DATA MINING COMO SUPORTE À TOMADA DE DECISÕES - UMA APLICAÇÃO NO 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

Fábio Mendonça Lopes e Andréia Smiderle
Doutorandos do Programa de Pós-Graduação em Métodos Numéricos em Engenharia-UFPR,
CP: 19081-CEP: 81531-990, Curitiba, PR; fmendoncal@uol.com.br; andreiasmiderle@brturbo.com.br

Resumo
Na área médica, a posse e uso de ferramentas que auxiliem na tarefa de classificação de pacientes em prováveis ictéricos com câncer ou ictéricos com cálculo, pode se tornar um fator crucial, em uma tentativa de otimizar todo o processo do diagnóstico, minimizando os riscos e os custos aos pacientes e, por outro lado, maximizando a eficácia nos resultados.
Com os métodos de *Data Mining* pode-se transformar os dados coletados dos pacientes em informações valiosas para auxiliar no processo decisório. Neste trabalho são analisados registros históricos de 118 pacientes do Hospital das Clínicas da cidade de Curitiba, PR, através de seis ferramentas de *Data Mining:* 3 delas enquadradas em Árvores de Decisão e as outras 3 em Regras de Classificação.
Estas técnicas permitem fazer o reconhecimento de padrões e a posterior classificação de novos casos. Os resultados foram satisfatórios, mostrando que, para este problema específico, as técnicas de Árvores de Decisão apresentaram uma taxa de classificação correta um pouco maior do que as técnicas de Regras de Classificação.
Palavras-Chaves: *Data Mining*, Árvores de Decisão, Regras de Classificação, Diagnóstico Médico.

Abstract
In medical studies, the search for techniques used to help in the task of classifying patients with cholestasis by cancer or cholestasis by gallstones might be a key element in optimizing the diagnosis process as a whole, minimizing risks and costs for patients and, enhancing the effectiveness of the results.
By the use of Data Mining, one can transform data collected from patients into valuable information to help in the decision process. In this paper historical data obtained from 118 patients at Hospital de Clínicas, the largest hospital in Curitiba, capital of the state of Paraná, are analyzed with the help of six Data Mining tools; 3 based on Decision Trees and 3 on Classification Rules.

These techniques were used to perform pattern recognition and classify new cases. The results were considered satisfactory for all methods tested and, at least for this specific situation, Decision Trees techniques provided a correct classification rate slightly greater than Classification Rule ones.
Keywords: Data Mining, Decision Trees, Classification Rules, Medical Diagnosis.

1. Introdução

Mineração de Dados ("*Data Mining*") é a principal etapa de um amplo processo denominado Descoberta de Conhecimentos em Bases de Dados ("*Knowledge Discovery in Databases – KDD*"). O objetivo do *KDD* é abordar técnicas e ferramentas que buscam transformar os dados armazenados, sejam de empresas, hospitais, telecomunicações e outros, em conhecimento (informação útil), enquanto que *Data Mining* refere-se a aplicação de algoritmos para extrair modelos dos dados.

Todo o processo de *KDD* é composto de um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados, formado por 5 etapas (FAYYAD et al., 1996): seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados ("*Data Mining*"); interpretação e avaliação dos resultados. Estas etapas e suas interconexões estão ilustradas na Figura 1 a seguir.

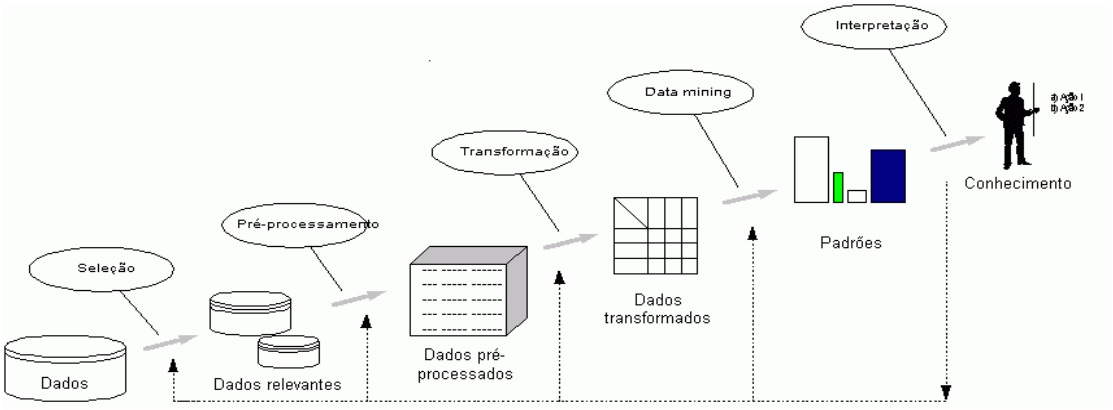

Figura 1. Processo *KDD* e o enquadramento da etapa de *Data Mining* (FAYYAD et al., 1996)

Segundo Freitas (FREITAS, 2000), o conhecimento a ser descoberto (processo *KDD*) deve satisfazer três propriedades: deve ser correto (tanto quanto possível); deve ser compreensível por usuários humanos; deve ser interessante / útil / novo. Além disso, os métodos de extração de padrões (métodos de *Data Mining*) devem apresentar as seguintes características: devem ser eficientes, genéricos (ou seja, aplicáveis a vários tipos de dados) e flexíveis (facilmente modificável).

O objetivo do presente trabalho é mostrar, através de um problema médico real descrito na seção 2 deste trabalho, a aplicação de seis técnicas de *Data Mining*, 3 de Árvores de Decisão e 3 de Regras de Classificação fazendo a comparação de performance entre as mesmas quanto a tarefa de classificação.

2. Descrição do Problema Médico (STEINER, 1995)

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:

- 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.
- Outras causas: São os casos em que o fígado traduz distúrbios sistêmicos, como anemias hemolíticas, hepatites, etc.

Somente o 1º grupo - das colestases - será objeto de estudo. Para fazer esta distinção inicial, o clínico se apoia geralmente em exames simples e rotineiros, que traduzem essencialmente as repercussões bioquímicas do obstáculo ao fluxo biliar, 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 cálculos;
- Obstrução por câncer.

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%. Mesmo que se visualizem cálculos com esses exames, há freqüentemente superposição de dados e até concomitância de doenças. Tem-se como exemplo, cálculos e câncer de vesícula biliar.

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. Um dos procedimentos mais utilizados é a colangiografia endoscópica retrógrada. Neste caso, quando há cálculos e estes são retirados por via endoscópica o risco de complicações é pequeno, porém quando ocorrem, a mortalidade atinge os 20%. Além disso, as complicações indiretas provocadas pela introdução de germes na árvore biliar obstruída podem ser significativamente superiores.

A outra alternativa, também de risco, é realizada através da injeção de contraste por via transparieto-hepática, ou ainda, por via transvesical. As biópsias por aspiração têm risco relativamente pequeno porém, é necessário que haja uma equipe muito bem treinada para evitar erros proibitivos de interpretação dos resultados.

Tendo em vista o risco do paciente e os altos custos envolvidos para um diagnóstico adequado, tem-se a justificativa para a utilização de técnicas de *Data Mining* 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 foram coletados dados históricos dos pacientes enquadrados nos dois casos anteriores os quais estão descritos na seção 2.1 a seguir.

2.1 Coleta de Dados

Para o desenvolvimento deste trabalho foram considerados dados de 118 pacientes, sendo que já se sabia que 35 pacientes possuíam câncer e 83 possuíam cálculo no duto biliar.

Consideraram-se 14 variáveis (medidas de exames clínicos) para cada um desses 118 pacientes, sugeridas por especialista da área: (entre parêntesis encontra-se a simbologia usada para cada variável)

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 consequências 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 que 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.;
14. Vg (vg): É a medida do volume globular.

Os conjuntos de dados dos pacientes ictéricos com câncer e ictéricos com cálculo, os quais objetiva-se discriminar, possuem respectivamente, 35 e 83 dados (padrões / exemplos).

3. Técnicas de *Data Mining* utilizadas neste Trabalho

A escolha das técnicas de *Data Mining* depende fundamentalmente do objetivo do processo de *KDD* (BERRY e LINOFF, 2000), que pode ser classificação, agrupamento ou associação de exemplos. Neste trabalho o objetivo se atém a classificação, ou seja, distinguir exemplos de clientes com câncer dos pacientes com cálculo no duto biliar. Para tanto, fez-se a abordagem de seis técnicas de *Data Mining* capazes de fazer a classificação (discriminação) entre os conjuntos: 3 enquadradas em Árvores de Decisão e 3 em Regras de Classificação.

3.1 Árvores de Decisão

As Árvores de Decisão utilizam a estratégia *dividir-e-conquistar* ("*divide-and-conquer*"), onde as árvores são construídas utilizando-se de apenas alguns atributos. As Árvores de Decisão são uma das técnicas de aprendizado de máquina ("*machine learning*"), onde um problema complexo é decomposto em subproblemas mais simples. Recursivamente a mesma estratégia é aplicada a cada sub-problema (GAMA, 2002).

Quinlan, da Universidade de Sidney, é considerado o "pai das Árvores de Decisão". A sua contribuição foi a elaboração de um novo algoritmo chamado *ID3*, desenvolvido em 1983. O *ID3* e suas

evoluções (*ID4*, *ID6*, *C4.5*, *See 5*) são algoritmos muito utilizados para gerar Árvores de Decisão. O atributo mais importante é apresentado na árvore como o primeiro nó, e os atributos menos importantes, segundo o critério utilizado, são mostrados nos nós subseqüentes. As vantagens principais das Árvores de Decisão são que elas "tomam decisões" levando em consideração os atributos que são considerados mais relevantes, segundo a métrica escolhida, 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.

Segundo Garcia (GARCIA, 2000), as Árvores de Decisão consistem de:

- nodos (nós) que representam os atributos;
- arcos (ramos), provenientes dos nodos e que recebem os valores possíveis para estes atributos (cada ramo descendente corresponde a um possível valor deste atributo) e
- nodos folha (folhas da árvore), que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe.

Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

O problema de construir uma Árvore de Decisão pode ser expresso recursivamente: primeiro selecione um atributo para colocar no nó raiz e faça um ramo para cada possível valor. Isto divide o problema em sub-conjuntos, um para cada valor do atributo. Agora o processo pode ser repetido recursivamente para cada ramo. Se a qualquer instante todos os exemplos em um nó têm a mesma classificação, pare de desenvolver aquela parte da árvore. Como determinar, no entanto, qual atributo dividir? Escolhe-se o atributo que gere uma árvore menor e que tenha chances de classificar melhor, ou seja, precisamos medir o grau de pureza de cada nó. Com isto, poderemos escolher o atributo que produz os nós filhos mais puros (CARVALHO, 2002). A medida de pureza mais utilizada é chamada de informação e é medida em *bits*.

Construção de uma Árvore de Decisão

O processo de construção de uma Árvore de Decisão inicia-se a partir de um conjunto de treinamento, que contém exemplos com classes previamente conhecidas (dados históricos).

Para gerar uma árvore de decisão com uma alta taxa de predição é necessário fazer a escolha correta dos atributos que serão usados como teste no agrupamento dos casos. Estes testes devem gerar uma árvore com o menor número possível de subconjuntos, fazendo com que cada folha da árvore contenha um número significativo de casos. O ideal é escolher os testes de modo que a árvore final seja a menor possível.

Como analisar todas as possibilidades seria algo absurdo, foram desenvolvidos vários métodos aplicados na escolha dos atributos e dos testes a serem utilizados, sendo que todos concordam em dois pontos: uma divisão que mantém as proporções de classes em todas as partições é inútil e uma divisão onde em cada partição todos os exemplos são da mesma classes tem utilidade máxima. Uma vez feita a escolha, as outras possibilidades não são mais exploradas (LEMOS, 2003).

Para melhor esclarecer os critérios que levam à escolha de um atributo, faz-se necessário o conhecimento de dois conceitos: **Entropia** e **Ganho de Informação (**CARVALHO, 2000**).**

Entropia: É a medida que indica a homogeneidade dos exemplos contidos em um conjunto de dados. Ela permite caracterizar a "pureza" (e impureza) de uma coleção arbitrária de exemplos (OSÓRIO, 2000).

Dado um conjunto *S* contendo exemplos positivos ("+") e exemplos negativos ("-") que definem o conceito a ser aprendido, a entropia relativa dos dados deste conjunto *S* é indicada pela expressão (3.1) a seguir (WITTEN e FRANK, 2000):

(3.1) $$Entropia(S) = - P_{(+)} \cdot log_2 P_{(+)} - P_{(-)} \cdot log_2 P_{(-)}$$

onde:

$P_{(+)}$ = Proporção entre os exemplos positivos e o total de exemplos do conjunto, ou seja, número de casos positivos / número total de casos.

$P_{(-)}$ = Proporção entre os exemplos negativos e o total de exemplos do conjunto, ou seja, número de casos negativos / número total de casos.

É assumido que: $0 \cdot log_2 0 = 0$, por definição.

A equação (3.1) apresentada é usada para calcular a entropia levando-se em conta duas classes. Fazendo a generalização para "N" Classes, tem-se a a equação (3.2):

(3.2) $$Entropia\ (S) = - \sum_{i=1}^{N} P_i \ log_2 \ P_i$$

A *Entropia (S)* tem máximo valor para $(log_2 P_i)$ se $P_i = P_j$ para qualquer $i \neq j$ (caso em que o número de casos positivos é igual ao número de casos negativos) e a *Entropia (S) = 0*, se existe um *i* tal que $P_i = 1$ (caso em que todos os exemplos são da mesma classe).

Ganho de Informação (Critério *GAIN*): Segundo OSÓRIO, 2000, o ganho de informação é a medida que indica o quanto um dado atributo irá separar os exemplos de aprendizado de acordo com a sua função objetivo (classes). O ganho de informação é a redução esperada no valor da Entropia devido à ordenação do conjunto de treinamento segundo os valores do atributo escolhido (ANTUNES, 2000).

GAIN (S, A) = Redução esperada na entropia de *S*, causada pelo particionamento dos exemplos em relação a um atributo escolhido (*A*).

(3.3) $$Gain\ (S, A) = Entropia\ (S) - \sum_{v=1}^{N} \frac{|S_v|}{|S|} \cdot Entropia(S_v)$$

onde:

A = Atributo considerado; *N* = Número de valores possíveis que este atributo pode assumir;

S_v = Sub-conjunto de *S* onde o atributo *A* possui o valor *V*.

O método de particionamento recursivo para geração das Árvores de Decisão, que subdivide o conjunto de casos de treinamento até que cada subconjunto em cada partição contenha casos de uma única classe ou até que nenhum outro teste ofereça qualquer melhora, pode gerar árvores complexas que acabam perdendo o seu poder de generalização. Faz-se necessário então, adotar algumas medidas para transformar árvores complexas em árvores mais simples (QUINLAN, 1993).

Existem dois caminhos pelos quais este particionamento recursivo pode ser modificado para produzir árvores mais simples: decidindo não continuar a dividir o conjunto de dados de treinamento ou removendo retrospectivamente alguma estrutura já construída pelo método. O primeiro caminho pode causar o

término da divisão antes que o benefício das divisões subseqüentes se tornem evidentes. Na segunda alternativa, o processo de *dividir-e-conquistar* segue até o fim e então, a árvore é "podada". Este processo é mais lento, mas muito mais seguro. O processo de poda irá, em geral, causar união de alguns exemplos de classes diferentes em um mesmo nó.

Pode-se dizer que uma das maiores motivações para podar Árvores de Decisão é no sentido de se evitar o ajuste demasiado / sobreajuste ("*overfitting*") da árvore aos dados. Neste caso, a árvore poderia se ajustar a peculiaridades dos dados, que talvez não ocorram em dados ainda não vistos (FREITAS, 2000). Ainda segundo FREITAS, 2000, deve-se, no entanto, ter cuidado para que a poda não seja muito agressiva, a fim de não gerar um sub-ajustamento ("*underfitting*") da árvore aos dados. A realização da "poda" ou simplificação das Árvores de Decisão é baseada em "erros".

Na Figura 2 a seguir é apresentada um exemplo de Árvore de Decisão para o problema real de diagnóstico médico apresentado na seção 2 deste trabalho.

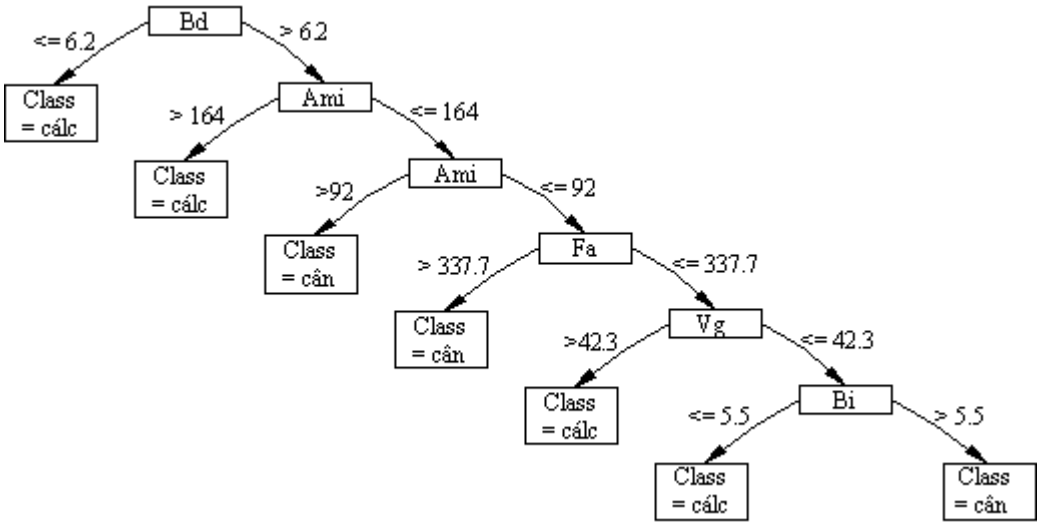

Figura 2. Um Exemplo de Árvore de Decisão para o Problema Médico abordado.

Informações adicionais sobre Árvores de Decisão: Muitos são os algoritmos de classificação que constróem Árvores de Decisão. Não há uma forma de determinar qual é o melhor algoritmo, sendo que um algoritmo pode ter melhor desempenho em determinada situação e outro pode ser mais eficiente em outros tipos de situações.

A utilização de Árvores de Decisão apresenta as seguintes vantagens: não assume nenhuma distribuição particular para os dados; as características ou atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos); pode construir modelos para qualquer função desde que o número de exemplos de treinamento seja suficiente; possui elevado grau de interpretabilidade.

Após a construção de uma Árvore de Decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento. Esta estratégia permite estimar como a árvore generaliza os dados e se adapta a novas situações, podendo, também, se estimar a proporção de erros e acertos ocorridos na construção da árvore (BRAZDIL, 2002).

Os algoritmos que geram Árvores de Decisão diferem em pequenos diversos aspectos, sempre na tentativa de obter árvores eficazes e simples para solucionar o problema. O algoritmo *J4.8*, por exemplo, é

a implementação em Java da Árvore de Decisão *C4.5*. Existe ainda uma versão melhorada da *C4.5* que é *C4.5 Revision 8*, que é a última versão pública desta família de algoritmos antes do *C5.0*, uma implementação comercial (WITTEN e FRANK, 2000). A essência do algoritmo para a Árvore de Decisão *C4.5* trabalha de acordo com o que foi relatado no decorrer desta sub-seção 3.1, sendo que o mesmo apresenta resultados satisfatórios para a maioria dos problemas.

O caso mais simples de árvore de decisão é chamado de "*stump*", o qual consiste em um único nó e dois nós folhas para predição. Da mesma forma que nas Árvores de Decisão, em uma Árvore de Decisão Alternada ("*Alternating Decision Tree*" ou "*ADTree*") um exemplo está mapeado em um caminho ao longo da árvore, desde a raiz até um dos nós folha. Entretanto, ao contrário das Árvores de Decisão, a classificação que é associada com o caminho não é o rótulo do nó folha, mas é obtida através da soma das previsões ao longo do caminho. Uma *ADTree* recebe este nome por ser composta de camadas alternadas de nós de previsão e de nós de particionamento. Uma das características mais interessantes da *ADTree* é que, além da classificação, ela fornece também uma medida de confiança denominada de margem.

Já uma Árvore de Decisão "Firme" ("*Decision Stump*") é uma Árvore de Decisão com apenas um particionamento. Ela é um algoritmo de aprendizagem muito simples, mas que é útil para ser usado como base de comparação com outros algoritmos. Duas das grandes vantagens do algoritmo *Decision Stump* são que ele aprende em um tempo que é proporcional ao número de exemplos de treinamento e que ele requer um espaço de memória proporcional ao número de classes vezes o número de atributos vezes o número de valores.

3.2 Regras de Classificação

As Regras de Classificação utilizam a estratégia *separar-e-conquistar* ("*separate-and-conquer*"), porque identificam uma regra que cobre exemplos de uma classe (e exclui outros que não estejam naquela classe), os separa, e o procedimento continua sobre aqueles exemplos que ainda não foram classificados (WITTEN e FRANK, 2000). Existem muitas critérios para a adição de regras a cada estágio que, obviamente, têm um efeito significativo no resultado final (regras produzidas finais).

A derivação de regras ocorre, em geral, após a construção de uma Árvore de Decisão. Essa transformação da Árvore de Decisão em Regras de Classificação, geralmente é feita no 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*"). As regras são escritas considerando o trajeto do nó raiz até uma folha da árvore. Árvores de Decisão e Regras de Classificação são métodos geralmente utilizados em conjunto. Devido ao fato das Árvores de Decisão tenderem a crescer muito, conforme a aplicação, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude das regras poderem ser facilmente modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras (INGARGIOLA, 1996).

Os algoritmos de Regras de Classificação objetivam gerar poucas regras independentes que decidam a designação de exemplos a diferentes classes e, por este motivo, estes algoritmos são muitas vezes chamados de Abordagem de Cobertura ("*Covering Approach*"), pelo fato de a cada estágio ser identificada uma regra que "cobre" alguns dos exemplos.

Um exemplo de derivação de regras está apresentada a seguir:

SE bt < 12.95 **ENTÃO** classe = cálculo
SE bt < 28.05 **ENTÃO** classe = câncer
SE bt < 38.35 **ENTÃO** classe = cálculo

SE bt ≥ 38.35 **ENTÃO** classe = câncer

Os algoritmos que geram Regras de Classificação diferem em pequenos detalhes, sempre na tentativa de obter de forma simples, regras eficazes para solucionar o problema. No "Método Zero R" ("*ZeroR*" ou zero regras), por exemplo, o algoritmo simplesmente contabiliza o número de exemplos pertencentes a cada classe. A classe que contiver o maior número de exemplos passa a conter todos os exemplos. Desta forma, tem-se 100% de acerto para uma das classes e 0% de acerto para a outra classe. Este método não deixa de ser interessante para o caso de problemas em que o número de exemplos pertencentes a determinada classe é bastante reduzido.

Já no "Método UmR" ("*OneR*" ou uma regra), um único atributo faz todo o trabalho de classificação. Este método gera uma árvore de decisão de apenas um nível, que é expressa na forma de um conjunto de regras em que todas testam um atributo particular. Este método mostra que regras simples freqüentemente atingem alta acurácia, funcionando bem em problemas em que apenas um atributo é suficiente para determinar a classe de um exemplo (WITTEN e FRANK, 2000).

O Método Tabela de Decisão ("*Decision Table*") também é bastante simples. Procuramos pelos atributos que possam decidir pelo melhor acerto do resultado; criar uma tabela de decisão envolve selecionar alguns dos atributos. Se algum(s) atributo(s) é irrelevante para a decisão, uma tabela menor (mais condensada), sem aquele(s) atributo(s) será uma "melhor guia" ao resultado. O problema, naturalmente, é decidir qual(is) atributo(s) retirar sem afetar a solução final (WITTEN e FRANK, 2000).

4. Implementação das Técnicas e Análise dos Resultados

Seis algoritmos foram utilizados para a realização dos experimentos, sendo 3 de Árvores de Decisão (*ADTree*, *DecisionStump* e *C4.5*) e 3 de Regras de Classificação (*Decision Table*, *ZeroR* e *OneR*). Para a realização dos experimentos utilizou-se o *software* livre *WEKA* (*Waikato Environment for Knowledge Analysis*, disponível no site www.cs.waikato.ac.nz/ml/weka), no qual tem-se os algoritmos citados, bem como diversos outros.

A metodologia para testes em todos os 6 métodos consistiu em aplicar o método *holdout* estratificado (WITTEN e FRANK, 2000) repetido 10 vezes. Para isso, dividiu-se os conjuntos de pontos (ictéricos com câncer e ictéricos com cálculo) aleatoriamente em 2 subconjuntos: um dos subconjuntos, para Treinamento, serviu para "treinar" o programa, e o outro subconjunto, para Teste, serviu para testar o "modelo treinado". Este procedimento (simulação) foi repetido 10 vezes, variando-se os dois subconjuntos (Treinamento e Teste), sendo que a percentagem de acertos foi calculada e é apresentada nos Quadros 1 e 2 a seguir. No Quadro 1 têm-se os algoritmos para geração de Árvores de Decisão e no Quadro 2, os algoritmos para a geração de Regras de Classificação. Observe-se que em todas as simulações o subconjunto de Treinamento e de Teste usado em cada um dos algoritmos foram os mesmos.

	ADTree		*DecisionStump*		*J4.8*	
	Treinamento	**Teste**	**Treinamento**	**Teste**	**Treinamento**	**Teste**
Simulação 1	97.87	87.50	84.04	70.83	93.62	75.00
Simulação 2	96.84	78.26	83.16	69.56	95.79	69.56
Simulação 3	100.00	83.33	84.04	66.67	93.62	66.67
Simulação 4	100.00	87.50	79.79	83.33	94.68	79.17
Simulação 5	97.87	79.17	80.85	87.50	97.87	87.50
Simulação 6	93.62	70.83	85.11	70.83	95.74	75.00
Simulação 7	100.00	69.56	82.10	69.56	96.84	65.21
Simulação 8	100.00	73.91	83.16	73.91	94.74	82.61
Simulação 9	96.84	78.26	81.05	82.61	92.63	82.61
Simulação 10	96.84	78.26	81.05	82.61	92.63	82.61
MÉDIA	**97.99**	**78.66**	**82.44**	**75.74**	**94.82**	**76.59**
DESVIO PADRÃO	**2.09**	**6.20**	**1.73**	**7.46**	**1.75**	**7.50**

Quadro 1. Percentuais de acerto para os 3 algoritmos de Árvores de Decisão abordados considerando os sub-conjuntos de Treinamento e de Teste

	Decision Table		*ZeroR*		*OneR*	
	Treinamento	**Teste**	**Treinamento**	**Teste**	**Treinamento**	**Teste**
Simulação 1	88.30	83.33	70.21	70.83	84.04	87.50
Simulação 2	89.47	78.26	70.53	69.56	86.31	73.91
Simulação 3	91.49	75.00	70.21	70.83	86.17	66.67
Simulação 4	91.49	75.00	70.21	70.83	81.91	75.00
Simulação 5	90.42	83.33	70.21	70.83	82.92	83.33
Simulação 6	89.36	70.83	70.21	70.83	86.17	70.83
Simulação 7	89.47	69.56	70.53	69.56	84.21	65.22
Simulação 8	91.58	82.61	70.53	69.56	84.21	78.26
Simulação 9	91.58	82.61	70.53	69.56	82.26	78.26
Simulação 10	91.58	82.61	70.53	69.57	85.26	78.26
MÉDIA	**90.47**	**78.31**	**70.37**	**70.20**	**84.35**	**75.72**
DESVIO PADRÃO	**1.23**	**5.38**	**0.17**	**0.67**	**1.63**	**6.96**

Quadro 2. Percentuais de acerto para os 3 algoritmos de Regras de Classificação abordados considerando os sub-conjuntos de Treinamento e de Teste

5. Conclusões e Recomendações

Na área médica, assim como em outras áreas, a posse e uso de ferramentas que auxiliem na tarefa de classificação pode ser crucial, em uma tentativa de otimizar todo o processo, minimizando riscos e custos e, por outro lado, maximizando a eficácia nos resultados.

São diversas técnicas apresentadas na literatura para a resolução de problemas de classificação (Reconhecimento de Padrões) como, por exemplo, Redes Neurais, Modelos de Programação Linear, Métodos Estatísticos e outras, que podem transformar os dados coletados em informações valiosas para auxiliar no processo decisório. Neste presente trabalho optou-se por utilizar técnicas de *Data Mining*, mais especificamente, 3 técnicas envolvendo Árvores de Decisão e 3 técnicas envolvendo Regras de Classificação.

Deve-se destacar que os métodos de *Data Mining* apresentam a vantagem de deixar claro ao usuário quais são os atributos que estão discriminando os padrões e de que forma (pontos de corte) a mesma está ocorrendo (compreensibilidade), característica altamente desejável em qualquer técnica de Reconhecimento de Padrões. A técnica que envolve Redes Neurais, também pode 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.

Com a implementação das técnicas abordadas ao problema médico apresentado neste trabalho, obteve-se os resultados contidos nos Quadros 1 e 2 nos quais nota-se, inicialmente, que todos os métodos para geração de Árvores de Decisão apresentaram percentuais relativamente altos de acerto para o sub-conjunto de Treinamento, com uma média de 91.55% ((97.39 + 82.44 + 94.82)/3), sendo que para os métodos de geração de Regras de Classificação este percentual ficou em 81.73%. Já para o sub-conjunto de Testes, estas médias ficaram em 77% e 74.74%, respectivamente, mostrando que o acerto neste caso, apesar de continuar sendo maior para as Árvores de Decisão, não é tão significativo. A Figura 3 faz uma ilustração comparativa entre estes percentuais.

Além disso, nota-se que a Árvore de Decisão *J4.8* (que é a técnica mais utilizada em trabalhos envolvendo o assunto de *Data Mining*) apresenta um percentual de acerto menor do que outras duas técnicas abordadas neste trabalho (*ADTree* e *Decision Table*).

De qualquer forma, todas as técnicas apresentaram performances satisfatórias, podendo ser utilizadas em casos reais como o problema médico aqui abordado. Desta forma, pode-se oferecer ao especialista da área (ao médico, neste caso), um sistema computacional contendo a técnica de melhor performance como uma ferramenta adicional para uma melhor prescrição de seus diagnósticos.

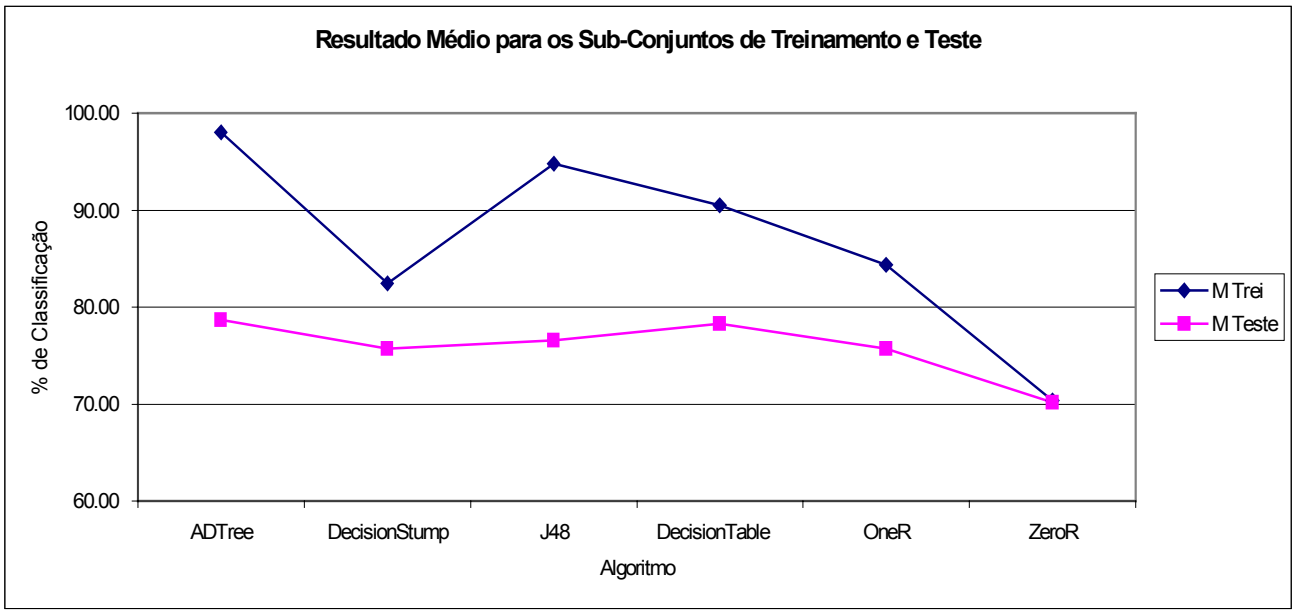

Figura 3. Comparação das Performances das técnicas abordadas (para os sub-conjuntos de Treinamento e de Teste)

Referências:

ANTUNES, C. M. Árvores de Decisão, 2002. Disponível em: <http://mega.ist.utl.pt/~ic.apr/doc/aulas/arvoresdecisão.pdf> Acesso em: 21 jul. 2002.

BERRY, M. J. A. and LINOFF, G. S. *Mastering Data Mining*. John Wiley & Sons, inc., New York, 2000.

BRADZIL, P. B. Construção de Modelos de Decisão a partir de dados, 1999. Disponível em: <http://www.nacc.up.pt/~pbrazdil/Ensino/ML/ModDecis.html> Acesso em: 21 jul. 2002.

CARVALHO, I. C. *Uma Contribuição ao Estudo do Efeito das Insconsistências em Bases de Dados usadas no Treinamento de Sistemas Simbólicos e Conexionistas.*Dissertação de Mestrado, CEFET-PR, Curitiba, PR, 2000.

CARVALHO, I. C. *Métodos de Mineração de Dados (Data Mining) como Suporte à Tomada de Decisão.*Dissertação de Mestrado, ITA, São José dos Campos, SP, 2002.

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.

GARCIA, S. C. *O uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde*. SEMANA ACADÊMICA. Universidade Federal do Rio Grande do Sul, 2000.

INGARGIOLA, G. Building Classification Models: ID3 and C4.5., 1996. Disponível em: http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html> Acesso em 05 jun. 2004.

LEMOS, E. P. *Análise de Crédito Bancário com o uso de Data Mining: Redes Neurais e Árvores de Decisão*. Dissertação de Mestrado, UFPR, Curitiba, PR, 2003.

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.

OSÓRIO, F. Sistemas Adaptativos Inteligentes - Indução de Árvores de Decisão, 2000. Disponível em: <http://www.inf.unisinos.br/~osorio/sadi.html> Acesso em: 12 ago. 2002.

QUINLAN, J. C. *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann, 1993. 302p.

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, Califórnia, 2000.