percentagens dos erros (3 testes) dos métodos no caso do exemplo médico, com a matriz sem ajuste M1 (118x14) e com a matriz ajustada M2 (111x13)

5. Conclusões

Analisando-se os resultados contidos no Quadro 1 nota-se que todos os métodos, com exceção do Conjunto para Teste da Árvore de Decisão, apresentaram uma melhoria significativa na sua performance com a adoção da análise estatística multivariada dos dados, preliminarmente à aplicação das técnicas de *Data Mining*. Este fato enfatiza a importância de se ter dados confiáveis e consistentes e, por conseguinte, dados multivariados analisados estatisticamente.

Além disso, pode-se observar que ao se utilizar a matriz M1 (dados originais, ou seja, sem ajuste) para verificar a performance de todos os 4 métodos, apenas a árvore de decisão (J48) é capaz de eliminar atributos considerados dispensáveis, sendo que os mesmos variam a cada teste efetuado. Apesar destes variarem de um teste para o outro, os atributos bt, sgot, tap, cre e le foram considerados dispensáveis nos 3 testes. Já ao se utilizar a matriz M2, os atributos descartados pela árvore de decisão foram: bd, st2, fa2n, vg2 e bt2. Note-se que para a matriz M1, dos 5 atributos descartados nos 3 testes pela árvore de decisão, 3 deles (tap, cr e le) coincidem com os atributos descartados pela análise multivariada que originou a matriz M2 (com ajuste nos dados).

Vale destacar ainda que de todas as técnicas de *Data Mining* aqui abordadas, apenas a árvore de decisão deixa claro ao usuário (compreensibilidade) quais são os atributos que estão discriminando os padrões e de que forma (pontos de corte) a mesma está ocorrendo, como pode-se observar pela Figura 2, que exemplifica um dos testes feitos. O percentual de erros relativamente alto desta técnica, se comparado com o das demais, é compensado por esta característica altamente desejável

(compreensibilidade). A técnica que envolve Redes Neurais, também poderia tornar-se compreensível, bastando para isto utilizar algum algoritmo de extração de regras a partir da rede neural treinada, conforme apresentado por (Lu et al., 1996), (Santos et al., 2000) e outros.

De qualquer forma, todas as técnicas apresentaram performances satisfatórias, podendo ser utilizadas em casos reais como o caso médico abordado neste trabalho. Desta forma, pode-se oferecer ao especialista da área (ao médico, no caso deste trabalho), um sistema computacional contendo as técnicas aqui apresentadas (análise estatística dos dados e técnicas de *Data Mining*) como uma ferramenta adicional para uma melhor prescrição de seus diagnósticos.

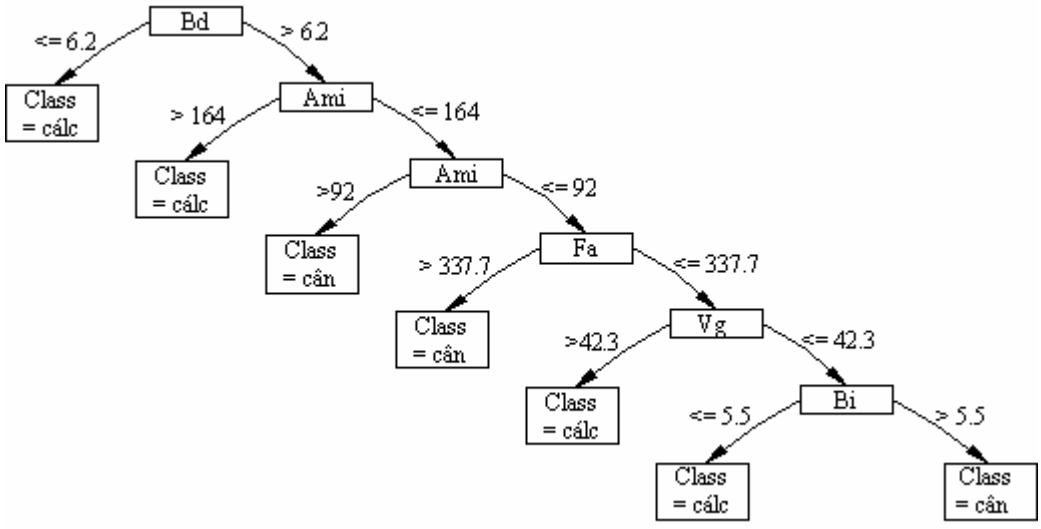

Figura 2. Um Exemplo de Árvore de Decisão para o Problema Médico abordado.

Referências:

BENNETT, K. P. & MANGASARIAN, O. L. "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", *Optimization Methods and Software*, 1992, vol.1, p. 23-34.

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

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R. *Advances in Knowledge Discovery & Data Mining.* AAAI/MIT, 1996.

FREITAS, A. A. *Uma Introdução a Data Mining. Informática Brasileira em Análise*. CESAR - Centro de Estudos e Sistemas Avançados do Recife. Ano II, n. 32, mai./jun. 2000.

GAMA, J. Árvores de Decisão, 2000. Disponível em: <http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html> Acesso em: 14 ago. 2002.

JOHNSON, R. A. & WICHERN, D. W. *Applied Multivariate Statistical Analysis*. New Jersey, Prentice-Hall, inc., 1988.

KRÖSE, B. J. A. & VAN DER SMAGT, P. P. *An Introduction to Neural Networks*. Amsterdam, University of Amsterdam, 1993.

LU, H.; SETIONO, R. and LIU H. "Effective Data Mining using Neural Networks". *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, 1996, p. 957-961.

MARDÍA, K. V.; KENT, J. P. & BIBBY, J. M. *Multivariate Analysis*. London, Academic Press, 1979.

QUINLAN, J. C. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.

SANTOS, R. T.; NIEVOLA, J. C. e FREITAS, A. A. "Extracting Comprehensible Rules from Neural Networks via Genetic Algorithms". *IEEE*, 2000, p. 130-139.

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

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