UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde

por

SIMONE CARBONI GARCIA

Dissertação submetida à avaliação,
como requisito parcial para a obtenção do grau de Mestre
em Ciência da Computação

Prof. Dr. Luis Otavio Campos Alvares
Orientador

Porto Alegre, outubro de 2003.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Garcia, Simone Carboni
O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde / por Simone Carboni Garcia. – Porto Alegre: PPGC da UFRGS, 2003.
87 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós Graduação em Computação, Porto Alegre, BR – RS, 2003. Orientador: Álvares, Luis Otavio Campos.

1. Descoberta de Conhecimento em Bases de Dados. 2. Mineração de Dados. 3. Classificadores. 4. Árvores de Decisão. 5. AIH. I. Álvares, Luis Otavio Campos. II.Título.

Agradecimentos

Agradeço ao meu orientador, professor Dr. Luis Otavio Alvares, pelo apoio, paciência e colaboração os quais me proporcionaram a realização deste trabalho.

Às médicas responsáveis pelo departamento de Controle e Avaliação da Secretaria Municipal da Saúde de Pelotas; também à administradora do hospital Beneficência Portuguesa de Pelotas e responsável pelo setor de Contas Médicas, a senhora Gilca Maria de Paula Fonseca, à médica Márcia Vaz e ao médico neurologista Florisberto Lambrecht, pela cooperação demonstrada e esclarecimentos transmitidos ao longo do desenvolvimento do estudo de caso realizado neste trabalho.

Ao senhor José Manuel R. da Fonseca, pela ajuda prestada.

Ao meu marido Daniel Garcia, pelo incentivo, paciência e, principalmente, carinho dispensados ao longo desta jornada.

E, em especial, a Deus pela oportunidade de estar aqui.

Sumário

Lista de Abreviaturas

AIHs	Autorizações de Internações Hospitalares
SUS	Sistema Único de Saúde
SMSP	Secretaria Municipal de Saúde de Pelotas
AHA	Authorizations of Hospital Admissions
DCBD	Descoberta de Conhecimento em Bases de Dados
KDD	Knowledge Discovery Databases
SGBD	Sistema Gerenciador de Banco de Dados
DM	Data Mining
AVC	Acidente Vascular Cerebral
CART	Classification and Regression Trees
SBC	Sistema de Base de Conhecimento
SADT	Serviços de Diagnose e Terapia
FPT	Fora de Possibilidade Terapêutica

Lista de Símbolos

dom(A)	Conjunto dos valores do atributo A
$\notin$	Não está presente em
$\in$	Está presente em
ε	Referência do conjunto de treino
f_i	Função de classificação
$\rightarrow$	Implicação
Σ	Somatório
$>$	Maior que
$\leq$	Menor ou igual
$=$	Igual
C	Conjunto de classes do problema
p_i	Probabilidade do acontecimento i
$\|S\|$	Cardinalidade do conjunto de exemplos S
$\|S_i\|$	Cardinalidade do conjunto de exemplos classificados na i-ésima partição
S_n	n-ésimo valor do vetor de atributos
$S_{n,m}$	Valor do n-ésimo atributo em relação ao m-ésimo exemplo
$\log_2$	Logaritmo na base 2
(S, A)	Condicionamento do acontecimento A em relação ao acontecimento S
x^2	Teste qui-quadrado
$\max_j f(j)$	Máximo de f(j) em relação a j

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Resumo

As árvores de decisão são um meio eficiente para produzir classificadores a partir de bases de dados, sendo largamente utilizadas devido à sua eficiência em relação ao tempo de processamento e por fornecer um meio intuitivo de analisar os resultados obtidos, apresentando uma forma de representação simbólica simples e normalmente compreensível, o que facilita a análise do problema em questão. Este trabalho tem, por finalidade, apresentar um estudo sobre o processo de descoberta de conhecimento em um banco de dados relacionado à área da saúde, contemplando todas as etapas do processo, com destaque à de mineração de dados, dentro da qual são aplicados classificadores baseados em árvores de decisão. Neste estudo, o conhecimento é obtido mediante a construção de árvores de decisão a partir de dados relacionados a um problema real: o controle e a análise das Autorizações de Internações Hospitalares (AIHs) emitidas pelos hospitais da cidade de Pelotas, conveniados ao Sistema Único de Saúde (SUS). Buscou-se encontrar conhecimentos que auxiliassem a Secretaria Municipal da Saúde de Pelotas (SMSP) na análise das AIHs, realizada manualmente, detectando situações que fogem aos padrões permitidos pelo SUS. Finalmente, os conhecimentos obtidos são avaliados e validados, possibilitando verificar a aplicabilidade das árvores no domínio em questão.

Palavras-chave: Descoberta de conhecimento em bases de dados, mineração de dados, classificadores, árvores de decisão, AIH.

TITLE: "THE USE OF DECISION TREES IN KNOWLEDGE DISCOVERY IN HEALTH RANGE"

Abstract

Decision trees are an efficient way to develop classifiers from a database. They are widely used due to its efficiency in what concerns the processing time and for putting up an intuitive way to analyze the results. They also present a form of simple symbolic representation and normally very comprehensible, thus facilitating the analysis of the problem. This research has the purpose to present a study about the process of knowledge discovery in a database related to the health area, contemplating all the areas of the process, with emphasis on the stage of data mining and on the classifiers based on decision trees. In this study, decision trees were applied to a real problem: the control and analysis of the authorization of hospitalization from hospitals of Pelotas connected with the Sistema Único de Saúde (SUS). From the data, information was sought to help the Secretaria Municipal de Saúde de Pelotas to detect situations that are out of the patterns allowed by the SUS analyzing the authorizations of hospitalization, which is manually done. Finally, the acquired knowledge are evaluated and validated, making possible to verify the applicability of the trees in the domain in question.

Keywords: Knowledge discovery in databases, data mining, classifiers, decision tree, AHA.

1 Introdução

Com o avanço da tecnologia de banco de dados, as organizações ou empresas, graças à maior capacidade de armazenamento e à maior segurança, têm tido a possibilidade de armazenar e recuperar maior quantidade de dados. Com isso, o processamento tornou-se cada vez mais difícil, devido ao grande volume de dados armazenado nas bases de dados e a capacidade de obtenção de informações igualmente tornou-se difícil com os sistemas gerenciadores de banco de dados (SGBD), pois estes conseguem obter somente informações explícitas sobre os dados. Estes dados armazenados podem esconder semelhanças impossíveis de serem identificadas pelos SGBD, as quais formam padrões identificados e extraídos por técnicas utilizadas na área de descoberta de conhecimento em banco de dados (DCBD), também conhecida como *Knowledge Discovery in Databases* (KDD). Estas técnicas possibilitam a obtenção de conhecimentos implícitos nos dados, disponibilizando informações até então desconhecidas pelo usuário.

Diante da imensa quantidade de dados que as grandes bases de dados têm capacidade de armazenar, torna-se impossível a manipulação e descoberta destas informações de forma manual [WRI2000]. A existência de uma forma de aquisição automática do conhecimento nelas contido, apresenta-se como de grande utilidade e tem despertado cada vez mais interesse devido aos bons resultados obtidos na sua utilização, possibilitando às organizações conhecerem melhor o seu funcionamento e, a partir deste conhecimento, tomarem melhores decisões, com o intuito de rever rotinas, traçar estratégias futuras, obter um melhor conhecimento da organização, entre outros.

O processo de DCBD conta com vários métodos de extração de informações: classificação, regressão, associação, clustering, padrões seqüenciais e detecção de desvios.

O método de classificação permite obter padrões através da construção de classificadores. Um dos tipos de classificadores mais aplicados baseia-se na utilização de árvores de decisão como uma forma clara de representar o conhecimento contido no conjunto de dados analisado.

As árvores de decisão apresentam inúmeras aplicações, sendo indicadas para problemas em que os exemplos são representados por pares de *atributo-valor* como, o atributo *sexo* e os valores *masculino* e *feminino* e quando o *atributo meta* tem como saída valores categóricos.

A técnica de árvores de decisão permite que os dados utilizados no treino da árvore contenham erros e também apresentem atributos com valores desconhecidos. Elas são hábeis na geração de regras compreensíveis, executam a tarefa de classificação sem requerer muito tempo de processamento, fornecem um meio intuitivo de analisar os resultados obtidos e manipulam tanto atributos categóricos como quantitativos.

Com base nestas características, as árvores de decisão apresentam inúmeras áreas de aplicação. Dentre as citadas na bibliografia, destacam-se as áreas de:

- medicina, na determinação de diagnósticos e tratamentos, no controle de gastos hospitalares, etc.;

- astronomia, em observações atmosféricas, em análise de informações espaciais, etc.;
- engenharia, no diagnóstico em automóveis, etc.;
- agricultura, na identificação de doenças em produções agrícolas;
- finanças, na detecção de fraudes em seguros; na detecção de uso indevido de cartões de crédito; na aprovação de crédito, etc.;
- marketing, na predição de vendas; na classificação de grupos econômico-sociais por comportamentos, de acordo com particularidades, etc..

A grande aplicabilidade das árvores de decisão e sua crescente utilização, tanto na área acadêmica como em aplicações comerciais, se dá devido à sua flexibilidade, robustez, interpretabilidade e velocidade de processamento.

Os classificadores baseados em árvore de decisão têm, como base, a estratégia *dividir para conquistar*, em que os dados de um problema são divididos em vários subconjuntos, de forma a cada subconjunto ser formado de acordo com características semelhantes dos dados. Desta maneira, os classificadores baseados em árvores de decisão buscam meios de dividir um conjunto de dados em vários subconjuntos, conhecidos como nodos. A classificação, através de uma árvore de decisão, ocorre à medida que são percorridos os caminhos descritos pelos nodos, até ser encontrado um nodo que contém a característica determinante do caminho seguido, recebendo então o nome de folha.

A seleção de algoritmos de classificação para construir árvores de decisão considera vários aspectos. Entre todos a serem ponderados, alguns se destacam pela sua importância apud [FON94]:

- o critério de escolha da característica a se utilizar em cada nodo;
- a forma de calcular o particionamento do conjunto de exemplos a ser utilizado;
- a determinação de um nodo como folha;
- a determinação do critério a utilizar-se na seleção da classe a ser atribuída a cada folha;
- a aplicação de processo de redução de árvores, comumente conhecido como poda.

O objetivo principal deste trabalho é a extração de conhecimento a partir de dados pertencentes a uma situação real, Autorizações de Internações Hospitalares (AIHs), mediante a construção de classificadores baseados em árvores de decisão. Os dados são gerados a partir de serviços prestados pelos hospitais da cidade de Pelotas, conveniados ao Sistema Único de Saúde (SUS). O conhecimento extraído tem como finalidade o auxílio à Secretaria Municipal de Saúde de Pelotas (SMSP) na administração dos serviços prestados pelos hospitais. Aplicações voltadas à administração hospitalar, ainda não são explorada pelos trabalhos relacionados a

construção de árvores de decisão. Na bibliografia consultada não foram encontrados relatos referentes ao tema, fato que tornou este estudo um desafio.

O estudo desenvolvido neste trabalho, contempla todo o processo de descoberta de conhecimento em bases de dados, tendo como parâmetro para o desenvolvimento do processo, a abordagem de [CAB97].

Como base para o desenvolvimento do estudo de caso proposto neste trabalho, no capítulo 2 é apresentado um embasamento teórico sobre o processo de descoberta de conhecimento em bases de dados, explanando as etapas que o compõem, os principais métodos que podem ser utilizados na etapa de mineração de dados, também conhecida por *Data Mining* (DM). Dentro dos métodos de mineração de dados é destacado o de classificação.

O capítulo 3 aborda o uso de classificadores baseados em árvores de decisão. Com o foco nas árvores de decisão são descritos os tipos de testes utilizados por esta técnica, as técnicas para sua construção, os critérios de seleção de atributos, como é realizada a determinação da classe que deve ser associada a uma folha da árvore e as técnicas de poda.

No capítulo 4 são expostos algoritmos que implementam árvores de decisão e as características e recursos da ferramenta Sipina-W, utilizada no estudo de caso realizado. Com esta ferramenta é possível trabalhar com diversos tipos de algoritmos de construção de árvores de decisão, gerar regras a partir da árvore construída e, também, verificar a confiabilidade da árvore gerada.

O capítulo 5 relata o estudo de caso realizado com base nos dados provenientes das Autorizações de Internações Hospitalares (AIHs), utilizados pelas Secretarias Municipais de Saúde, no gerenciamento dos serviços prestados pelos hospitais conveniados ao Sistema Único de Saúde (SUS). No estudo de caso são descritos os resultados obtidos com as análises executadas por meio de árvores de decisão, a partir dos dados das AIHs referentes à patologia Acidente Vascular Cerebral (AVC), sendo a construção das árvores e a sua validação efetuadas pela ferramenta Sipina-W.

As conclusões do trabalho realizado são apresentadas no último capítulo.

2 Descoberta de Conhecimento em Bases de Dados

Neste capítulo é apresentada uma revisão bibliográfica sobre o processo de descoberta de conhecimento em bases de dados (DCBD), destacando a etapa mineração de dados, bem como, o método de classificação. Após algumas considerações iniciais, a seção 2.2 introduz uma conceituação referente ao processo de descoberta de conhecimento e descreve as etapas que o compõe. A seção 2.3 explana com mais detalhes a etapa de mineração de dados e descreve os métodos que podem ser utilizados na descoberta de padrões. O método de classificação é detalhado na seção 2.4.

2.1 Considerações Iniciais

O crescente interesse pelo campo de descoberta de conhecimento em bases de dados é impulsionado pela redução do custo dos dispositivos de armazenagem de dados, pela grande capacidade dos sistemas de computação em gerar e armazenar dados, e pela necessidade de transformar estes dados, até então desconhecidos pelas organizações e empresas, em conhecimento.

O termo "descoberta de conhecimento em bases de dados" foi formalizado nos anos 80, tendo, como finalidade, encontrar padrões, similaridades e conhecimentos a partir de dados, facilitando a análise das informações a serem utilizadas por pessoas responsáveis pela tomada de decisões [CAB97].

A descoberta do conhecimento em bases de dados surge como uma nova área de pesquisa para a busca de padrões que podem ser definidos como uma afirmação sobre uma distribuição probabilística e sobre informações anteriormente desconhecidas, existentes em grandes bases de dados [JOH97][UEA99][FAY2000]. A descoberta de conhecimento combina técnicas de inteligência artificial, reconhecimento de padrões e estatística.

O campo da descoberta de conhecimento em bases de dados tem produzido muitas aplicações práticas em áreas como detecção de fraudes, análises de diagnósticos médicos, predição do comportamento de clientes, predição de interesses de usuários da Web, otimização de processos de manufatura, etc. Ela também tem conduzido para um conjunto de fascinantes questões científicas sobre como computadores podem aprender automaticamente a partir de experiências [MIT2000].

2.2 O Processo de Descoberta de Conhecimento em Bases de Dados

O processo de DCBD é definido por Fayyad e outros [FAY96], como um processo não-trivial, proveniente de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados.

Nesta definição o processo é dito não-trivial, pois envolve um certo grau de complexidade, devido às condições dos dados encontrados nas bases de dados, os quais

podem ser de grande volume, possuírem ruídos, serem esparsos, terem informações incompletas e redundantes. Os padrões descobertos durante o processo devem ser válidos para novos dados com algum grau de confiabilidade e também úteis e compreensíveis às pessoas envolvidas no processo.

Autores como [ADR96], [BRA96], [FAY96], [CAB97] têm, de uma maneira geral, a mesma visão global do processo de DCBD, assim como a maioria das propostas existentes, havendo apenas pequenas diferenças entre elas.

Nas propostas dos autores citados, o processo envolve várias etapas necessárias à obtenção de um conhecimento válido. Elas compreendem a definição dos objetivos a serem alcançados; a coleta e preparação dos dados indispensáveis ao alcance dos objetivos; a realização de uma análise dos dados coletados com o auxílio de técnicas de extração de padrões; a avaliação dos resultados e a assimilação do conhecimento obtido.

Uma das propostas mais citadas na literatura é a encontrada em [FAY96]. Para o autor, o processo de descoberta de conhecimento é iterativo, envolvendo várias etapas executadas seqüencialmente, sendo, muitas vezes, necessário o retorno a etapas anteriores para poder-se fazer ajustes, obtendo o resultado esperado. O processo é também interativo, sendo necessária a tomada de decisão por parte do analista durante a execução das etapas.

Nas propostas apresentadas por [ADR96] e [CAB97], o processo é iterativo, podendo retornar a uma ou mais etapas para se fazerem ajustes necessários ao bom desenvolvimento do processo. As propostas de [ADR96], [FAY96] e [CAB97] de uma maneira geral são muito similares, com diferenças em alguns detalhes.

Na proposta apresentada por [BRA96], é dada ênfase ao envolvimento das pessoas que interagem no processo de descoberta de conhecimento em bases de dados. O autor trata o processo como uma complexa interação entre o usuário (analista) e uma grande base de dados, durante a qual o próprio processo dá suporte a quem o está conduzindo, através de um conjunto heterogêneo de ferramentas (ferramentas de consulta, Estatísticas e Inteligência Artificial, visualização, apresentação e transformação). O processo é apresentado por etapas que o analista pode repetir tantas vezes quantas for necessário, podendo, do mesmo modo, retornar a etapas anteriores. O autor também enfatiza os resultados e aplicações de um processo de descoberta de conhecimento em bases de dados.

2.2.1 Etapas do processo de descoberta de conhecimento em bases de dados

Este trabalho apresenta mais detalhadamente a abordagem descrita por [CAB97]. Para o autor, o processo é composto por cinco etapas que seguem o fluxo apresentado na figura 2.1.

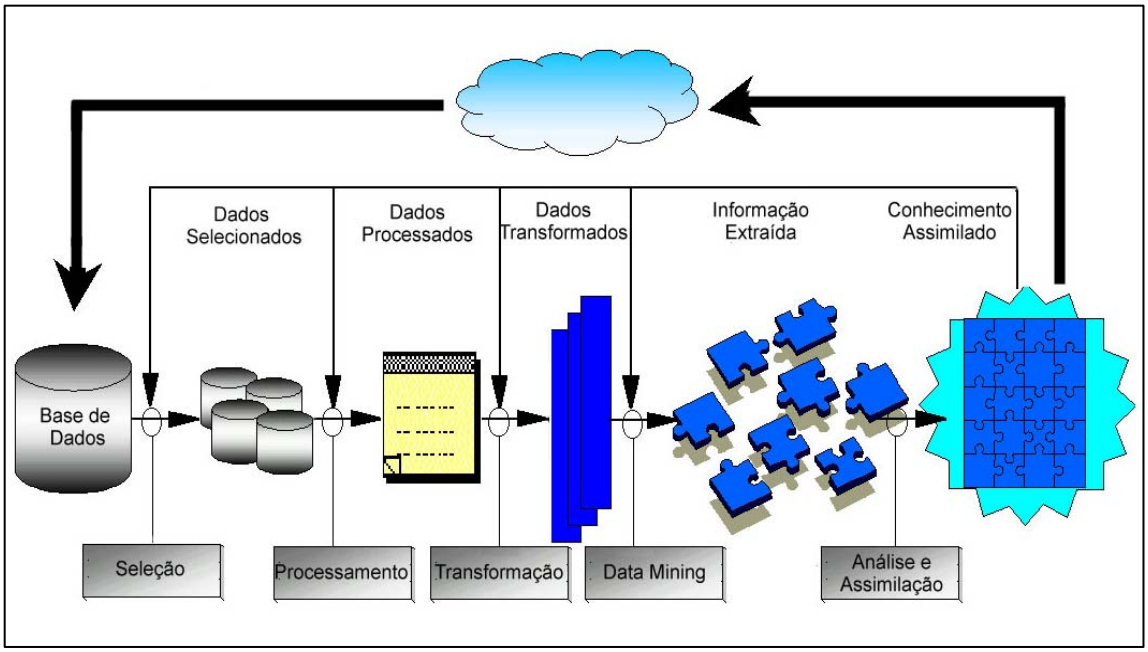

FIGURA 2.1 – O processo de descoberta de conhecimento em bases de dados [CAB97]

Nesta abordagem, os objetivos determinados para a aplicação dirigem todo o processo. Eles são a base na qual o projeto inicial é estabelecido e o meio de avaliar os resultados finais, devendo, também, guiar constantemente a equipe envolvida durante todas as etapas do processo.

Embora as etapas sejam executadas na ordem descrita na figura 2.1, o processo é iterativo, possivelmente com muitos laços de retorno a uma ou mais etapas. Além disso, o processo não é autônomo, sendo necessária a intervenção da equipe na tomada de decisão durante a execução das etapas.

2.2.1.1 Determinação de objetivos

Nesta etapa são determinados os objetivos a serem alcançados no decorrer do processo. Eles devem ser claros e bem definidos, para o processo ser favorável e o resultado ter algo a acrescentar à organização ou empresa. Pela indicação dos objetivos, é estabelecido o problema em questão, sendo esta etapa essencial em qualquer projeto de descoberta de conhecimento.

Nela, são também avaliadas, com base no conhecimento do domínio e dos dados, a viabilidade e a possibilidade de se construir uma aplicação através do processo de DCBD.

2.2.1.2 Preparação de dados

A preparação dos dados é a etapa que comumente consome mais recursos, ocupando, segundo [CAB97], em média 60% do projeto como um todo. É composta de três fases:

- *seleção dos dados* – o objetivo da seleção dos dados é identificar a origem dos dados disponíveis e extrair os dados necessários à etapa de mineração de dados, tendo, como base, os objetivos estabelecidos anteriormente. Feita a seleção dos dados, é necessário compreender o significados dos atributos, seus possíveis valores, o formato dos dados, entre outras informações. Nos dados é possível encontrar dois tipos principais de atributos, nomeados em [CAB97] como:
  - *categóricos* – os valores são finitos e diferem no tipo, dividindo-se em nominais e ordinais. Os nominais nomeiam o tipo do objeto a que se referem, não havendo ordem entre os possíveis valores; por exemplo, o estado civil (solteiro, casado, divorciado). Os atributos ordinais têm uma ordem entre os possíveis valores; por exemplo, a avaliação de crédito de um cliente (bom, regular, ruim).
  - *quantitativos* – nos atributos quantitativos existe diferença entre os possíveis valores, podendo ser contínuos e discretos. Os valores dos atributos contínuos são números reais e os valores dos atributos discretos são inteiros.

- *pré-processamento dos dados* – o objetivo do pré-processamento é assegurar a qualidade dos dados selecionados. A limpeza dos dados e sua compreensão são pré-requisitos para a etapa de mineração de dados ser bem sucedida. A fase de pré-processamento dos dados tem início com uma revisão geral da estrutura dos dados e a verificação da qualidade deles. Com poucas exceções, os dados selecionados apresentam inconsistências, sendo as mais comuns:
  - *dados com ruído* – envolvem atributos com valores que não estão de acordo com o esperado deles. A ocorrência de ruídos pode resultar de erro humano; por exemplo, a idade de uma pessoa é gravada como 650 ou uma renda é negativa. Neste caso, o valor pode ser substituído por um valor válido ou o registro com este valor pode ser retirado da análise. A medida, em caso de se verificar a presença de dados com ruído, depende da natureza deste ruído; verificada sua natureza, deve ser aplicada aquela mais adequada ao caso.
  - *valores desconhecidos* – os valores desconhecidos compreendem os valores inexistentes nos dados selecionados que podem ter sido suprimidos durante a detecção do ruído. Este tipo de situação pode ocorrer por um erro humano, por a informação não estar disponível durante sua entrada no sistema ou porque os dados foram selecionados através de fontes heterogêneas, assim criando más combinações. Para tratar os valores desconhecidos, utilizam-se diversas técnicas, não tendo uma que seja ideal. Uma delas é a de

eliminar os valores desconhecidos, sendo de fácil aplicação, mas na qual podem ser perdidos dados de grande importância. Outra técnica se dá pela substituição do valor desconhecido por um valor mais provável.

- *transformação dos dados* - nesta fase, os dados processados são modelados para produzir um modelo analítico, isto é, um modelo informativo dos dados, que representa a consolidação, integração e reestruturação dos dados selecionados e pré-processados. Logo após a construção do modelo, os dados são transformados de acordo com as exigências do formato de entrada do algoritmo de mineração de dados a ser utilizado. As técnicas usadas na transformação podem variar das conversões no formato dos dados à redução dos dados, ou seja, reduzir o número total de atributos através da combinação de diversos atributos existentes em um novo atributo.

2.2.1.3 Mineração de dados

A mineração de dados é considerada o núcleo do processo de DCBD. Nesta etapa, é aplicado o algoritmo ou a combinação apropriada de algoritmos de mineração aos dados tratados na etapa anterior [ARU99]. Os algoritmos são selecionados a partir dos objetivos determinados no início do processo.

2.2.1.4 Análise dos resultados

Nesta etapa, são realizadas a interpretação e a avaliação do resultado obtido na etapa anterior. A análise dos resultados e a mineração de dados estão descritas de forma separada, como mostra a figura 2.1, mas são inseparáveis, estando ligadas por um processo interativo. Devido à natureza exploratória do processo de descoberta de conhecimento, a análise dos resultados obtidos é feita sempre na busca de algo interessante e válido, sempre retornando à etapa de mineração se estes objetivos não foram atingidos.

2.2.1.5 Assimilação do conhecimento

A assimilação do conhecimento encerra o ciclo, que foi iniciado quando o conjunto de objetivos estabelecidos deu início ao processo. A finalidade esta etapa é encerrar o processo de descoberta de conhecimento, incorporando a demanda de informações ganhas nas diversas etapas. Nesta etapa, são apresentadas as novas descobertas e são formuladas as maneiras de como as novas informações podem ser melhor exploradas.

Neste trabalho será dado destaque à etapa de mineração de dados, apresentada como outro subtítulo novamente, por ser o cerne do estudo.

2.3 Mineração de Dados

A mineração de dados é uma etapa que objetiva encontrar padrões escondidos nos dados, envolvendo freqüentemente uma repetida aplicação iterativa de métodos de mineração de dados [SOU98]. Estes são implementações específicas de algoritmos utilizados na procura de padrões em conjuntos de dados.

O uso de um tipo de método não está associado necessariamente a um tipo de aplicação específica e vice-versa [CAB97]. No entanto, existem algumas associações bem estabelecidas entre aplicações e métodos como, por exemplo, estratégias de marketing, executadas quase sempre por meio do método de clustering. Entretanto, a detecção de fraudes pode ser implementada por vários tipos de métodos, dependendo da natureza do problema.

Os métodos de mineração de dados não são necessariamente utilizados de forma exclusiva, podendo ser aplicado primeiramente um método à base de dados, para melhor prepará-los para a aplicação e, então, aplicar outro no conjunto de dados resultante.

A seleção do método ou métodos baseia-se nos objetivos determinados para a aplicação e, geralmente, é realizada conjuntamente entre o analista de dados e as pessoas da organização ou empresa envolvidas no processo.

A literatura descreve muitos métodos de mineração de dados. Este trabalho apresenta, os mais comumente encontrados: classificação, regressão, associação, clustering, padrões seqüenciais e detecção de desvios.

2.3.1 Classificação

Classificação é um método de mineração de dados cujo objetivo é classificar elementos de um conjunto de dados em diferentes classes, baseado em propriedades que estes elementos têm em comum (atributos) [SOU98], para poder obter-se um modelo de classificação e, a partir deste modelo, predizer classes de novos elementos de um conjunto de dados.

O método de classificação pode ser utilizado em aplicações que incluem diagnósticos médicos, avaliação de risco em empréstimos, detecção de fraudes, etc.

A seção 2.4 trata com mais detalhes o método de classificação, que é o método utilizado no estudo de caso descrito no capítulo 5.

2.3.2 Regressão

Regressão refere-se à descoberta de padrões preditivos em que o atributo meta possui valor real [JOH97]. Este é um método matemático utilizado para ajuste de curvas, pois dado um conjunto de pontos, ele calcula fórmulas capazes de fornecer pontos intermediários, anteriores e posteriores. Neste caso podem ser determinadas, por exemplo, similaridades em séries de tempo.

O uso do método de regressão se adapta a muitas aplicações [FAY96], entre elas, predizer a quantidade de biomassa presente em uma floresta; estimar a probabilidade de um paciente morrer com base em resultados de um conjunto de diagnósticos ou predizer a demanda de consumo para um novo produto em função das despesas com publicidade.

2.3.3 Associação

O método de associação busca estabelecer relacionamentos entre um conjunto de dados, a fim de encontrar afinidades entre eles. A partir de uma transação, as regras de associação tentam encontrar itens que envolvem a presença de outros itens.

Tendo, como exemplo, uma base de dados de compras, na qual cada compra (transação) consiste de vários itens de artigos adquiridos por um cliente, pôde-se extrair uma regra de associação hipotética apresentada na figura 2.2, mostrando que, em 70% das compras de uma camisa, ocorre também a compra de uma gravata.

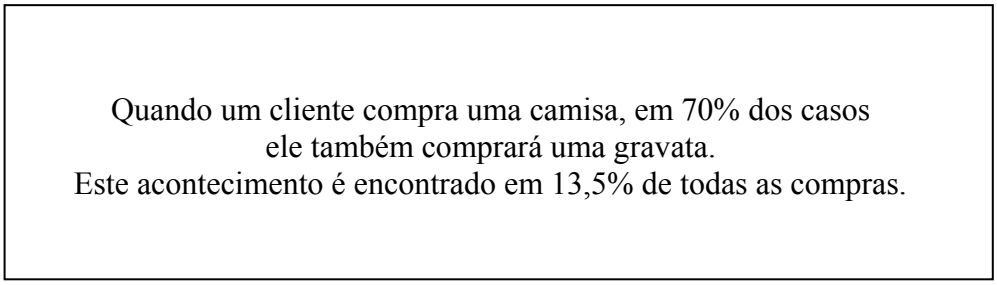

FIGURA 2.2 – Exemplo de regra de associação [CAB97]

O método de associação produz regras com o formato “Se X, então Y”, em que X e Y são conjuntos de itens. Estas regras podem ser avaliadas através dos parâmetros *fator de suporte* e *fator de confiança* [AGR93][CAB97].

O fator de suporte indica a ocorrência relativa das regras de associação dentro do conjunto de dados da transação. Este fator é determinado pela divisão do número das transações que suportam a regra pelo número total das transações. Uma transação suporta a regra “quando X então Y” se os itens X e Y na regra ocorrerem também na transação. Na figura 2.2, o fator de suporte da regra é tido como 13,5% dos registros da base de dados.

O fator de confiança de uma regra de associação é o grau no qual a regra é verdadeira ao considerar os registros de forma individual. É calculado dividindo o número das transações que suportam a regra pelo número de transações que suportam somente a primeira parte da regra (X). No exemplo da figura 2.2, o fator da confiança é 70%.

Após a aplicação de regras de associação a um conjunto de dados, o resultado obtido é uma lista de padrões que indicam afinidades entre itens.

2.3.4 Clustering

Clustering, também conhecido como *segmentação de base de dados* ou *agrupamento*, tem, como objetivo, particionar uma base de dados dentro de grupos de registros similares, isto é, registros que têm em comum um número de propriedades e, deste modo, são considerados homogêneos [CAB97].

O método de clustering tem seu uso difundido pelas empresas que necessitam aprender mais sobre quem são os clientes, para poderem melhorar seu marketing ou desenvolver novos produtos [JOH97].

A figura 2.3 mostra um exemplo hipotético de um modelo de clustering, obtido por meio de uma base de dados de compradores de carros esporte. Para cada comprador são conhecidos o sexo, a idade e a renda. No exemplo, a população foi agrupada em conjuntos (indicados por círculos), representando subpopulações significativas dentro da base de dados. Pode-se verificar que existem três conjuntos de compradores: compradores jovens, compradores na grande maioria do sexo masculino com renda média-alta englobando várias faixas etárias e compradores do sexo masculino com faixa etária entre os 40 e 50 anos.

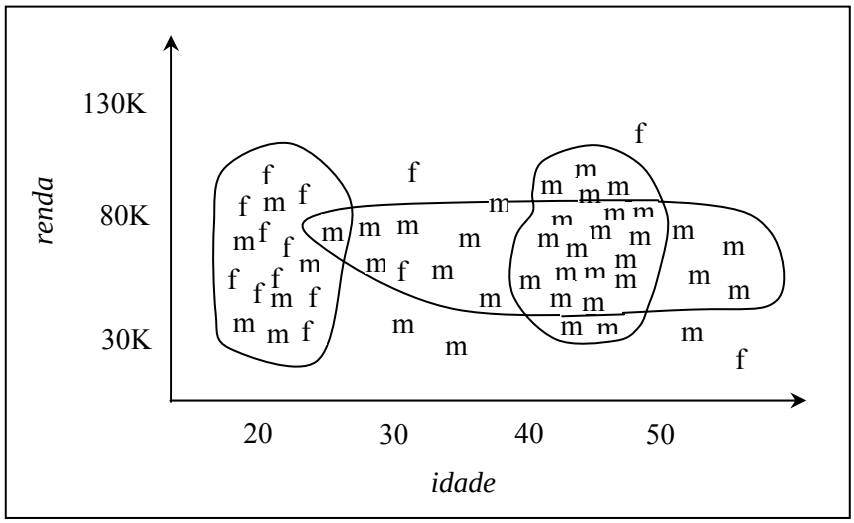

FIGURA 2.3 – Exemplo do método de clustering [JOH97]

2.3.5 Padrões seqüenciais

O método de padrões seqüenciais detecta tendências entre transações, sendo estas transações associadas a um período de tempo. Por exemplo, dada uma base de dados de compras e clientes, cada transação consiste em um conjunto de itens, no qual cada transação é identificada por seus dados, uma data e hora. Cada item é identificado por um único item identificador. Cada cliente é identificado por um único identificador de cliente.

A tabela 2.1 mostra parte de uma base de dados de uma loja de bebidas, com detalhes de algumas transações efetuadas. Os dados são classificados por cliente e transação. Por exemplo, o cliente *B. Moore* visitou a loja em três dias consecutivos. No

primeiro o cliente comprou cerveja; no dia seguinte ele comprou vinho e cidra e no terceiro, conhaque.

As seqüências do cliente são organizadas pela transação, como mostra a tabela 2.2. Cada conjunto de parênteses nas seqüências indica uma transação que pode possuir um ou mais artigos. Por exemplo, B. Moore comprou dois artigos: vinho e cidra, na segunda transação.

A técnica faz uma contagem da freqüência para cada combinação da transação, que pode ser produzida a partir das seqüências do cliente e indica aqueles padrões seqüenciais cuja ocorrência relativa é maior do que o nível de suporte mínimo requerido; isto é mostrado na tabela 2.3.

Na tabela 2.3, o padrão seqüencial é estimado com 40% de suporte, "cerveja é comprada em uma transação antes que o conhaque seja comprado em uma transação subseqüente", se isto ocorrer para dois dos cinco clientes.

TABELA 2.1 – Padrão Seqüencial: Base de Dados de Transações [CAB97]

Cliente	Tempo da transação	Itens comprados
B. Adams	21 junho, 1994 17:27	Cerveja
B. Adams	22 junho, 1994 10:34	Conhaque
J. Brown	20 junho, 1994 10:13	Suco, Refrigerante
J. Brown	20 junho, 1994 11:47	Cerveja
J. Brown	21 junho, 1994 9:22	Vinho, Água, Cidra
J. Mitchell	21 junho, 1994 15:19	Cerveja, Gim, Cidra
B. Moore	20 junho, 1994 14:32	Cerveja
B. Moore	21 junho, 1994 18:17	Vinho, Cidra
B. Moore	22 junho, 1994 17:03	Conhaque
F. Zappa	20 junho, 1994 11:02	Conhaque

TABELA 2.2 – Padrão Seqüencial: Seqüências dos Clientes [CAB97]

Cliente	Seqüências do cliente
B. Adams	(Cerveja)(Conhaque)
J. Brown	(Suco, Refrigerante)(Cerveja)(Vinho, Água, Cidra)
J. Mitchell	(Cerveja, Gim , Cidra)
B. Moore	(Cerveja)(Vinho, Cidra)(Conhaque)
F. Zappa	(Conhaque)

TABELA 2.3 – Padrão Seqüencial: Suporte ≥ 40% [CAB97]

Padrões seqüenciais com suporte ≥ 40%	Clientes Suportados
(Cerveja)(Conhaque)	B. Adams, B. Moore
(Cerveja)(Vinho, Cidra)	J. Brown, B. Moore

2.3.6 Detecção de desvios

Detecção de desvios é um método de descoberta de objetos que não seguem padrões de valores. Para isso, é necessária a criação de padrões de forma prévia, comumente chamados de normas [MAT96].

Um desvio, de uma maneira geral, é um contraste entre uma observação realizada e um referencial adotado. Os desvios podem ser instâncias não enquadradas nas classes; superposições entre classes; mudanças no valor em um período de tempo; discrepância entre valores observados e valores esperados previstos pelo modelo.

2.4 O Método de Classificação

O método de classificação baseia-se na modelagem preditiva, a qual se assemelha à experiência humana, que usa observações para dar forma a um modelo, baseando-se nas características essenciais subjacentes de um fenômeno. Sendo assim, para o entendimento e a comunicação com o mundo, o ser humano está constantemente classificando, categorizando e graduando vários elementos ao seu redor. Por exemplo, a partir de características que nos fazem determinar o que é um cachorro, podemos identificar outros cachorros pertencentes às mais variadas raças; a partir de características dos povos, podemos determinar suas raças [BER97][CAB97].

Neste método, utiliza-se um modelo preditivo para analisar uma base de dados existente e determinar algumas características essenciais dos dados. O modelo deve refletir uma resposta correta, referente a alguns exemplos já conhecidos, para que, a partir deste modelo, seja possível realizar pareceres sobre novos exemplos.

Os exemplos a serem classificados são representados por registros (casos) de uma base de dados; a tarefa de classificação (de um classificador) consiste em relatar

casos ocorridos anteriormente e organizá-los dentro de categorias (classes). Estas classes são baseadas em propriedades comuns (conteúdo dos atributos) do conjunto de casos da base de dados em análise. Os casos organizados dentro de classes, constituem um modelo que é utilizado para classificar novos casos ainda não classificados.

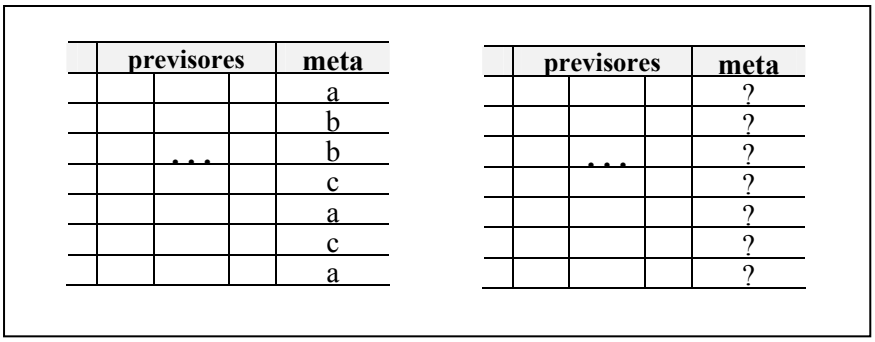

FIGURA 2.4 – Conjunto de treino e conjunto de teste [FRE97]

A figura 2.4 ilustra dois conjuntos de dados, em que cada coluna corresponde a um atributo e cada linha a um registro. Cada registro é composto por um grupo de atributos previsores e por um atributo meta, determinando a classe a que pertence o registro. Na figura, o conjunto da esquerda é utilizado na construção do classificador, sendo conhecido como conjunto de treino. Nele, os valores (*a*, *b*, *c*)do atributo meta são conhecidos. A tarefa do classificador é prever com base nos atributos previsores e no atributo meta do conjunto de treino, o valor do atributo meta do conjunto da direita, ou seja, a classe a que pertence cada um dos registros do conjunto de teste.

Aplicações adequadas ao método de classificação incluem [QUI86][BER97] [SOU98]: a detecção de fraudes, quando fraudes são descobertas a partir de padrões similares que já tenham ocorrido no passado; alvos em marketing, com os quais, a partir de características de uma população, é possível traçar campanhas de marketing para determinados tipos de produtos; a aprovação de crédito a clientes, o crédito podendo ser classificado como de risco baixo, médio ou elevado, de acordo com concessões realizadas anteriormente; a determinação de tratamentos apropriados, cujo diagnóstico e tratamento aplicado a pacientes podem ser classificados a partir dos sintomas de doenças e terapias possíveis; observações atmosféricas, pelas quais é possível prever se um temporal severo é improvável, possível ou provável.

Outro exemplo de aplicação do método de classificação é citado por Agrawal, Imielinski e Swami [AGR93]. Visa a solucionar o problema da localização de uma loja. Supõe-se que o sucesso da loja é determinado pelas características de sua vizinhança, e a empresa está interessada em identificar as vizinhanças que devem ser as principais candidatas a uma investigação, para ser escolhida a localização de uma nova loja. A empresa tem acesso a uma base de dados das vizinhanças e categoriza primeiramente suas lojas como *bem sucedidas*, *médias* e *mal sucedidas*. Baseado nos dados das vizinhanças destas lojas, é determinado um perfil para cada uma delas, desenvolvendo-se, assim, uma função de classificação para cada categoria de loja. A partir da função de classificação de lojas bem sucedidas, são recuperadas as vizinhanças em que a nova loja obterá sucesso.

Estas aplicações, entre outras, têm utilizado como meio de solução para os problemas apresentados, o método de classificação, obtendo bons resultados.

2.4.1 Descrição formal

A descrição formal da tarefa de classificação para um conjunto de exemplos, conforme [AGR2000], é apresentada como: seja G um conjunto de m grupos denominados $\{G_1, G_2, ..., G_m\}$. Seja A um conjunto de n atributos (características) $\{A_1, A_2, ...A_n\}$. Seja $dom(A_i)$ referente ao conjunto de possíveis valores para o atributo A_i. É dada uma grande base de dados de objetos D na qual cada objeto é um n-registro de forma $<v_1, v_2, ...v_n>$ onde $v_i \in dom(A_i)$ e $G \notin A_i$, em outras palavras, o grupo chamado de objetos em D não é conhecido. É dado um conjunto de objetos ε no qual cada objeto é um $(n + 1)$-registro na forma $<v_1, v_2, ..., v_n, g_k>$ onde v_i dom (A_i) e $g_k \in G$, ou seja, os objetos em ε têm os mesmos atributos que os objetos em D e, adicionalmente, têm grupos associados a eles. O problema consiste em obter m funções classificadoras, uma para cada grupo G_j, usando a informação em ε, com a função classificação f_j para o grupo G_j presente em f_j: A_1 x A_2 x ... $A_n \rightarrow G_j$ para $j = 1, ..., m$. Os exemplos do conjunto ε são referidos como o conjunto de treino e a base de dados D como o conjunto de dados de teste.

2.4.2 Avaliação da qualidade de um classificador

A avaliação da qualidade de um classificador é de suma importância, para que se possa estimar quanto este classificador é preciso na classificação de futuros exemplos.

A precisão de um classificador é medida pelo número de erros ocorridos durante a classificação realizada em um conjunto de exemplos. A equação 2.1 define o erro de classificação, ou seja, o percentual de exemplos que foram classificados de forma incorreta.

$$Taxa\ de\ erros = \frac{Número\ de\ erros}{Número\ de\ casos\ testados} \qquad (2.1)$$

A taxa de erros é determinada pelo percentual de exemplos do conjunto de teste mal classificados. Um fator que pode influenciar a estimativa da taxa de erro é o número de exemplos utilizados. Se o número de exemplos for grande, a taxa de erro num conjunto de teste apresentará uma boa estimativa de generalização do classificador, sendo o conjunto de teste independente do conjunto de treino e ambos os conjuntos suficientemente grandes. Se o número de exemplos é limitado, situação considerada mais comum, é necessário estimar a taxa de erro, tendo o cuidado de não torná-la por demais otimista ou pessimista.

Uma outra situação colocada na avaliação da qualidade de um classificador são os custos do erro. Verifica-se que existem variados tipos de custos associados aos possíveis erros encontrados [FON94][BEM97][UTG97]. Neste trabalho, o custo é apenas associado à atribuição de classes à cada nodo folha. Como exemplo, pode-se

citar o caso de um sistema de diagnóstico médico no qual é geralmente considerado mais grave classificar um doente como *saudável*, do que classificar um saudável como *doente*.

A seguir são descritas técnicas que podem ser utilizadas na avaliação da qualidade de um classificador.

2.4.2.1 Avaliação por estimativa de custos

A estimativa de custos de um erro em um conjunto de exemplos considera como ponto de avaliação a minimização os custos do problema em questão. Nesta metodologia, é necessária a determinação de pesos que deverão ser atribuídos aos exemplos mal classificados. Para isto, é necessário o uso de uma matriz conhecida como matriz de custos, contendo o custo de cada tipo de erro possível. Esta matriz terá n^2 elementos, em que n representa o número de classes. A tabela 2.4 ilustra um exemplo de uma matriz de custos possível para um problema com três classes.

TABELA 2.4 – Exemplo de uma matriz de custos de erros [FON94]

Classe atribuída pelo classificador	**Classe real**		
	1	*2*	*3*
1	0	2	12
2	10	0	8
3	5	6	0

Para avaliar a qualidade do classificador baseado em custos, é necessária a utilização da matriz de confusão referente ao problema. A matriz de confusão descreve os acertos e os erros ocorridos durante a classificação. A tabela 2.5 apresenta os resultados do teste do classificador em termos do número de exemplos da classe *i* aos quais foi atribuída a classe *j* com *i* e *j* variando entre *1* e *n*.

TABELA 2.5 – Exemplo de uma matriz de confusão [FON94]

Classe atribuída pelo classificador	**Classe real**		
	1	*2*	*3*
1	92	6	2
2	5	89	6
3	10	15	75

O custo do classificador é determinado por:

$$Custo = \sum_{i=1}^{n}\sum_{j=1}^{n} - C_{i,j} M_{i,j} \qquad (2.2)$$

Onde: $C_{i,j}$ é o valor da linha i, coluna j da matriz de custos

$M_{i,j}$ é o valor da linha i, coluna j da matriz de confusão

Baseado nas tabelas 2.4 e 2.5, o valor do custo para o exemplo é:

$$Custo = 0*92 + 2*6 + 12*2 + 10*5 + 0*89 + 8*6 + 5*10 + 6*15 + 0*75 = 274$$

Neste exemplo, o custo médio por decisão é:

$$Custo\_médio = \frac{Custo}{Número\_de\_classificações} = \frac{274}{301} = 0{,}91$$

2.4.2.2 Avaliação por estimativa de erros

2.4.2.2.1 Ressubstituição

O método de estimação por ressubstituição é utilizado para avaliar o modelo de classificação, usando, para isto, um dado conjunto de exemplos, já utilizados para o treino.

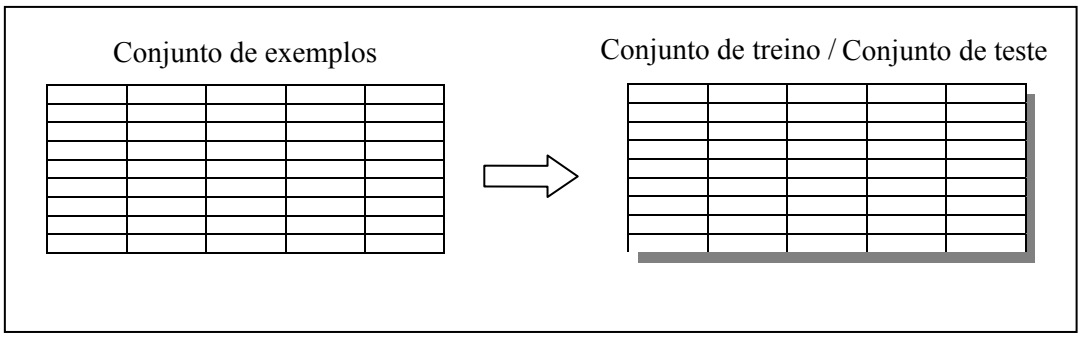

FIGURA 2.5 – Relação entre o conjunto de treino e de teste (Ressubstituição)

A proporção de resultados incorretos obtidos durante a estimação por ressubstituição determina a taxa de erro de classificação.

2.4.2.2.2 Treino e Teste

Na estimação por treino e teste, o conjunto de exemplos é particionado em dois subconjuntos, o de treino e o de teste. A maior parte do conjunto de exemplos é aproveitada para o treino. Geralmente 2/3 dos casos são utilizados no conjunto de treino e 1/3 no conjunto de teste [FON94][MOU93][WEI98]. A figura 2.6 ilustra a partição do conjunto de exemplos em treino e em teste.

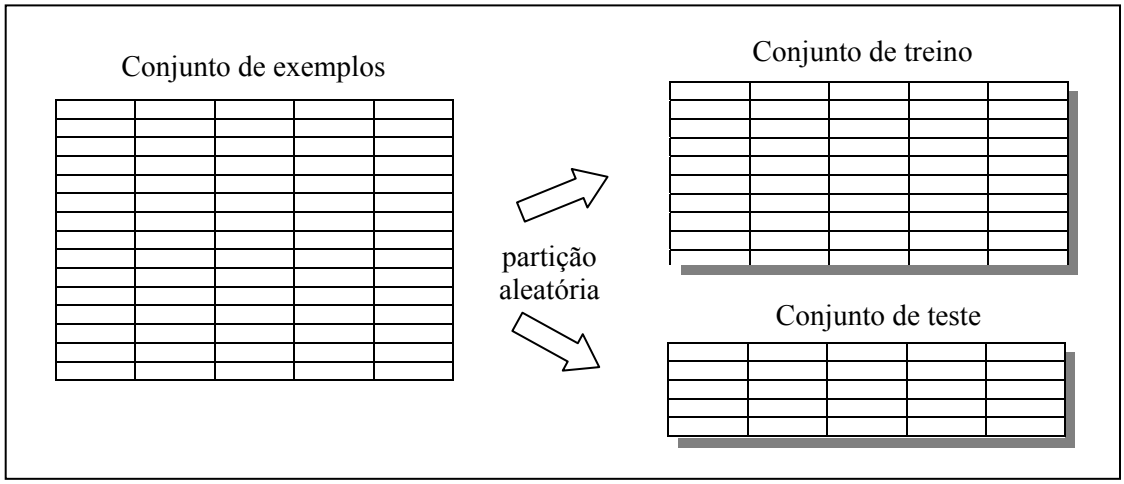

FIGURA 2.6 – Relação entre o conjunto de treino e de teste (Treino e Teste)[WEI98]

O método de classificação não tem nenhum acesso aos dados de teste. Uma vez encontrada a solução baseada no conjunto de treino, o modelo de classificação construído é avaliado, sendo medido seu desempenho através do conjunto de teste. O desempenho do conjunto de teste é uma estimativa do desempenho futuro do modelo de classificação em relação a novos casos.

2.4.2.2.3 Validação Cruzada

Na estimação por validação cruzada, os exemplos são particionados dentro de *n* subconjuntos, de forma que o número de exemplos e a distribuição das classes sejam o mais uniforme possível.

O conjunto de treino é constituído utilizando *n-1* partições, e a partição restante constitui o conjunto de teste, sendo este utilizado para estimar o erro de classificação.

Com base nas partições realizadas, em que são determinados subconjuntos de dados, ocorrem *n* permutações diferentes, sendo construídos classificadores parciais, normalmente 10. Este número é considerado bom para se obter uma boa precisão do modelo de classificação gerado.

A estimação do erro é, portanto, determinada pela utilização de um conjunto de teste independente para cada um dos classificadores parciais. O valor estimado para o

erro do classificador final é a média dos erros estimados para cada um dos *n* classificadores parciais calculados.

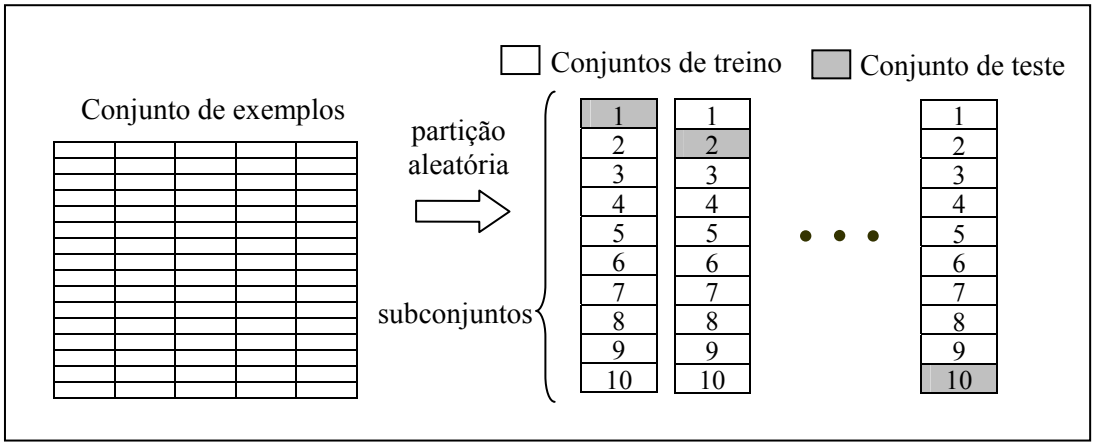

FIGURA 2.7 – Relação entre o conjunto de treino e de teste (Validação Cruzada)

2.4.2.2.4 Bootstrap

A estimação por Bootstrap baseia-se na reamostragem de um conjunto de exemplos, sendo indicada para situações em que este conjunto é bastante limitado.

A partir de um conjunto de exemplos de tamanho *n*, é formado um novo conjunto de treino com a retirada aleatória de elementos do conjunto de exemplos original. Os exemplos omitidos são utilizados na constituição do conjunto de teste.

Logo após o modelo de classificação ter sido gerado e estimada a taxa de erro, os exemplos são recolocados no conjunto original, sendo este processo repetido diversas vezes, geralmente 10, usando cada vez uma amostra diferente de Bootstrap.

A estimativa da taxa de erro no conjunto de dados é dada pela média das taxas de erro em cada iteração realizada durante a execução da técnica de Bootstrap.

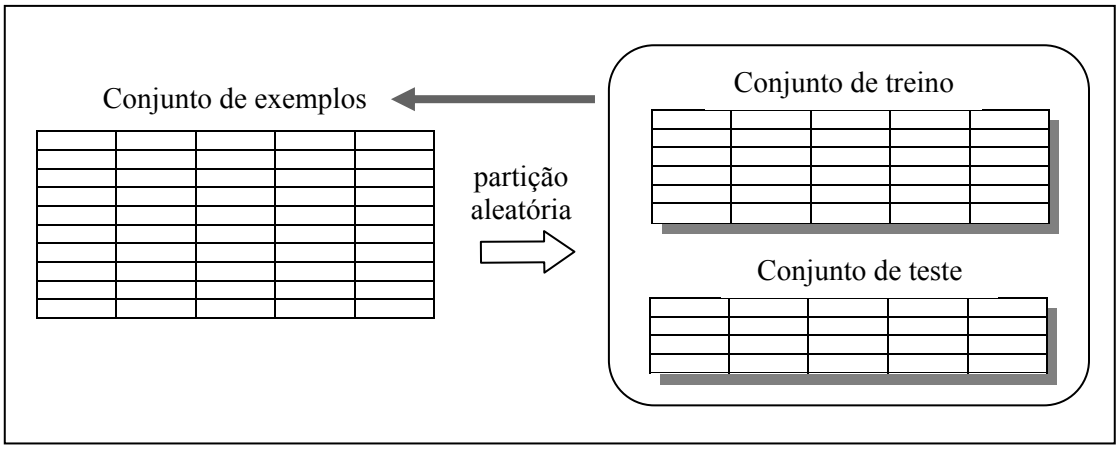

FIGURA 2.8 – Relação entre o conjunto de treino e de teste (Bootstrap)

2.4.3 Classificadores baseados em árvores de decisão

Um classificador tem, como base, um algoritmo de aprendizagem, e este tem como entrada, um conjunto de exemplos, constituído de valores de atributos, sendo seu objetivo a saída de um esquema de classificação, que irá predizer classes baseadas nos valores dos atributos.

Um meio eficiente para representar classificadores a partir de dados, é o uso de árvores de decisão. Esta técnica tem, como ponto forte, a sua eficiência em termos de tempo de processamento e o fornecimento de um meio intuitivo de analisar os resultados, por apresentar como estrutura final do classificador, uma forma de representação simbólica simples e normalmente bastante compreensível, o que facilita a compreensão do problema em análise.

Os classificadores baseados em árvores de decisão surgiram no final dos anos 50, tendo, como principal referência, o trabalho de Hunt [QUI93][FON94], que apresenta vários experimentos de indução. Posteriormente, Friedman, trabalhando em novas pesquisas, desenvolveu o algoritmo CART [QUI93][FON94]. Também o algoritmo ID3, desenvolvido por Quinlan, teve grande influência na área de pesquisa em aprendizado de máquina, tendo, como sucessores, os algoritmos C4.5 [QUI93] e C5.0 [RUL2000].

Os fundamentos dos classificadores baseados em árvore de decisão são idênticos, embora sejam muitas as possibilidades existentes para a sua construção, os algoritmos baseiam-se na sucessiva divisão de um problema em vários subproblemas de menores dimensões, até que a solução para cada um dos subproblemas tenha sido encontrada. Desta maneira, os classificadores baseados em árvores de decisão buscam meios de dividir um problema (conjunto de exemplos) em vários subproblemas (nodos). Esta divisão ocorre até cada um destes subconjuntos conter apenas uma classe, ou até uma das classes demonstrar ser majoritária, não necessitando mais divisões, gerando, em qualquer uma das duas situações, um nodo folha.

A estrutura de uma árvore de decisão é formada por nodos que representam os atributos, por ramos (ligações) provenientes dos nodos, que recebem os possíveis valores do atributo em questão, e de nodos folha, que representam as diferentes classes de um conjunto de exemplos (dados).

A figura 2.9 ilustra a representação da estrutura de uma árvore de decisão, onde são apresentados seus componentes: nodos, arcos e nodos folha.

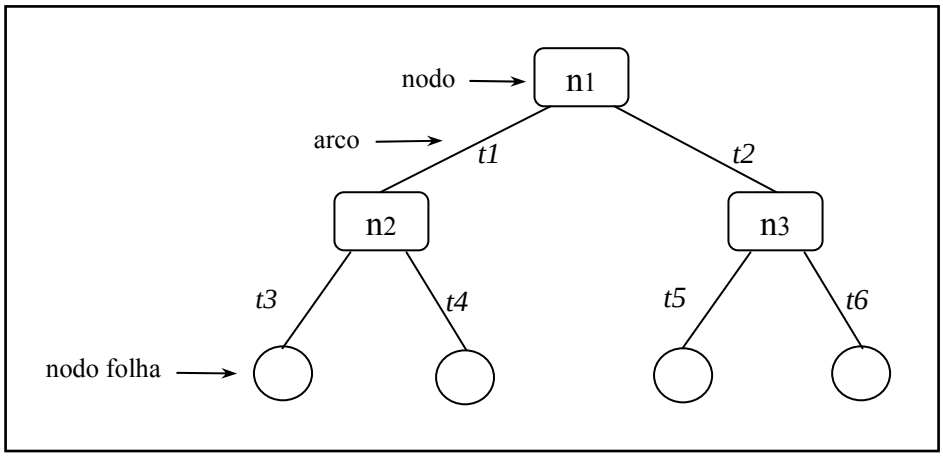

FIGURA 2.9 – Árvore de decisão

Dentro da filosofia de árvores de decisão, a classificação consiste em seguir o caminho determinado pelos sucessivos nodos dispostos ao longo de uma árvore, até ser alcançada uma folha que contém a classe a ser atribuída a um novo exemplo de um conjunto de exemplos.

3 Árvores de Decisão

Neste capítulo refere-se às técnicas utilizadas para a solução dos problemas que se apresentam à construção de árvores de decisão. São apresentados os critérios de seleção do melhor atributo a ser atribuído a cada nodo, os métodos para se determinar a classe que deve ser associada a uma folha da árvore e as técnicas de poda.

3.1 Considerações Iniciais

Árvores de decisão são uma forma simples e eficaz de representar o conhecimento. Elas baseiam-se na abordagem *dividir para conquistar* [QUI93], ou seja, na sucessiva divisão do conjunto de exemplos utilizado para o treino, em vários subconjuntos, até cada um destes subconjuntos pertencer a uma mesma classe, ou até uma das classes ser majoritária, não havendo necessidade de novas divisões.

Os resultados dos vários subconjuntos obtidos com a construção de uma árvore de decisão são dados organizados de maneira compacta, utilizados para classificar novos exemplos.

A figura 3.1 apresenta a representação de um classificador baseado em árvore de decisão, no qual são testados atributos quantitativos e categóricos.

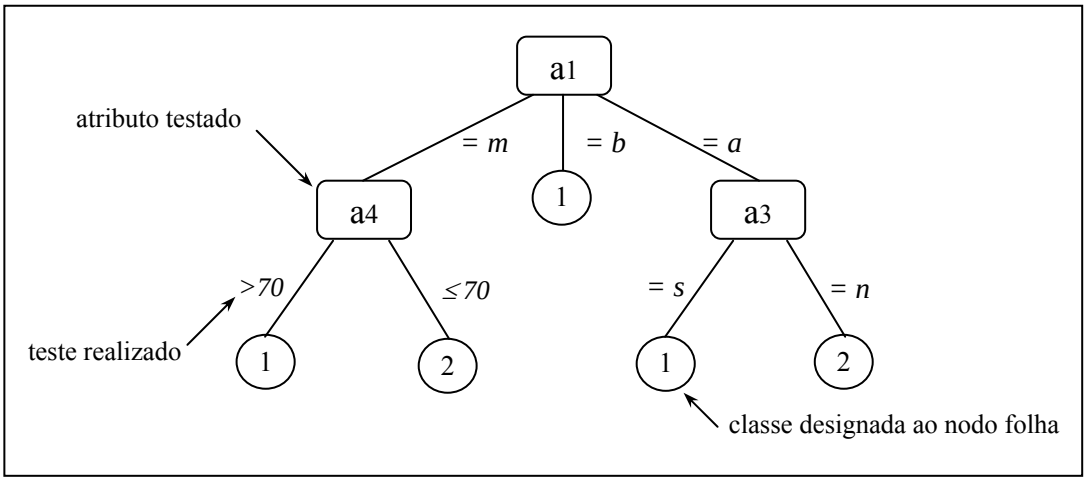

FIGURA 3.1 – Exemplo de um classificador utilizando árvore de decisão

Neste exemplo, os nodos são representados pelos atributos *a1*, *a3* e *a4*, dispostos na árvore de acordo com seu nível informativo. Nos arcos são testados os valores do atributo designado ao nodo a que pertencem. Os testes são realizados de acordo com os valores dos atributos que, quando categóricos, são representados por uma igualdade, por exemplo, = *m*, onde *m* é um valor do atributo. Quando os atributos possuem valores quantitativos, são representados por um intervalo de valor, por

exemplo, > *70*, sendo este intervalo obtido através de cálculo. Cada círculo ao final dos ramos da árvore indica a classe associada aos nodos folha, considerando, no exemplo, *1* como uma classe positiva e *2* como negativa.

A classificação nesta árvore exemplo, como em uma outra qualquer, ocorre ao se percorrer o caminho que se inicia no nodo raiz (*a1*) e se estende até as folhas. A situação é mostrada na figura a seguir.

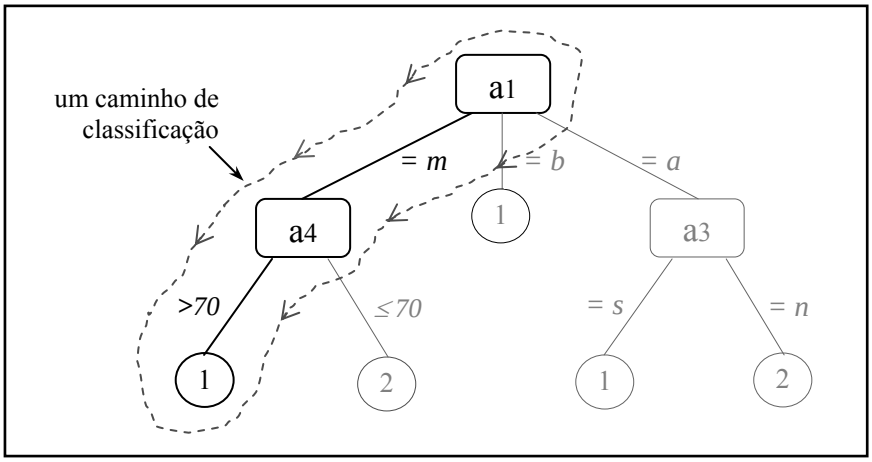

FIGURA 3.2 – Exemplo de um caminho de classificação

O caminho destacado na figura 3.2 indica uma classificação, cujo caso em análise se enquadra na situação apresentada, tendo o atributo denominado *a1* valor igual a *m* e o atributo *a4* valor maior que *70*.

Através dos caminhos descritos por uma árvore de decisão é possível derivar regras. As árvores e as regras são geralmente utilizadas em conjunto. Devido ao fato de as árvores tenderem a crescer muito de acordo com algumas aplicações, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude de as regras poderem ser facilmente modularizadas [ING2000].

A regra extraída do caminho destacado na árvore da figura 3.2, é dada como:

"Se *a1 = m* e se *a4 > 70* então *1"*

3.2 Tipos de Testes

Os testes a serem realizados em um nodo de uma árvore de decisão dependem das características do atributo designado a este nodo, sendo utilizado apenas um atributo por nodo na realização de cada teste, tornando, assim, a estrutura da árvore de fácil compreensão. Com base nos testes definidos e em um conjunto de exemplos, é decidido qual o caminho a percorrer na árvore durante o processo de classificação.

Na determinação dos testes, é possível trabalhar com atributos dos tipos quantitativos, categóricos, ou ainda, com valores desconhecidos.

3.2.1 Atributos quantitativos

Os atributos que possuem características quantitativas permitem grande variedade de testes, implicando, de forma geral, uma certa complexidade de cálculo.

A técnica de construção de árvores de decisão baseada em atributos quantitativos, baseia-se em testes do tipo "*atributo* ≤ *ponto_de_quebra*" ou "*atributo* > *ponto_de_quebra*", nos quais devem ser considerados todos os valores pertencentes ao atributo, pois cada um deles pode representar um possível teste.

Para a utilização deste tipo de atributo, primeiramente é necessária a ordenação de todos os valores do atributo que está sendo trabalhado, em ordem crescente ou decrescente, sendo o atributo ordenado como $\{v_1, v_2, ... v_m\}$. Após a ordenação, seleciona-se o valor (teste) que mais reduz a informação necessária [QUI93][LIU94][BRA2000].

Tomando como exemplo de utilização de atributos quantitativos os valores de um atributo denominado salário: 78 – 98 – 86 – 76 – 125 – 321 – 456 – 99 –125 – 678 – 890 – 67 – 567 – 67, obtém-se, após ordenar estes valores e realizar o cálculo para obter o melhor valor para o teste, o valor 99. A partir dele pode ser montada a estrutura da árvore, exemplificada na figura 3.3.

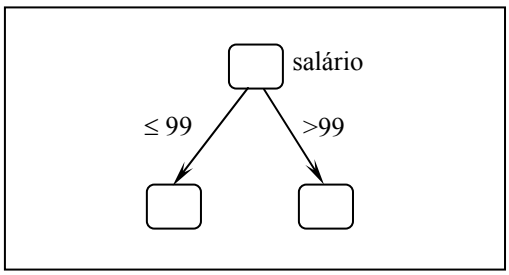

FIGURA 3.3 – Subárvore característica de atributos contínuos

3.2.2 Atributos categóricos

O tratamento de atributos com características categóricas é distinto daquele aplicado a atributos com características quantitativas. Ao tratá-los, devem ser consideradas e avaliadas as abordagens a seguir.

- *Criar um ramo para cada valor do atributo* – esta abordagem torna a árvore bastante detalhada, mas tem desvantagem de criar um grande número de ramos, tornando-a, muitas vezes, de grande dimensão. A figura 3.4 ilustra esta abordagem.

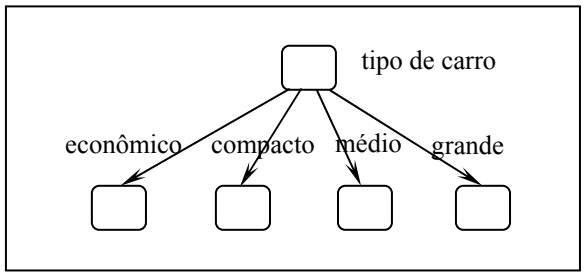

FIGURA 3.4 – Subárvore característica de atributos categóricos: um ramo para cada valor do atributo [SOM98]

- *Criação de nós binários* - a solução apresentada por Hunt [FON94], ilustrada na figura 3.5, sugere a criação de nós binários, atribuindo, a um dos ramos, um dos valores da característica eleita e ao outro, todos os demais valores. Esta solução é naturalmente limitada, não aproveitando todo o poder de discriminação da característica. Apresenta, no entanto, a vantagem da grande simplicidade e inteligibilidade resultante, especialmente útil quando os resultados da geração do classificador automático se destinam a serem interpretados por pessoas consideradas não-especialistas.

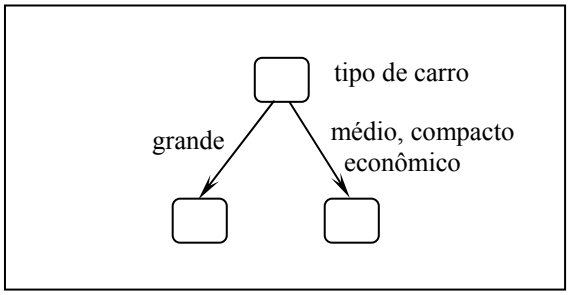

FIGURA 3.5 – Subárvore característica de atributos categóricos: nós binários

- *Características ordenadas* – esta abordagem estabelece uma ordem entre os valores do atributo, possibilitando a construção de árvores binárias apud [FON94].

- *Agrupamento de valores de características em dois conjuntos* – apresentado por Breiman apud[FON94], baseia-se na criação de dois subconjuntos de valores associados respectivamente ao ramo esquerdo e ao direito do nodo em desenvolvimento, sendo uma outra forma de partição binária. Uma ilustração desta abordagem é apresentada na figura 3.6.

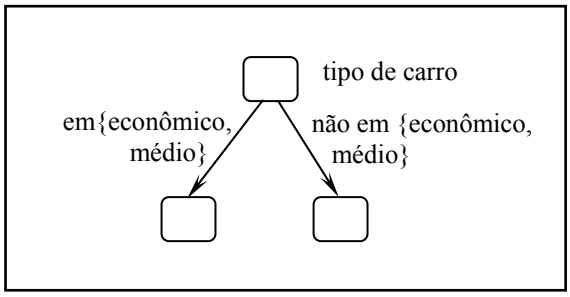

FIGURA 3.6 – Subárvore característica de atributos categóricos: agrupamento de valores de características em dois conjuntos [SOM98]

3.2.3 Valores desconhecidos

O tratamento de valores desconhecidos tem, como objetivo, por um lado, o uso do máximo número possível de exemplos para o treino do classificador, mesmo quando estes contêm valores desconhecidos para algumas características e, por outro, a possibilidade de classificar novos casos ainda que estes se encontrem incompletos [FON94]. Várias soluções são propostas na literatura para o tratamento destes tipos de valores que podem ser:

- simplesmente descartados;
- receber um valor que seja considerado mais provável (mais comum) no conjunto de treinamento;
- considerados como outro possível valor do atributo, de forma que, durante a construção da árvore, cada nodo pode possuir um ramo designado ao atributo testado que possui um valor desconhecido;
- tratados, também, através da probabilidade da distribuição dos valores do atributo. Esta probabilidade é estimada com base nas freqüências observadas dos possíveis valores para o atributo, nos exemplos do nodo em questão, sendo estes valores utilizados para calcular sua impuridade.

3.3 Técnicas de Construção de Árvores de Decisão

De uma maneira geral, na construção de classificadores baseados em árvores de decisão, após terem sido avaliadas e identificadas as características dos dados disponíveis, deve-se determinar o critério para a escolha da característica (atributo) a ser utilizada em cada nodo, assim como a determinação de quando um nodo é considerado folha e a classe a ser atribuída a ele.

3.3.1 Algoritmos de árvores de decisão

O algoritmo para a construção de uma árvore de decisão seleciona um atributo de particionamento e divide o conjunto de exemplos, criando um ramo para cada valor deste atributo. A cada ramo é criado um nodo e novamente selecionado um novo atributo de particionamento para o subconjunto de exemplos atribuído ao novo nodo e, assim, sucessivamente. Este processo tem, como objetivo, separar os exemplos em classes, de modo que os exemplos de distintas classes tendam a serem atribuídos a partições diferentes.

O algoritmo 3.1 mostra um algoritmo genérico para construir árvores de decisão, em que S representa o conjunto de exemplos aplicado à árvore, sendo que inicialmente S contém todos os exemplos de treino.

---

(1) *Se* todos os exemplos no atual conjunto de exemplos S satisfazem um critério de parada
então
(2) cria um nodo folha com algum nome da classe e pára;
senão
(3) seleciona um atributo A para ser utilizado como um atributo de particionamento e cria um nodo com o nome do atributo de particionamento;
(4) escolhe um teste sobre os valores de A, com resultados mutualmente exclusivos e coletivamente exaustivos $R_1,...,R_k$, e cria um ramo, a partir do nodo recentemente criado, para cada teste;
(5) particiona S nos subconjuntos $S_1,...S_k$, tal que cada S_i, i=1..,k, contenha todos os exemplos em S com resultado R_i do teste escolhido;
(6) aplica este algoritmo recursivamente para cada subconjunto S_i, i=1,...k;
fim_senão
fim_se

ALGORITMO 3.1 – Algoritmo genérico para construção de árvores de decisão [FRE97]

Como forma de melhor explicar o funcionamento do algoritmo 3.1, cada passo do algoritmo é identificado por um número seqüencial entre parênteses à sua esquerda. Os passos *(1)* e *(3)* são os passos mais relevantes do algoritmo, por requererem acesso aos dados para avaliar uma regra candidata. Estes são os passos que consomem mais tempo do algoritmo quando a descoberta de conhecimento se dá em grandes bases de dados.

O passo *(1)* do algoritmo 3.1 consiste em decidir o momento de parar o particionamento recursivo. O processo é parado se todos os exemplos no nodo atual possuem a mesma classe. No entanto, se esta condição não for verdadeira, pode ser interessante parar o particionamento no atual nodo, para evitar que a árvore se expanda

muito. Esta forma de parar o particionamento é conhecida como pré-poda e é discutida com mais detalhes na seção 3.4.1

No passo *(2)* se todos os exemplos no recém-criado nodo folha têm a mesma classe, o algoritmo dá ao nodo o nome da classe; contudo, o algoritmo pode dar ao nodo folha o nome da classe mais freqüente ocorrida neste nodo.

No passo *(3)* é computado o melhor atributo candidato à partição, avaliando como, efetivamente, os valores do atributo candidato discriminam as classes dos exemplos.

No passo *(4)* é criado um arco para cada valor distinto, resultante do particionamento do atributo selecionado no passo *(3)*.

O passo *(5)* consiste em designar cada exemplo a um dos arcos criados, de acordo com o valor do atributo de particionamento.

Finalmente, no passo *(6)*, o algoritmo é aplicado recursivamente para cada subconjunto do conjunto de exemplos *S*.

3.3.2 Tabela de freqüências

Para ilustrar o uso de classificadores baseados em árvores de decisão, tomamos como exemplo uma aplicação de *avaliação de concessão de crédito*.

Na tabela 3.1 é apresentado um conjunto de exemplos, que descrevem a realidade dos dados referentes à pessoas que receberam ou não um empréstimo de um banco. Estes dados, extraídos de um banco de dados com informações referentes à créditos, foram selecionados por serem mais relevantes na tomada de uma decisão de conceder ou não novos créditos, tendo como atributos previsores o montante do empréstimo solicitado, a idade da pessoa solicitante, o seu salário e se ela possui ou não conta no banco.

TABELA 3.1 – Conjunto de exemplos de treino [BRA2000]

	montante	**idade**	**salário**	**conta**	**empréstimo**
1	médio	sênior	baixo	sim	não
2	médio	sênior	baixo	não	não
3	baixo	sênior	baixo	sim	sim
4	Alto	média	baixo	sim	sim
5	Alto	jovem	alto	sim	sim
6	Alto	jovem	alto	não	não
7	baixo	jovem	alto	não	sim
8	médio	média	baixo	sim	não
9	médio	jovem	alto	sim	sim
10	alto	média	alto	sim	sim
11	médio	média	alto	não	sim
12	baixo	jovem	baixo	não	sim
13	baixo	sênior	alto	sim	sim
14	alto	média	baixo	não	não

Cada atributo na tabela apresenta valores a serem considerados. O atributo montante é definido dentro de uma faixa que compreende os valores *baixo*, *médio* e *alto*; o atributo idade nas categorias *jovem*, *média* e *sênior*; o salário é considerado dentro de dois grupos estipulados pelo banco, *baixo* e *alto*; o atributo conta identifica se a pessoa possui ou não conta no banco e o atributo empréstimo destaca as situações já ocorridas, em que pessoas foram categorizadas como aptas ou não-aptas a receber empréstimos, tendo o atributo empréstimo dois valores, um deles positivo (*sim*) e o outro negativo (*não*), valor este indicando uma situação em que os requisitos exigidos não são satisfatórios à concessão de créditos. O atributo empréstimo é tomado como atributo meta ou classe, sendo através dele categorizadas as concessões de crédito.

Para posterior utilização deste exemplo na seção que trata da seleção de atributos para a construção de uma árvore de decisão, é tomada a tabela 3.2 como forma de melhor visualizar as situações apresentadas nos dados. Ela mostra uma estrutura geral de visualização dos atributos e o enquadramento de seus valores dentro das classes, que são os valores do atributo meta. Este tipo é conhecido como tabela de freqüências [FRE97][QUI93][WHI94].

TABELA 3.2 – Estrutura geral de uma tabela de freqüências [FRE97]

	C_1	**. . . . .**	**C_m**	**Total**
A_1	S_{11}	. . . . .	S_{1m}	S_{1+}
.	.	. . . . .	.	.
.	.	. . . . .	.	.
.	.	. . . . .	.	.
A_n	S_{n1}	. . . . .	S_{nm}	S_{n+}
Total	S_{+1}	. . . . .	S_{+m}	$\|S\|$

A tabela 3.2 é uma matriz *n* x *m* que inclui os totais das colunas e linhas, onde *n* é o número de distintos valores de um atributo previsor e *m* é o número dos distintos valores do atributo meta. Cada célula S_{ij} (i=1..n, j=1..m) desta matriz contém o número de exemplos do atributo previsor correspondentes ao valor A_i e que satisfazem o atributo meta de valor C_j. A soma dos valores ao longo das linhas (i=1..n) é denotado como S_{i+} e a soma ao longo das colunas (j=1..m) é denotado como S_{+j}. O número total de exemplos que satisfazem o conjunto de valores do atributo em análise (previsor) é denotado por $|S|$.

Na seleção de atributos para a construção de uma árvore de decisão é interessante que para cada atributo previsor, seja montada uma tabela como a 3.2, onde as linhas contenham os valores do atributo previsor; as colunas, os valores do atributo meta e cada célula da tabela contenha o número de exemplos para a respectiva combinação de valores do atributo previsor e do meta, como mostra a tabela de freqüências 3.3. Estando esta tabela elaborada, fica mais fácil, a partir dela, determinar a heterogeneidade dos atributos previsores para a construção da árvore de decisão.

TABELA 3.3 – Tabela de freqüências do atributo Montante

	Sim (+)	**Não (-)**	**Total**
médio	2	3	*5*
baixo	4	0	*4*
alto	3	2	*5*
Total	*9*	*5*	*14*

A árvore da figura 3.7 corresponde à tabela de freqüência 3.3, elaborada a partir da tabela de exemplos de treino 3.1, no qual a partição efetuada, baseada no atributo montante, resulta em três subárvores, cada uma delas correspondendo aos valores deste atributo. O valor médio possui cinco exemplos: dois positivos de uma classe *sim* e três negativos de uma classe *não*. O valor baixo possui somente valores positivos e o valor alto, cinco exemplos; três positivos e dois negativos.

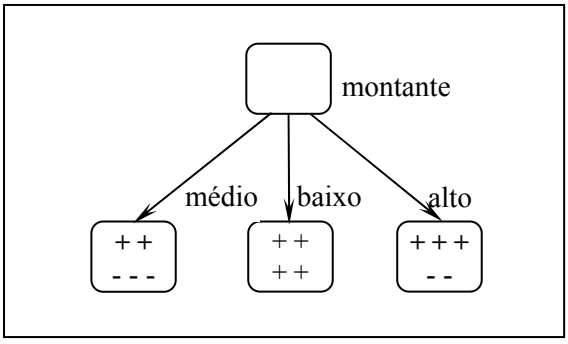

FIGURA 3.7 – Subárvore atributo montante

3.3.3 Critérios de seleção de atributos

Os critérios para a seleção de atributos são utilizados para determinar qual atributo pertencente ao conjunto de exemplos melhor se enquadra no nodo em análise, durante a construção da árvore de decisão. Deste modo avalia-se a melhor partição a ser realizada, de acordo com a capacidade informativa do atributo.

3.3.3.1 Ganho de Informação

Um dos mais antigos e conhecidos critérios de seleção de atributos é o *ganho de informação* (Gain), utilizado pelos também conhecidos algoritmos ID3 e C4.5.

O ganho de informação tem, como base geradora, uma medida conhecida como *entropia* [QUI93][MIT97][DAN2000]. O ganho pode ser conceituado como a redução esperada da entropia, tendo, como função, a seleção de atributos utilizados no particionamento de um conjunto de dados.

A entropia empregada para a obtenção do ganho de informação tem sua origem na teoria da informação e baseia-se no trabalho realizado por Claude Shannon and Warren Weaver, em 1949 [BER97][JAY2000]. A entropia usa, como estratégia, a redução da impuridade, ou seja, mede a quantidade de informação necessária para codificar uma situação encontrada em um nodo [CAM2000][MAR2000]. A impuridade é máxima se todas as classes de um nodo têm igual prioridade e mínima quando existe apenas uma classe. Na teoria da informação, a informação é medida em *bits*.

A entropia é dada pela fórmula (3.1) que determina o número de exemplos de *S* pertencentes à classe C_j, podendo o atributo ter *m* possíveis valores:

$$Entropia(S) = \sum_{j=1}^{m} - p_j \log_2 p_j \qquad (3.1)$$

Onde: S é o conjunto de exemplos

m é o número de classes

p_j é a proporção de S pertencer à classe *j, tendo então:*

$$p_j = \frac{|S_j|}{|S|} \qquad (3.2)$$

Onde: $|S_j|$ é o número de exemplos classificados na j-ésima partição

$|S|$ é o número total de exemplos do conjunto S

Tomando o exemplo ilustrado na tabela 3.3, em que o conjunto *S* é uma coleção de 14 exemplos com 9 instâncias positivas e 5 negativas, a entropia no conjunto de exemplos é:

$$Entropia\ (S) = -\ (9/14)\ \log_2 (9/14) - (5/14)\ \log_2 (5/14) = 0.940$$

A entropia no atributo montante é calculada para cada um dos seus valores, sendo:

$$Entropia\ (montante, médio) = -\ (2/5)\ \log_2 (2/5) - (3/5)\ \log_2 (3/5) = 0{,}971$$

$$Entropia\ (montante, baixo) = -\ (4/4)\ \log_2 (4/4) - (0/4)\ \log_2 (0/4) = 0$$

$$Entropia\ (montante, alto) = -\ (3/5)\ \log_2 (3/5) - (2/5)\ \log_2 (2/5) = 0{,}971$$

Com base na medida da entropia, uma classificação é considerada perfeita, se todos os membros de um conjunto *S* pertencem a uma mesma classe, sendo a entropia igual a *zero*. Por exemplo, se todos os membros são positivos, $p_+ = 1$, então $p_- = 0$ (zero). Quando a entropia é igual a um, é dito que os membros de um conjunto foram classificados ao acaso, pois o conjunto possui número igual de exemplos positivos e

negativos. Se o conjunto contiver números diferentes de exemplos positivos e negativos, a entropia estará entre 0 (zero) e 1 (um). A figura 3.8 mostra a forma da curva da entropia, com p_+ variando entre 0 e 1.

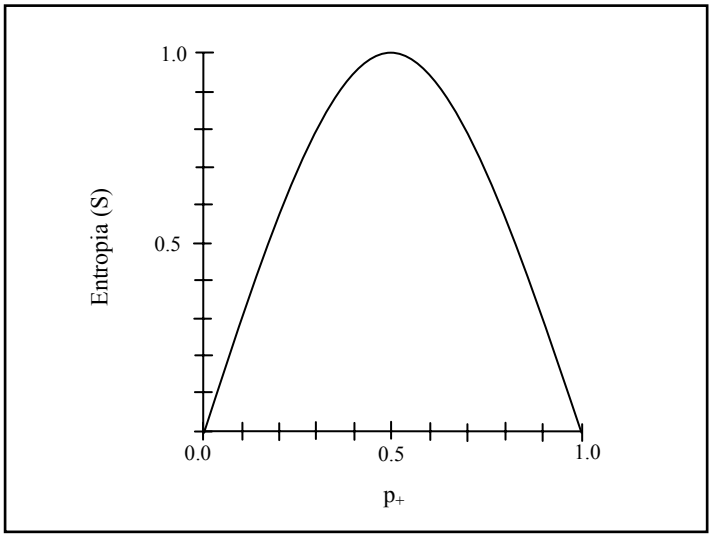

FIGURA 3.8 – A medida da entropia para exemplos positivos variando entre 0 e 1 [MIT97]

O ganho de informação (Ganho(S,A)) – resultante da adoção de uma partição baseada num dado atributo será assim dado por:

$$Ganho(S, A) = Entropia(S) - \sum_{j=1}^{m} \frac{|S_j|}{|S|} Entropia(S_j) \qquad (3.3)$$

Onde: *Ganho (S, A)* é o ganho do atributo *A* sobre o conjunto *S*

$|S_j|$ é o subconjunto de *S* no qual o atributo *A* tem valor *j*

A partir destes cálculos é possível escolher a partição que permite obter um maior ganho de informação.

O calculo do ganho do atributo montante é:

*Ganho(S, montante) = 0,940 – ((5/14)*0,971+(4/14)*0+(5/14)*0,971)*

= 0,246

Para a análise do atributo mais informativo é necessário o cálculo do ganho de todos os atributos envolvidos na análise. Abaixo são mostrados os ganhos dos atributos idade, salário e conta.

*Ganho(S, idade) = 0,940 – ((4/14)*1+(5/14)*0,971+(5/14)*0,722) = 0,049*

*Ganho(S, salário) = 0,940 – ((7/14)*0,592+(7/14)*0,985) = 0,151*

*Ganho(S, conta) = 0,940 – ((8/14)*0,811+(6/14)*1) = 0,047*

De acordo com o ganho de informação, o atributo com maior ganho, portanto mais informativo dentre os calculados e que melhor prediz o atributo meta, é o montante com valor de 0,246 bits, designado como nodo raiz.

3.3.3.2 Razão do Ganho de Informação

Para atributos que possuem muitos valores o ganho de informação apresenta a desvantagem de tender a ser muito grande, fazendo gerarem-se árvores muitos largas, que não são desejáveis. Para contornar esta situação, foi proposta por Quinlan [QUI86] uma alternativa baseada na *razão do ganho de informação* (*RGanho*).

$$RGanho(S, A) = \frac{Ganho(S, A)}{Info(S, A)} \qquad (3.4)$$

A informação do atributo *A* em relação ao conjunto *S* é dada por:

$$Info(S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \qquad (3.5)$$

Onde: S_i é o subconjunto de exemplos resultante do particionamento de *S* pelos *n* valores do atributo *A*

3.3.3.3 Gini

O Gini é um critério utilizado na seleção de atributos, sendo aplicável a árvores m-árias, cujo objetivo é a minimização da impuridade [AGR97][BIO97].

Para um dado conjunto de dados *S* contendo exemplos de *m* classes, *Gini(S)* é definido como:

$$Gini(S) = 1 - \sum_{j=1}^{m} p_j^2 \qquad (3.6)$$

Onde: p_j é a freqüência relativa da classe *j* em *S*

Se dividirmos *S* em dois subconjuntos, S_1 e S_2, com n_1 e n_2 exemplos respectivamente, o novo índice de divisão dos dados, $Gini_{split}(S)$, é dado por:

$$Gini_{split}(S) = \frac{n_1}{n} Gini(S_1) + \frac{n_2}{n} Gini(S_2) \qquad (3.7)$$

3.3.3.4 Qui-quadrado

O critério qui-quadrado (x^2), desenvolvido pelo estatístico inglês Karl Pearson, é um teste estatístico comumente usado para comparar dados observados com dados esperados, sendo o resultado obtido de acordo com uma hipótese específica [WHI94][BEM97][HOL2000].

Para o desenvolvimento do calculo do x^2 é necessário, primeiramente, criar uma tabela de freqüências para o atributo em análise como a apresentada na seção 3.3.2. Com base na tabela é calculada a freqüência esperada para cada célula por meio da fórmula (3.8). Com as freqüências estabelecidas pode-se então aplicar a fórmula (3.9) e obter o x^2.

$$E_{ij} = \frac{S_i \times S_j}{|S|} \qquad (3.8)$$

$$x^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(S_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.9)$$

Onde: E_{ij} é a freqüência esperada na célula S_{ij} da matriz apresentada na tabela 3.2

3.3.3.5 Twoing

O critério Twoing aplica-se a árvores binárias, defindo a impuridade de um nodo como:

$$i(S) = \frac{p_L p_R}{4}\left[\sum_j \left| p(j/S_L) - p(j/S_R) \right|\right]^2 \qquad (3.10)$$

Onde: P_L é a probabilidade do nodo descendente esquerdo

P_R é a probabilidade do nodo descendente direito

$P(j/S_L)$ é a probabilidade da classe j no nodo descendente esquerdo

$P(j/S_R)$ é a probabilidade da classe j no nodo descendente direito

3.3.4 Determinação da classe associada à folha

Durante a construção de uma árvore de decisão, ao se deparar com um nodo considerado folha, é necessário determinar qual classe deverá estar associada a ele.

A determinação da classe associada à folha pode ser realizada através da atribuição da classe que minimiza a taxa de erro da classificação ou a atribuição da classe que minimiza os custos da classificação.

3.3.4.1 Atribuição da classe mais provável

Nesta aproximação é atribuída à folha a classe mais provável dentro dos exemplos que se encontram associados a ela [FON94].

$$\max_j (p_i) = \max_j \frac{N_j}{N} \qquad \text{com j=1..m} \qquad (3.11)$$

Onde: N é o número total de exemplos na folha

N_i é o número de exemplos da classe

m é o número de classes

3.3.4.2 Determinação baseada na noção de custo

A determinação de uma classe baseada em custos, tem, como objetivo, a minimização dos custos provenientes da adoção de uma determinada classe e não a probabilidade do erro resultante. Para determinar a classe a atribuir a cada nodo folha, deve-se considerar o conceito de matriz de custos apresentado na seção 2.4.2.1, sendo o custo dado por:

$$Custo(m) = \sum_{i=1}^{m} p_i S_{i,j} \qquad (3.12)$$

Onde: p_i é a probabilidade da classe *i*

$S_{i,j}$ é o valor da linha *i*, coluna *j* da matriz de custos

m é o número de classes

3.4 Técnicas de Poda

Os métodos utilizados para a construção de árvores de decisão, geralmente, resultam em árvores de grandes dimensões, com isso tornando-as complexas e, muitas vezes, comprometendo o seu desempenho.

A poda é considerada uma parte importante do processo de construção de árvores de decisão, pois visa a limitar as dimensões da árvore, removendo partes que não contribuem para uma classificação mais precisa, produzindo desta forma uma estrutura menos complexa e, conseqüentemente, fazendo a árvore apresentar melhor desempenho e ser de mais fácil compreensão.

De forma a contornar o problema da demasiada dimensão e da precisão da classificação, as técnicas de poda podem ser aplicadas de duas formas: a primeira interrompe a construção da árvore quando algum critério de parada é satisfeito e é conhecida como pré-poda. A segunda, a pós-poda, é aplicada somente após a conclusão da árvore, ou seja, quando todos os exemplos do conjunto de exemplos tenham sido distribuídos ao longo da árvore e visa a reduzir as dimensões da árvore até serem consideradas ideais.

As duas formas de poda buscam remover partes da árvore que não contribuam para o rigor da classificação, produzindo estruturas menos complexas. Embora a pré-poda tenha como vantagem o fato de se evitar a construção de uma estrutura de árvore que virá posteriormente a ser destruída, a pós-poda é mais confiável, devido ao fato de utilizar todos os exemplos na construção da árvore.

3.4.1 Pré-poda

Na pré-poda, o particionamento da árvore de decisão pode chegar ao fim durante a fase de construção. Usualmente, o critério de parada é calculado para dar uma estimativa do ganho esperado na continuação da construção da árvore, sendo o teste x^2 muito utilizado na obtenção deste critério. A construção é encerrada quando um limite mínimo de ganho não é o esperado.

Nesta técnica a poda é realizada concomitantemente com a construção da árvore. Isso evita a construção de subárvores que venham a ser podadas posteriormente à sua constituição. Mas a pré-poda tem, como inconveniente, que as decisões de poda têm de ser tomadas ao mesmo tempo em que a construção da árvore é realizada, decisões estas que podem ser muito complexas, o que onera o custo do algoritmo, muitas vezes anulando a vantagem da poda antecipada.

3.4.2 Pós-poda

A pós-poda é efetuada nos nodos não folha, a partir do cálculo do erro de uma árvore e de todas as suas subárvores, onde são examinados cada um dos nodos não folha, começando por baixo, ou seja, pelos nodos mais próximos das folhas. Se a sua substituição por uma folha ou pelo seu ramo mais utilizado conduzir a um menor erro, este é substituído.

Para se verificar a necessidade de uma árvore de decisão ser podada, é preciso estimar a taxa de erro apresentada pela árvore e também a taxa de erro introduzida pela poda nos testes ao longo da árvore, assim como nos nodos folha.

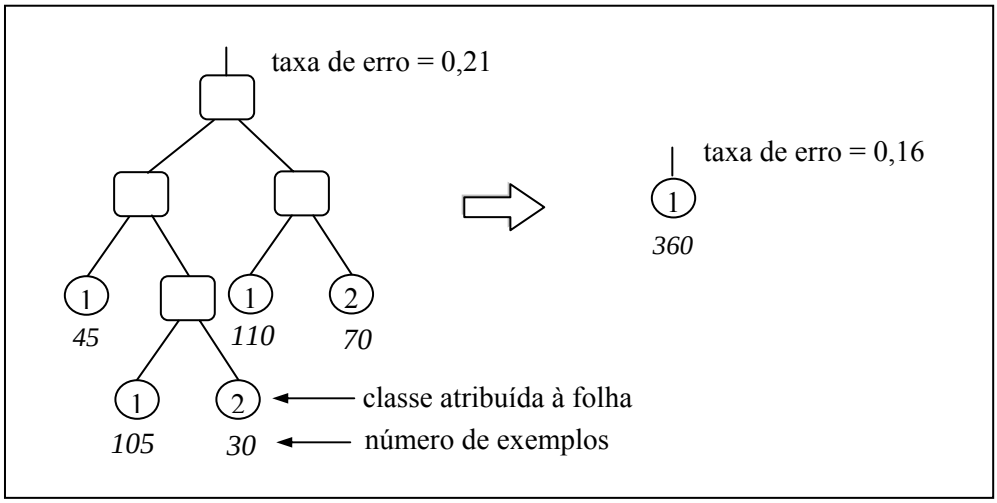

FIGURA 3.9 – Exemplo de remoção de nodos

A figura 3.9 ilustra a poda de ramos da árvore que não contribuem para uma boa classificação. Neste exemplo, é tomado como meio de medir a eficiência da árvore

a medida taxa de erro. A taxa de erro na subárvore antes da poda era de 0,21 e, após a poda, passou a 0,16, fazendo os exemplos anteriormente distribuídos entre as duas classes, 1 e 2, passarem a ser atribuídos à classe 1, devido à sua maior relevância.

A vantagem desta técnica está em muitas vezes, só ser possível perceber que uma árvore deve ser podada após a sua construção, para qual se pode analisar o contexto das classificações que a árvore proporciona.

4 Algoritmos de Indução de Árvores de Decisão

Neste capítulo são apresentados os algoritmos de indução de árvores de decisão: ID3, CART e C4.5. Estes algoritmos são referências fundamentais ao se tratar de geração automática de árvores de decisão. A ferramenta utilizada no desenvolvido do estudo de caso descrito no capítulo 5 denomina-se Sipina-W. Tanto ela, como os algoritmos são explanados nas seções a seguir.

4.1 Algoritmos de Árvore de Decisão

Esta seção apresenta uma breve explanação sobre os algoritmos ID3, CART, C4.5 e Sipina, descrevendo alguns pontos fundamentais, como: suas origens; suas características e as metodologias aplicadas à construção de árvores de decisão.

4.1.1 ID3

No final dos anos 70, Ross Quinlan introduziu um algoritmo de árvore de decisão chamado ID3. Este foi um dos primeiros algoritmos de árvore de decisão, tendo sua elaboração baseada em sistemas de inferência e em conceitos de sistemas de aprendizagem já existentes na época. O ID3 foi usado inicialmente para tarefas de aprendizagem, como no jogo de xadrez, no qual a estratégia é fundamental. Desde então, o ID3 foi aplicado a uma grande variedade de situações acadêmicas e industriais.

O ID3 foi desenvolvido visando à resolução de problemas que contenham atributos categóricos. Este algoritmo necessita que os valores dos atributos não possuam ruídos, sendo assim, estes valores devem ser tratados previamente. Ele adota o critério ganho de informação para a escolha da característica (atributo) a ser atribuído a cada nodo. A estrutura da árvore gerada pelo algoritmo é bastante simples, cada atributo permite a divisão do conjunto de treino num número de subconjuntos igual a sua cardinalidade.

4.1.2 CART

O algoritmo CART (Classification and Regression Trees) foi apresentado pelos estatísticos Leo Breiman, Jerone Friedman, Richard Oslen e Charles Stone, no trabalho intitulado “Classification and Regression Trees", publicado em 1984. Este trabalho é um importante referencial no que se refere à aprendizagem automática, sendo citado na maioria da bibliografia sobre mineração de dados.

Uma das principais características deste algoritmo está na capacidade de gerar árvores de reduzidas dimensões, de elevado desempenho, possuindo grande capacidade de generalização.

Com a metodologia empregada pelo CART para a construção de árvores de decisão é possível trabalhar com atributos previsores categóricos ou quantitativos. No particionamento baseado em atributos categóricos são testadas todas as possibilidades de formação de dois subconjuntos com os possíveis valores. No particionamento relacionado à atributos quantitativos é adotada a técnica de pesquisa exaustiva do ponto de divisão, descrita na seção 3.2.1.

A árvore resultante é baseada na técnica recursiva de divisão binária. O processo é binário porque cada nodo é separado sempre em exatamente dois subconjuntos e à medida que se percorre a árvore, da raiz às folhas são respondidas questões simples do tipo sim/não. A recursividade se dá a cada subconjunto gerado, até que não seja possível ou não seja necessário mais efetuar partições na árvore. A eleição da melhor característica é efetuada geralmente com base num dos dois critérios: Gini ou Entropia, e a atribuição de uma classe a cada folha é realizada com base no critério da classe mais provável ou da minimização dos custos.

Na ferramenta Sipina-W, este algoritmo utiliza os critérios: Gini e Twoing para gerar árvores de decisão.

4.1.3 C4.5

O algoritmo C4.5 foi apresentado por Ross Quinlan em seu trabalho intitulado “C4.5: Programs for machine learning”, publicado ano de 1993 [QUI93]. Trata-se de um aprimoramento do algoritmo ID3, trabalhando com atributos categóricos e quantitativos, bem como aqueles de valores desconhecidos, e adotando o sistema de poda.

Durante o processo de construção da árvore de decisão o C4.5 os atributos categóricos podem ser particionados de duas maneiras: um ramo distinto a cada valor do atributo ou a formação de agrupamentos de valores em vários conjuntos. Para as partições efetuadas com base em atributos contínuos é utilizado o método de pesquisa exaustiva do ponto de divisão, gerando árvores binárias.

A eleição da melhor característica pode ser efetuada pelo critério ganho de informação ou pelo critério da razão do ganho de informação. A cada folha é atribuída a classe mais provável, ou seja, a majoritária.

Na ferramenta Sipina-W, o C4.5 efetua partições com base no critério razão do ganho de informação.

4.2 A Ferramenta Sipina-W

Sipina-W é uma ferramenta de mineração de dados, cuja finalidade é extrair conhecimento a partir de base de dados pelo uso de classificadores, sendo o conhecimento obtido representado por árvores de decisão e por regras. Esta ferramenta foi desenvolvida na Universidade de Lyon, França, em 1995, por um grupo de pesquisadores coordenados pelo professor D. A. Zighed, sendo seu uso gratuito para fins acadêmico e de pesquisa.

Atualmente, são muitas as ferramentas que realizam o processo de descoberta de conhecimento em bases de dados. O artigo "A Survey of Data Mining and Knowledge Discovery Software Tools" [GOE99], escrito por Michael Goebel e Le Gruenwald, mostra uma visão geral destas ferramentas até então existentes. Neste artigo, para cada ferramenta são apresentadas suas características gerais, sua conectividade a bancos de dados e suas características de mineração de dados.

A ferramenta Sipina-W foi a escolhida para a realização do estudo de caso proposto neste trabalho, por: implementar diversos algoritmos geradores de árvores de decisão; possibilitar a escolha por executar a mineração de dados de forma automática ou interativa, ambas possibilitando a obtenção de informações dos nodos, mas a interativa com a possibilidade de "forçar" o uso de um atributo que conste do conjunto de dados utilizado; apresentar, durante a obtenção do conhecimento, uma forma gráfica clara da árvore de decisão, tornando bastante compreensível o resultado apresentado e possibilitando gerar regras a partir da árvore obtida; ser uma ferramenta de uso livre para o uso acadêmico, entre outras características e recursos mostrados por ela.

4.2.1 Interface

A interface do Sipina-W (figura 4.1) é amigável. Durante sua execução, permanecem abertas três janelas, a de dados, a da matriz de classificação e a do gráfico. A janela dos dados é uma espécie de planilha no qual são trabalhados os parâmetros e os dados do conjunto de treino e de teste, sendo possível, a partir dela, executar o processo de aprendizado automático.

A janela do gráfico destina-se à visualização dos resultados da aprendizagem, demonstrados através de uma árvore de decisão. Cada retângulo da árvore é um nodo e, no interior de um nodo está a distribuição das classes. Ao lado dos nodos fica o seu nível e o seu número, ambos representam a identificação do nodo. Também, acima do nodo, em seu ramo, consta o nome do atributo o qual foi particionado e, abaixo dele, aparece o nome do atributo que ensejou criar um próximo ramo. Em cada nodo é possível visualizar a distribuição dos atributos em relação às classes, sendo esta informação útil na aprendizagem interativa, possibilitando a verificação de quais atributos são mais informativos.

A janela matriz de classificação mostra os resultados da aprendizagem. Nela, podemos visualizar como a análise realizada pelo Sipina_W pôde predizer os casos analisados. Nesta janela, é apresentada uma matriz de confusão, onde são enquadrados os casos bem e mal classificados.

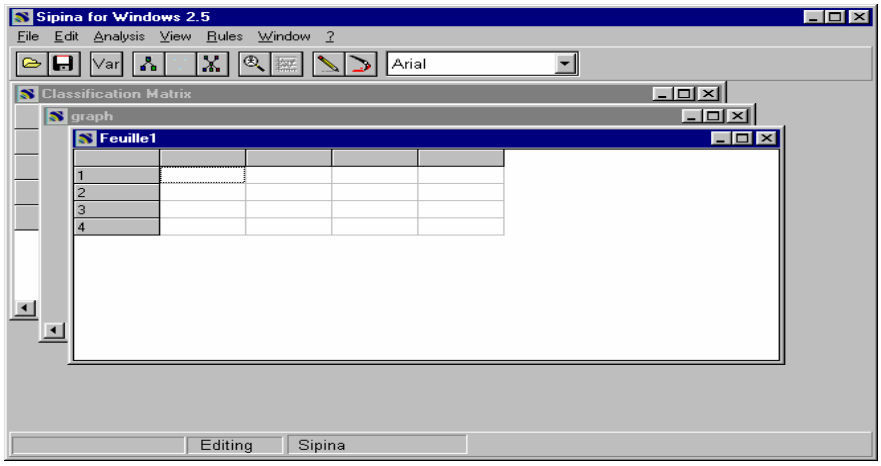

FIGURA 4.1 – Interface da ferramenta Sipina-W

4.2.2 Características

Os arquivos utilizados pelo Sipina_W têm o formato ASCII. Para a obtenção dos arquivos neste formato, pode-se utilizar a planilha eletrônica que acompanha a ferramenta (figura 4.2). Esta planilha importa dados com formato do Microsoft Excel (.xls) e com o formato texto (.txt).

Para gerar os arquivos utilizados na mineração, os dados inseridos na planilha devem ser selecionados e, então, executado o comando que exporta o conjunto de dados para o Sipina-W. Com a execução do comando, são gerados arquivos no formato (.dat), para os dados, e outro cujo, teor são os parâmetros referentes a estes dados (.par). Também, é possível, gerar a partir da planilha, o arquivo de validação (.val).

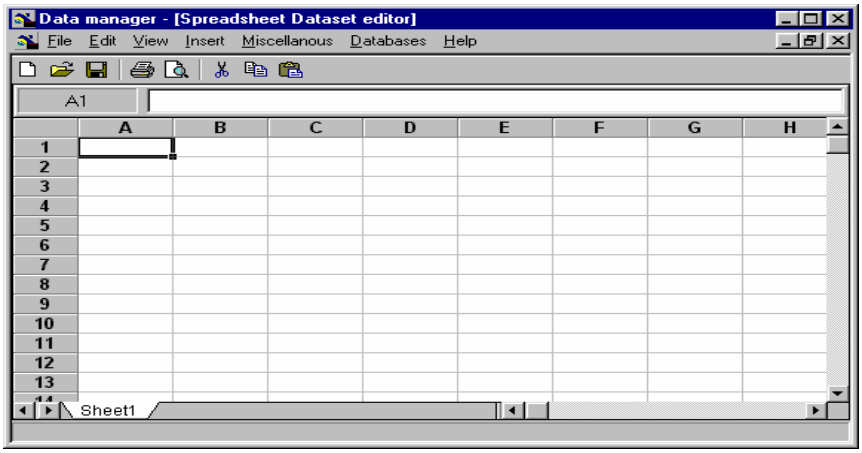

FIGURA 4.2 – Planilha da ferramenta Sipina-W

O procedimento de aprendizagem pode ser realizado de forma automática ou interativa. Ao usar o procedimento de aprendizagem automática, somente é necessário escolher o algoritmo desejado e executar a análise. A forma de aprendizagem interativa permite que sejam executadas operações "forçadas" (divisão ou fusão de nodos), assim como a escolha dos atributos a serem usados em cada nodo. A cada nodo selecionado para uma nova partição, é possível verificar a distribuição das classes, a função de distribuição em cada atributo, quais atributos são mais qualificados para serem utilizados.

O resultado da aprendizagem que é representado por árvores de decisão, pode ser traduzido de árvores para regras de produção, sendo as regras armazenadas em um "Sistema de Base de Conhecimento" (SBC). A utilização das regras de produção tem, por finalidade, a validação e a generalização da aprendizagem.

Para efetuar os testes e a avaliação dos dados, pode-se utilizar um arquivo de treino e um de teste, sendo primeiramente trabalhado o arquivo de treino, gerando a árvore de decisão e, em seguida, as regras correspondentes a árvores seguindo-se, a validação a partir do arquivo de teste. O teste e a avaliação dos dados podem ser efetuados também pelo uso de validação cruzada.

4.2.3 Limitações

A limitação do SIPINA-W é de 16.384 atributos e de 500.000.000 exemplos. Esta limitação é considerada teórica, pois todo o conjunto de dados é carregado na memória do computador antes de ser realizado o aprendizado, sendo a memória do computador a real limitação.

Outras limitações se referem aos atributos, que não podem conter mais de 10 valores; à análise, tecnicamente limitada a 25 níveis; ao número de nodos, a cada nível limitado a 50.

5 Estudo de Caso

O estudo de caso descrito neste capítulo tem, como finalidade, a extração de conhecimento a partir de bases de dados associadas à área da saúde, mediante a construção de árvores de decisão. O estudo contempla todo o processo de descoberta de conhecimento em bases de dados, e baseia-se na seqüência de etapas propostas em [CAB97], iniciando pela determinação dos objetivos a serem atingidos e finalizando com a etapa de análise dos resultados obtidos, já comentadas nas seções 2.2.1.2 a 2.2.1.5. A abordagem de [CAB97] foi adotada para o desenvolvimento do estudo de caso, por apresentar clara seqüência de desenvolvimento das etapas e, para relatar uma outra abordagem ao leitor deste trabalho, uma vez que, habitualmente, são citadas na literatura as abordagens de autores como [ADR96], [BRA96] e [FAY96].

A seguir, é apresentada uma explanação do domínio trabalhado no estudo de caso. As seções subseqüentes relatam o processo de descoberta de conhecimento desenvolvido. Ao final do capítulo é apresentada uma avaliação do processo e dos conhecimentos obtidos.

5.1 Descrição do Domínio

Para a compreensão do domínio utilizado no estudo de caso, é necessário, primeiramente, compreender o que é uma AIH e sua finalidade. A AIH é um documento utilizado no controle de internações hospitalares, no qual se pode identificar, entre outras informações, os dados do paciente, o diagnóstico que motivou o tratamento hospitalar e os serviços prestados durante a sua internação, por profissionais e pelo hospital, tendo em vista os procedimentos realizados. É por este documento que hospitais, profissionais de saúde e serviços de diagnose e terapia (SADT) são habilitados a receber pelos préstimos fornecidos.

Mensalmente os hospitais encaminham os seus faturamentos correspondentes ao período à Secretaria Municipal de Saúde de Pelotas (SMSP), para poderem receber pelos serviços prestados em cada internação. Este faturamento é composto por um conjunto de documentos: as AIHs; os laudos médicos, documentos utilizados na solicitação de internações de pacientes; pelos prontuários e pelos lançamentos das contas hospitalares que discriminam gastos com os procedimentos realizados e outras despesas hospitalares. Ao receber esta documentação, a SMSP, através do departamento de Controle e Avaliação, revisa e faz uma avaliação dos faturamentos dos hospitais, localizando “erros” nos dados informados. Ao avaliar os itens dos faturamentos, o departamento responsável busca identificar através da experiência das especialistas e das regras de emissão de AIHs (anexo 1) se algum item possui alguma cobrança indevida. Esta avaliação é realizada manualmente, consumindo um tempo significativo, devido ao número de AIHs produzidas mensalmente.

Basicamente o foco da AIHs são os custos dos serviços prestados e os procedimentos realizados durante uma internação, estes, agrupados de acordo com as especialidades médicas afins. As especialidades são definidas dentro das áreas: de

clínica médica, cirurgia geral, obstetrícia, psiquiatria, pediatria, tisiologia, crônico e FPT (fora de possibilidade terapêutica), reabilitação e psiquiatria hospital/dia.

Com base em um estudo efetuado sobre os procedimentos mais usuais e importantes que constam nas AIHs, optou-se por trabalhar neste estudo de caso com os procedimentos referentes à patologia "acidente vascular cerebral" (AVC), por julgar que com os dados associados a esta patologia, seria possível extrair um maior número de padrões.

O AVC, comumente conhecido como derrame, é uma patologia compreendida como um súbito déficit na irrigação sangüínea do cérebro, causando lesão celular e danos às funções neurológicas, visto o cérebro ser um órgão que exige suprimento sanguíneo adequado e constante. Este fenômeno ocorre quando as artérias carótidas (figura 5.1), vasos do pescoço que levam sangue para o cérebro, ficam bloqueadas por placas de aterosclerose (placas de gordura, de cálcio, de coágulos sanguíneos, etc), podendo, de tempos em tempos, desprender-se, indo para o cérebro e entupindo vasos menores, ocasionando falta de circulação e matando as células cerebrais que receberiam sangue. Dependendo dos locais onde tais células morrem, o paciente sofrerá seqüelas maiores ou menores, podendo até mesmo falecer [OSU93] [BAR2002].

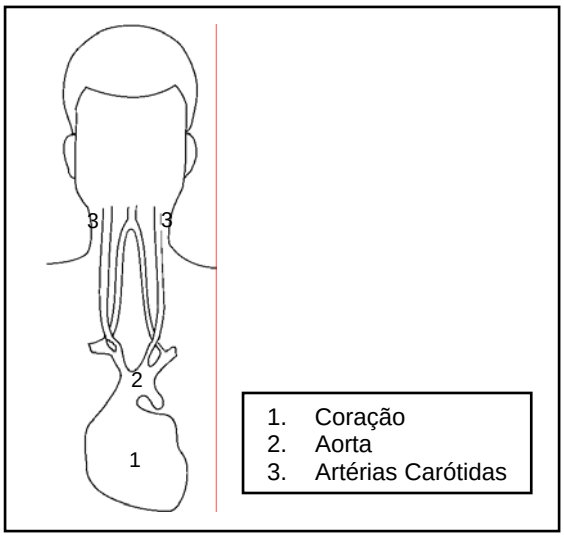

FIGURA 5.1 – Ilustração que mostra as Artérias Carótidas [SBN2002]

O AVC é a uma das principais causas de morte no mundo e de incapacidade física e mental. Sua incidência aumenta significativamente com a idade, podendo afetar tanto mulheres como homens quase igualmente. O AVC pode apresentar-se de duas formas:

- Isquêmico: quando não há passagem de sangue para determinada área, por uma obstrução no vaso ou redução no fluxo sangüíneo do corpo.
- Hemorrágico: quando o vaso sangüíneo se rompe, extravasando sangue.

Entre estas duas formas, o AVC hemorrágico, segundo a literatura e ao médico neurologista consultado, é o maior causador de mortes.

5.2 O Processo de DCBD

Para o desenvolvimento do processo de DCBD, é fundamental o conhecimento do domínio a ser trabalhado. Em [CAB97], a compreensão do domínio não é tratada como uma etapa, nesta abordagem a compreensão está associada à determinação dos objetivos, tornando o objetivo definido para o processo, diretamente ligado à compreensão.

Os dados estudados foram obtidos no Centro de Processamento de Dados da SMSP. Uma cópia do manual das AIHs foi cedida pelo departamento de Controle e Avaliação. Após um prévio estudo dos dados e do funcionamento do sistema de AIHs, realizaram-se entrevistas com as médicas responsáveis pelas revisões e análises, sendo passadas informações gerais sobre a rotina de AIHs e fornecidos alguns esclarecimentos sobre os dados.

5.2.1 Determinação dos objetivos

Com base nas informações sobre o domínio e os dados, foram realizados estudos e testes para verificar a possibilidade de obtenção de padrões válidos mediante a construção de classificadores baseados em árvores de decisão. Inicialmente parecia um pouco difícil estabelecer padrões entre os dados disponíveis, pois as informações que constam nas tabelas do sistema de AIHs não são muito relevantes para tal aplicação. Mas, ao longo de vários testes, surgiram alguns problemas classificatórios que poderiam ser trabalhados.

A partir das entrevistas efetuadas com as especialistas do departamento de Controle e Avaliação, se pôde obter uma visão do sistema existente, compreendendo o fluxo de uma AIH, os serviços prestados a pacientes internados pelo SUS, o significados dos dados disponíveis, as análises e revisão nas contas hospitalares. As informações obtidas durante a entrevista realizada com a responsável pelo departamento de Contas Médicas de um hospital de Pelotas, possibilitaram compreender melhor que serviços são agregados aos itens cobrados em uma AIH.

Para a compreensão da patologia AVC e seus tratamentos, foi entrevistado um médico neurologista e, também, houve a compreensão por meio de literatura. Pôde-se, ainda, obter informações bastante elucidativas no manual [SUS98] referente a normas para emissão de AIHs.

Para a determinação dos objetivos ou problemas, foram analisadas as tabelas que compõem o sistema de AIHs e selecionadas as mais relevantes:

- movimento da AIH (DSMS010) – esta tabela contém informações correspondentes às características do paciente (idade, sexo, residência), da internação (número da AIH, hospital, especialidade, procedimento

solicitado e realizado, diagnóstico, data de internação e alta, motivo de cobrança), entre outras;

- valores da AIH (DSMS160) – contém os valores cobrados referentes a cada AIH. Discrimina o número da AIH, o hospital, o procedimento realizado, os valores de serviços hospitalares, serviços profissionais, serviços auxiliares de diagnose/terapia, valor pago por permanência a maior, valor de sangue, recém-nato, valor de UTI, valor de diárias de acompanhante, UTI neo-natal, valor de transplante e valor de neurologia;
- atos médicos (DSMS030) – contém registros correspondentes a cada ato profissional autorizado nas AIHs;
- atos (DAIH050) – contém informações referentes aos atos médicos, com sua identificação, representados pelo código e pela descrição do ato, o etempo de permanência hospitalar, o sexo para o qual o ato pode ser aplicado, as idades mínima e máxima no qual se pode aplicar os atos;
- CID (DAIH150) – contém a descrição dos diagnósticos de acordo com a tabela CID (Classificação Internacional de Doenças);
- procedimentos especiais (DSMS020) – contém registros correspondentes a cada procedimento especial autorizado de AIHs nos municípios no período. Procedimentos especiais são exames a que o paciente é submetido durante sua internação.

As informações contidas nas tabelas referem-se basicamente a dados do faturamento da AIH, que englobam, além dos valores cobrados, os serviços prestados. Dados complementares e importantes para uma aplicação de classificadores, encontram-se somente no prontuário do paciente, no qual o acesso não foi possível.

Com base nas informações disponíveis, foi possível resolver os problemas classificatórios:

- Avaliar o bloqueio de AIHs: visa solucionar o problema do pouco tempo disponível para a realização de revisões nas AIHs. As revisões realizadas buscam identificar se uma AIH deve ou não ser bloqueada por não estar de acordo com as normas estabelecidas pelo SUS. Com base em um conjunto de AIHs caracterizadas como *pagas* e *bloqueadas* é traçado um perfil para estas duas situações, sendo possível a partir destes perfis avaliar outras AIHs e rotulá-las como *liberadas* e *bloqueadas*;
- Avaliar o tipo de internação: objetiva identificar as internações cuja patologia diagnosticada é AVC, e que não se caracterizam como tal, devido ao tipo de internação. Os tipos de internações são caracterizados como *eletiva*, para casos que não apresentam gravidade, e como *urgente*, para casos graves. As internações associadas a AVC devem ser de caráter urgente, devido à gravidade da patologia. Uma internação eletiva, neste caso, é considerada uma "irregularidade". Com base em um conjunto de AIHs, relativas à

AVC, caracterizadas como *eletiva* e *urgente*, é traçado um perfil, sendo possível identificar AIHs "irregulares".

5.2.2 Preparação dos dados

Seguindo a abordagem de [CAB97], a etapa a ser vencida após a definição dos objetivos, é a da preparação dos dados. Esta etapa é desenvolvida com base nos objetivos especificados na etapa anterior. Nela são selecionados os dados que formarão os conjuntos de dados utilizado na mineração e, também, avaliada a qualidade dos dados, sendo realizada sua limpeza, operação quase sempre necessária.

A base de dados utilizada conta com 555 registros de AIHs emitidas a partir dos atendimentos efetuados pelos hospitais de Pelotas aptos a tratar AVC isquêmico, compreendendo o período de janeiro de 2001 a fevereiro de 2002. Os registros referentes a AVC hemorrágico formam descartados por possuírem características distintas do AVC isquêmico, não podendo ser tratados juntos, e por terem ocorrências muito reduzidas.

Com o auxílio dos especialistas a base de dados foi analisada, sendo identificados os dados: considerados importantes, os menos relevantes e os que não deveriam ser utilizados, por conterem informações cuja veracidade é duvidosa.

Selecionados os atributos relevantes das tabelas do sistema de AIHs, foram estudados os seus formatos e seus valores, identificando possíveis alterações necessárias para a obtenção de um modelo de classificação preciso.

Como os dados estavam dispostos em várias tabelas, para a união dos atributos identificados como mais discriminantes, foi necessário utilizar ferramentas de apoio. As ferramentas utilizadas no tratamento dos dados foram o Microsoft Access e o Microsoft Excel.

Foram importadas as tabelas para o Access e estabelecidos relacionamentos entre as tabelas DSMS010, DSMS020, DSMS030 e DSMS160, com a finalidade de obter os atos médicos, os exames e os valores dos serviços prestados correspondentes a cada AIH emitida, sendo todas informações agregadas à tabela DSMS010. Também foi efetuado o relacionamento entre a tabela DSMS010 e as tabelas DSMS050 e DSMS150 para identificar, respectivamente, os procedimentos e atos realizados nas AIHs, pois na DSMS010 somente constam códigos, e para identificar os dados por códigos inicialmente seria difícil.

Com os relacionamentos estabelecidos, foi gerada uma tabela contendo todos os atributos identificados, anteriormente, como mais relevantes. Os atributos do arquivo utilizado na mineração dos dados estão representados na tabela 5.1. O atributo meta utilizado para a resolução do primeiro problema (avaliar o bloqueio de AIHs), denomina-se *situação_aih*, caracterizando as AIHs como *bloqueadas* e *liberadas*. O atributo meta utilizado para a resolução do segundo problema (avaliar o tipo de internação), denomina-se *tipo_internação*, caracterizando as AIHs de acordo com o tipo de internação, sendo *eletiva* e de *urgência*. Os demais atributos que constam na tabela são utilizados como atributos previsores.

TABELA 5.1 – Descrição dos atributos selecionados para o conjunto de mineração

Atributo	Valores
tipo_internação	eletiva; urgência
hospital	A; B; C; D
faixa_etária	A: <20; B: 20-29; C: 30-39; D: 40-49; E: 50-59; F: 60-69; G: 70-79; H: 80-89; I: >89
procedimento_solicitado	avc_agudo; outro_procedimento
óbito	sim; não
permanência	baixa; normal; alta
uti	sim; não
sangue	sim; não
25001019	sim; não
17018030	sim; não
17023041	sim; não
17042046	sim; não
17055040	sim; não
17064040	sim; não
17059046	sim; não
04001010	sim; não
17009065	sim; não
21005060	sim; não
21003068	sim; não
17041040	sim; não
17012015	sim; não
17019044	sim; não
32028040	sim; não
17027039	sim; não
17019036	sim; não
17063043	sim; não
23005025	sim; não
17023033	sim; não
17066042	sim; não
17008018	sim; não
17061040	sim; não
23004029	sim; não
17007011	sim; não
17018048	sim; não
17011043	sim; não
17056047	sim; não
17009049	sim; não
97004057	sim; não
97006009	sim; não
97007005	sim; não
97011002	sim; não
97013013	sim; não
97013021	sim; não
97014001	sim; não

TABELA 5.1 – Descrição dos atributos selecionados para o conjunto de mineração - continuação

Atributo	Tipo / Valores
42004039	sim; não
situação_aih	bloqueada; liberada

A qualidade dos dados é considerada razoável. Os atributos apresentam um considerável nível de ruído. Foram detectados atributos com valores inválidos, pois os valores que constam, segundo o departamento de Controle e Avaliação, são preenchidos sem levar em consideração a realidade. Por exemplo, o atributo *grau de instrução* do paciente e o atributo *número de filhos* são preenchidos com valores que não é possível determinar se reflete a realidade. A utilização destes atributos teve que ser descartada.

Alguns atributos foram tratados a partir da eliminação de redundâncias. Na tabela DSMS010, consta o atributo *data de nascimento* e o atributo *idade do paciente*. O atributo *data de nascimento* foi descartado e selecionado o atributo *idade*. Este possuía um formato *30idade*. O número 30 inserido no início do valor do atributo foi excluído, ficando somente a idade do paciente. Por ser pouco discriminante, o atributo *idade* foi, posteriormente, substituído pelo atributo *faixa_etária*, que indica a faixa etária em que o paciente se enquadra, possuindo os valores: *A*: para idades inferiores à 20 anos; *B*: para idades entre 20 e 29 anos; *C*: para idades entre 30 e 39 anos; *D*: para idades entre 40 e 49 anos; *E*: para idades entre 50 e 59 anos; *F*: para idades entre 60 e 69 anos; *G*: para idades entre 70 e 79 anos; *H*: para idades entre 80 e 89 anos e *I*: para idades iguais ou superiores à 90 anos.

O atributo *motivo de cobrança* é multivalorado. Pelo grande número de valores é difícil a utilização deste atributo na obtenção de padrões. Basicamente, encontram-se duas situações nos valores, o paciente pode obter alta por recuperação ou por óbito. Então, os valores que se referem a óbitos (41, 42, 43, 51, 52 e 53) foram substituídos pelo valor *sim,* e os demais (12, 13, 14, 16, 18, 19, 22 33 e 39) valores substituídos por *não*.

Outro atributo multivalorado, *diagnóstico principal*, possui um total de quatorze valores. Este atributo foi descartado. Com exceção, de um de seus valores, que se destaca, os demais representam um pequeno número de casos, não influenciando na mineração de dados.

O atributo referente ao procedimento realizado foi extraído por ter valor único, *AVC Agudo*, também, não influenciando na mineração. O atributo referente ao procedimento solicitado teve seus valores agregados, passando a ter os valores *avc_agudo* e *outro_procedimento*.

Outros atributos tiveram seus valores modificados: o atributo *permanência* foi definido a partir do tempo de internação dos pacientes, cujos valores: *baixa* refere-se à permanência menor ou igual a quatro dias; o valor *normal* refere-se ao período de cinco a quatorze dias e o valor *alta* refere-se a mais de quatorze dias. O atributo *sangue* indica se o paciente sofreu ou não transfusão de sangue, seus valores foram determinados com base no atributo *val_sangue,* da tabela referentes aos custos da AIH (DSMS160), quando verificados valores diferentes de zero para este atributo, ele recebeu o valor *sim*, para valores iguais a zero, o valor *não*. Os valores do atributo *uti*,

que indica a internação ou não de um paciente na UTI, foram transformados de qualitativos para quantitativos, os valores acima de zero foram substituídos pelo valor *sim* e os iguais a zero pelo valor *não*.

O atributo hospital teve seus valores alterados. Foram substituídos os CGCs, que os identificam, por letras: A, B, C e D. Cada letra representando um hospital

Os atributos da tabela 5.1, que são nomeados por números, correspondem aos atos médicos e procedimentos especiais aplicados aos pacientes. O significado de cada atributo consta no anexo 2. Estes atributos foram gerados a partir das tabelas DSMS020 e DSMS030, associadas à tabela DSMS010. Cada ato médico e procedimento especial correspondem a um registro das duas primeiras tabelas. Com o auxílio do Access, as tabelas foram associadas e, então, gerado um arquivo que contém todos os atos e procedimentos referentes a cada AIH. Com o auxílio da planilha Excel, foi possível transformar os valores correspondentes aos atos e aos procedimentos, em atributos, valorados como *sim* e como *não*, de acordo com sua existência ou não na AIH.

Após a definição final do arquivo de mineração, foi necessário utilizar a planilha eletrônica pertencente à ferramenta do Sipina-W, para gerar os arquivos utilizados pelos algoritmos geradores de árvores de decisão. Os dados gerados para o Sipina-W têm os formatos *.dat*, relativo aos dados, e o formato *.par*, relativo aos parâmetros gerados a partir dos dados.

O arquivo de mineração gerado foi dividido de forma aleatória em dois conjuntos de dados. Para constituir o conjunto de treino foram extraídos 67% dos casos do conjunto exemplos e para a validação do classificador gerado a partir do conjunto de treino, foi definido o conjunto de teste com 33% dos casos, distribuindo, assim, 372 casos para treino e 183 casos para a validação da árvore.

Esta etapa demanda bastante atenção por parte das pessoas envolvidas no processo, sendo, trabalhosa, por exigir muitas análises sobre os dados e por serem tratados nela, muitos detalhes. Todos os requisitos desta etapa devem ser trabalhados cuidadosamente para que na mineração de dados se obtenha resultados válidos.

5.2.3 Mineração dos dados e análise dos resultados obtidos

Como mencionado, a ferramenta Sipina-W foi utilizada como apoio durante a execução da etapa de mineração de dados. Os algoritmos implementados pela ferramenta e utilizados nesta etapa, compreendem o C4.5 e o CART. Estes algoritmos foram os escolhidos por terem, na maioria das vezes, os seus fundamentos utilizados como base para a elaboração de outros algoritmos e, também, por terem grande aceitação entre a comunidade acadêmica.

A partir das árvores geradas pelos algoritmos, obtiveram-se regras correspondentes aos caminhos de classificação apresentados. Com base nas regras, os conhecimentos extraídos foram validados por um conjunto de teste e pela avaliação dos especialistas.

5.2.3.1 Problema: Avaliar o bloqueio e a liberação de AIHs

Para que uma AIH seja paga, é necessário que, ao ser emitida, seus dados estejam de acordo com normas estabelecidas pelo SUS, abrangendo situações que não podem ocorrer, por exemplo, a idade do paciente não estar dentro da idade permitida para o procedimento cobrado. Devido a esta e muitas outras situações, é necessário realizar revisões nas AIHs, sendo verificado o motivo de tais ocorrências, liberando ou não seus pagamentos.

A partir do problema da avaliação das AIHs, deseja-se obter um classificador binário, com o atributo *situacao_aih* como: meta. Para este problema, a árvore de decisão é construída para gerar padrões que distingam entre *liberada*, a AIH que não necessita ser revisada, e *bloqueada*, a AIH que necessita ser avaliada detalhadamente e determinada sua rejeição ou não.

Foram criados dois modelos classificatórios. O primeiro modelo foi construído com a aplicação do algoritmo C4.5 ao conjunto de treino e, o outro, com a aplicação do algoritmo CART. A partir dos modelos criados obteve-se padrões para caracterizar as AIHs como: *liberadas* ou *bloqueadas*.

5.2.3.1.1 Algoritmos C4.5

A árvore de decisão gerada pelo algoritmo C4.5 é ilustrada na figura 5.2. Ela apresentou uma taxa de precisão de 100% ao classificar os exemplos do conjunto de treino, gerando a matriz de classificação apresentada na tabela 5.2.

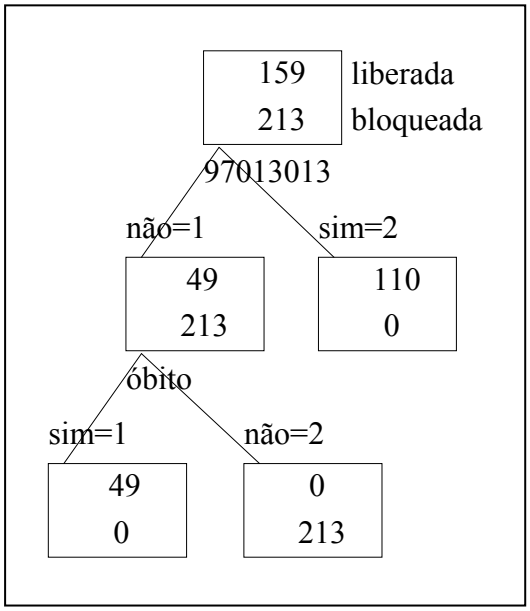

FIGURA 5.2 – Árvore de decisão gerada pelo algoritmo C4.5 para o problema: avaliar o bloqueio de AIHs

TABELA 5.2 – Matriz de confusão do algoritmo C4.5 para o problema: avaliar o bloqueio de AIHs

	Liberada	Bloqueada	Não classificado
Liberada	159	0	0
Bloqueada	0	213	0
Total	159	213	0

Três regras foram extraídas da árvore de decisão, determinando padrões de liberação e de bloqueio. São elas:

- Se o paciente não realizou o exame Tomografia Computadorizada de Crânio (97013013) e teve óbito então a AIH é *liberada* (49 casos – 100% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de Crânio (97013013) e teve sobrevida então a AIH é *bloqueada* (213 casos – 100% precisão);
- Se o paciente realizou o exame Tomografia Computadorizada de Crânio (97013013) então a AIH é *liberada* (110 casos – 100% precisão).

A validação do modelo gerado foi efetuada com base no conjunto de teste definido. Pôde-se verificar pela matriz de confusão (tabela 5.3), obtida durante a validação, uma precisão de 100% para os casos avaliados, significando que todos os casos foram classificados com sucesso, tornando o modelo aplicável a novos casos.

TABELA 5.3 – Matriz de confusão do C4.5 para o problema: avaliar o bloqueio de AIHs

	Liberada	Bloqueada	Não classificado
Liberada	116	0	0
Bloqueada	0	67	0
Total	116	67	0

O conhecimento obtido com a aplicação do algoritmo revelou um padrão de bloqueios de AIHs. Foi descoberto que dois fatores são de fundamental importância no bloqueio ou liberação de uma AIH, o fator exame de Tomografia Computadorizada de Crânio e fator óbito.

5.2.3.1.2 Algoritmo CART

Utilizando o mesmo conjunto de treino aplicado ao algoritmo C4.5, a árvore gerada pelo algoritmo CART (figura 5.3) obteve precisão de 86%, dando origem a matriz de classificação mostrada na tabela 5.4. Trinta e nove casos que no conjunto de treino estão definidos como bloqueados, foram classificados como liberados.

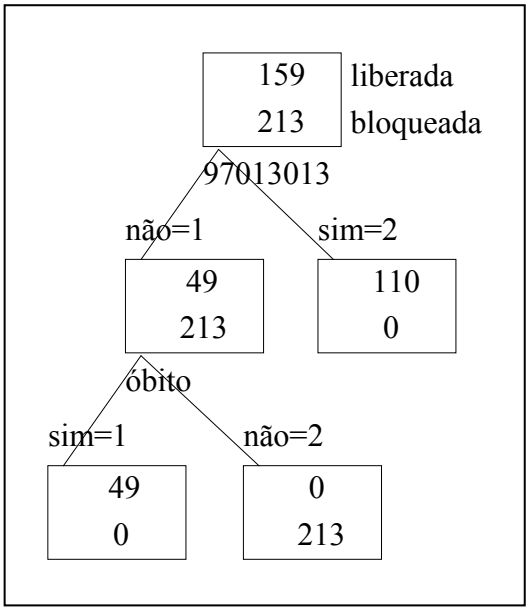

FIGURA 5.3 – Árvore de decisão gerada pelo algoritmo CART para o problema: avaliar o bloqueio de AIHs

TABELA 5.4 – Matriz de confusão do algoritmo CART para o problema: avaliar o bloqueio de AIHs

	Liberada	Bloqueada	Não classificado
Liberada	159	0	0
Bloqueada	0	213	0
Total	159	213	0

A partir da árvore do gerada pelo CART, obtêm-se as mesmas regras que com o algoritmo C4.5.

A validação do modelo gerado pela árvore, também foi efetuada com base no conjunto de teste aplicado ao algoritmo C4.5. Pôde-se verificar pela matriz de confusão (tabela 5.5) gerada durante a validação, boa precisão, indicando que 100% dos casos avaliados foram classificados com sucesso.

TABELA 5.5 – Matriz de confusão do CART para o problema: avaliar o bloqueio de AIHs

	Liberada	Bloqueada	Não classificado
Liberada	67	0	0
Bloqueada	0	116	0
Total	67	116	0

A árvore gerada pelo algoritmo CART não revelou conhecimentos além dos já apresentados pela árvore gerada pelo algoritmo C4.5. Constatou-se os mesmos fatores de bloqueio e liberação de AIHs.

5.2.3.2 Problema: Avaliar o tipo de internação

O quadro clínico de um paciente que é internado pelo procedimento AVC Agudo, é considerado muito grave e o surgimento do quadro é súbito. Desta forma, a internação deste paciente é de urgência ou de emergência. Uma internação eletiva não caracteriza um quadro de AVC, não sendo possível este tipo de internação ser atribuída a uma AIH cujo procedimento realizado é AVC Agudo. Pelas análises efetuadas na base de dados, constatou-se que são registradas hospitalizações do tipo eletiva, caracterizando uma "irregularidade" no preenchimento das AIHs.

Com base nesta problemática, procurou-se estabelecer padrões para os dois tipos de internações, aplicando aos dados os algoritmos C4.5 e CART e identificando as AIHs que se enquadram dentro de *eletiva* ou *urgente*. Para a resolução deste problema, foi utilizado como atributo meta o tipo_internação.

5.2.3.2.1 Algoritmo C4.5

Os mesmos conjuntos de treino e de teste utilizados no problema anterior foram utilizados para a resolução do problema: *avaliar o tipo de internação*. A aplicação do algoritmo C4.5 aos dados do conjunto de treino, gerou a árvore de decisão apresentada na figura 5.4.

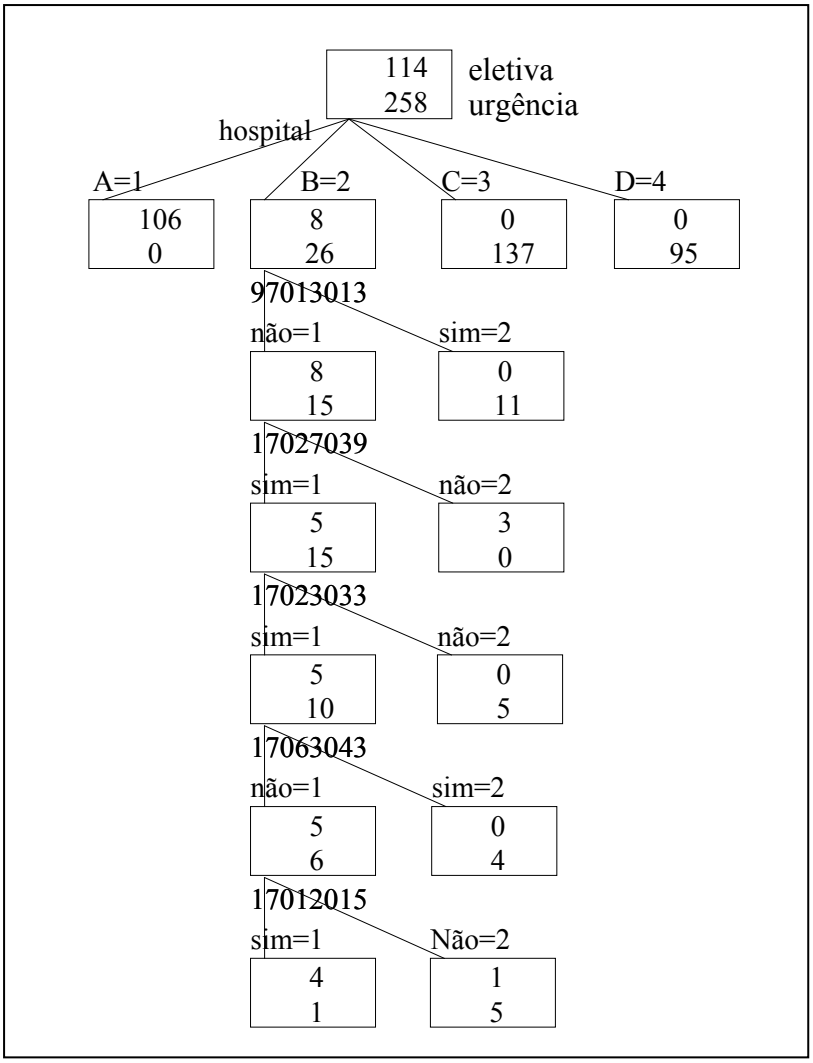

FIGURA 5.4 – Árvore de decisão gerada pelo algoritmo C4.5 para o problema: avaliar o tipo de internação

A matriz de classificação obtida a partir da árvore de decisão gerada pelo C4.5, é apresentada na tabela 5.6. O algoritmo classificou erroneamente dois casos do conjunto de treino. Um caso cuja internação é definida como eletiva, foi classificada como urgência, e um caso que possui internação de urgência, foi classificado como eletiva.

TABELA 5.6 – Matriz de confusão do algoritmo C4.5 para o problema: avaliar o tipo de internação

	Eletiva	Urgência	Não classificado
Eletiva	113	1	0
Urgência	1	257	0
Total	114	258	0

Nove regras foram extraídas da árvore de decisão, determinando padrões de internação para as AIHs. São elas:

- Se o hospital é o A então o tipo de internação é *eletiva* (106 casos 100% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de crânio (97013013) e não realizou o exame de triglicerídios (17063043) e realizou os exame de protombina (17027039) e de contagem de plaquetas (17023033) e de urina (17012015) e o hospital é o B então o tipo de internação é *eletiva* (5 casos – 80% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de crânio (97013013) e não realizou de protombina (17027039) e o hospital é o B então o tipo de internação é *eletiva* (3 casos – 100% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de crânio (97013013) e não realizou exame de triglicerídios (17063043) e efetuou os exames de protombina (17027039) e de contagem de plaquetas (17023033) e não realizou exame de urina (17012015) e o hospital é o B então o tipo de internação é *urgência* (6 casos – 83% precisão);
- Se o hospital é o C então o tipo de internação é *urgência* (137 casos – 100% precisão);
- Se o hospital é o D então o tipo de internação é *urgência* (95 casos – 100% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de crânio (97013013) e realizou o exame de protombina (17027039) e não realizou o exame de contagem de plaquetas (17023033) e o hospital é o B então o tipo de internação é *urgência* (5 casos – 100% precisão);
- Se o paciente realizou o exame Tomografia Computadorizada de crânio (97013013) e o hospital é o B então o tipo de internação é *urgência* (11 casos – 100% precisão);
- Se o paciente não realizou o exame Tomografia Computadorizada de crânio (97013013) e realizou exames de triglicerídios (17063043) e de protombina (17027039) e de contagem de plaquetas (17023033) e o hospital é o B então o tipo de internação é *urgência* (11 casos – 100% precisão).

Os conhecimentos obtidos para o problema relacionado aos tipos de internações, referem-se: a cobrança de internação relacionadas a patologia AVC, que muitas vezes é realizada por internações do tipo eletiva, o correto para este tipo de patologia são internações do tipo urgência; aos hospitais que cobram pelo tipo de internação eletiva de forma "equivocada".

A validação do modelo gerado foi efetuada com base no conjunto de teste definido. Verificou-se pela matriz de confusão (tabela 5.7) gerada durante a validação, uma precisão de 92%, indicando o percentual de casos que foram classificados com sucesso.

TABELA 5.7 – Matriz de confusão do C4.5 para o problema: avaliar o tipo de internação

	Eletiva	Urgência	Não classificado
Eletiva	52	0	0
Urgência	14	117	0
Total	66	117	0

5.2.3.2.2 Algoritmos CART

Aplicando o algoritmo CART aos dados do conjunto de treino, foi gerada a árvore de decisão apresentada na figura 5.5. Nela, são mostrados os padrões encontrados pelo CART para o problema da avaliação das internações.

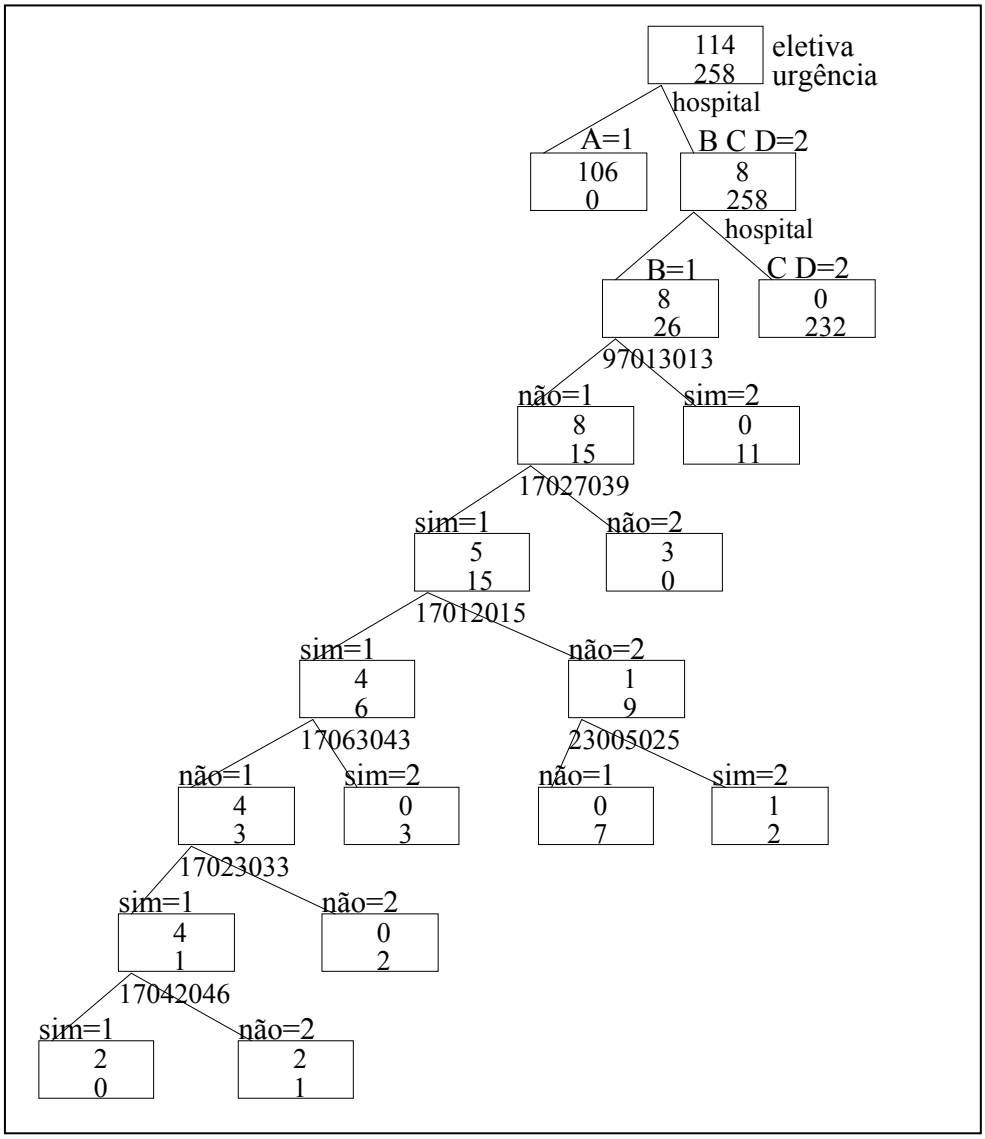

FIGURA 5.5 – Árvore de decisão gerada pelo algoritmo CART para o problema: avaliar o tipo de internação

TABELA 5.8 – Matriz de confusão do algoritmo CART para o problema: avaliar o tipo de internação

	Eletiva	Urgência	Não classificado
Eletiva	113	1	0
Urgência	1	257	0
Total	114	258	0

O resultado da validação do modelo gerado é apresentado pela matriz de confusão da tabela 5.9. Dos casos avaliados, 92% foram classificados com sucesso, indicando uma ótima precisão.

TABELA 5.9 – Matriz de confusão do CART para o problema: avaliar o tipo de internações

	Eletiva	Urgência	Não classificado
Eletiva	52	0	0
Urgência	14	117	0
Total	66	117	0

O principal conhecimento obtido após a aplicação do algoritmo CART aos dados relacionado ao problema em questão, foi quanto aos hospitais que cobram por internações eletivas para a patologia AVC, como citado, esta patologia não admite este tipo de internação.

5.3 Análise do Processo de DCBD

Nos próximos itens são feitas considerações referentes à experiência obtida a partir do processo de DCBD tratado anteriormente. A proposta deste trabalho em efetivar o processo de DCBD e extrair padrões de dados associados a um domínio de administração hospitalar, mediante a aplicação de algoritmos de árvores de decisão, utilizou como referencial os critérios sugeridos por [CAB97], de forma geral muito semelhantes a propostas de autores como [ADR96], [BRA96] e [FAY96].

Como indicado pela proposta de [CAB97], grande parte dos esforços realizados durante o processo ocorrem durante a etapa de preparação dos dados, quando são selecionados os dados necessários ao alcance dos objetivos pré-determinados, o pré-processamento dos dados, na busca e correção de ruídos, muito freqüentemente encontrados nas bases de dados, e também a fase de transformação dos dados, na adaptação de seus formatos ao formato algoritmo de mineração a ser empregado, sendo o tempo destinado à aplicação dos algoritmos de mineração menor pois, já se tendo todos os dados tratados, a aplicação do algoritmo não leva tanto tempo.

Pôde-se verificar a interatividade e a iteratividade do processo, uma vez que foi necessário retornar várias vezes a etapas anteriores. Como ressaltado por [BRA96], o

papel das pessoas envolvidas no processo, em especial o analista, é fundamental para o desenvolvimento do processo. É indispensável uma grande interação entre o analista e a base de dados em questão, sendo imprescindível o uso de ferramentas que apóiem esta interação, como as de consulta, as de visualização, as estatísticas e de Inteligência Artificial e as de apresentação. O analista deve interagir com as demais pessoas envolvidas no processo, para que ao seu término, possam ser apresentados bons resultados.

5.3.1 As etapas de compreensão do domínio e determinação dos objetivos

Como mencionado, a abordagem adotada como referencial não inclui uma etapa dedicada à compreensão do domínio. Sua importância é indicada na etapa de determinação dos objetivos, indicando que a compreensão do domínio tem uma relação direta com o objetivo definido para o processo.

Para um objetivo claro e viável para o desenvolvimento do processo é fundamental obter um bom entendimento do domínio, pois um erro na definição dos objetivos, cuja finalidade é guiar o processo, levará à obtenção de resultados pouco confiáveis.

Para a determinação do objetivo do processo de descoberta de conhecimento em AIHs, além da interação com especialistas, foi essencial a interação com o próprio banco de dados para chegamos a um objetivo possível de ser alcançado, isto é, sem um bom conhecimentos dos dados existentes seria muito difícil obter, ao final do processo, resultados válidos.

Com o conhecimento dos dados das AIHs, inicialmente parecia um pouco difícil estabelecer padrões entre os dados que pudessem ser classificados, pois as informações que constam nas tabelas existentes não são muito discriminantes para tal aplicação. Mas, no decorrer de análises efetuadas, surgiram problemas classificatórios, para os quais foi possível a construção de um modelo de classificação.

Dentro desta etapa de compreensão e determinação dos objetivos, as interações com os especialistas foram fundamentais para o andamento do processo. A cada entrevista efetuada eram esclarecidas dúvidas sobre o sistema de AIHs e sobre os dados. Durante o estudo dos dados para a definição do objetivo, as entrevistas auxiliaram na compreensão dos dados, pois muitos deles possuem valores agregados que puderam ser mostrados.

5.3.2 A etapa de preparação de dados

Em [CAB97], após a definição do objetivo, deve-ser realizar a preparação dos dados. Nesta etapa, o objetivo estabelecido guia a formação de um conjunto de dados a ser utilizado nas próximas etapas. Este conjunto pode ter alguns problemas quanto à qualidade, sendo necessária à realização de uma limpeza nos dados do conjunto.

Os dados coletados estavam dispostos em várias tabelas. Utilizando como ferramenta de apoio o Access e o Excel, foram estabelecidos relacionamentos entre elas e gerada uma tabela única com os atributos julgados importantes para, a partir dos dados, ser possível encontrar padrões que alcançassem o objetivo determinado.

Com as tabelas do sistema de AIHs contendo dados discriminantes do faturamento de cada AIH, pôde-se contar apenas com aqueles ligados aos atos médicos e procedimentos especiais realizados, os valores cobrados pela realização destes atos e procedimentos, assim como valores referentes à UTI, a transfusões de sangue, à permanência do paciente. Contou-se também com poucas informações cadastrais do paciente. Disponíveis somente os atributos *idade*, *sexo*, *cidade* e *cep*. Informações como grau de instrução e número de filhos não puderam ser utilizadas devido à baixa qualidade da informação.

A qualidade dos dados pode ser definida como razoável. Basicamente, o que foi detectado de ruído, foram valores inseridos com erro de digitação e alguns atributos com valores que não podem ser considerados para a análise pois, segundo a responsável pelo departamento de Controle e Avaliação, eles não são confiáveis.

Muitos atributos tiveram que ser tratados por serem multivalorados, situação que compromete o desempenho do modelo criado. Os atos médicos, assim como, os procedimentos especiais, constam como valores nas tabelas DSMS030 e DSMS020 respectivamente. Além das associações com a tabela de movimento de AIHs (DSMS010), foi necessário transformar estes valores em atributos. Com o auxílio do Excel foi possível realizar esta transformação bastante trabalhosa.

Os dados não necessitaram sofrer transformações em seu formato, tanto as ferramentas utilizadas como apoio na seleção e pré-processamento dos dados, quanto a ferramenta de mineração Sipina-W, trabalham com o formato (xls), não ocasionando problemas neste sentido.

Como mencionado em [CAB97], esta etapa tomou grande parte do tempo gasto no processo, inicialmente pelo relacionamento das tabelas, a seleção dos procedimentos relativos à patologia AVC e pela criação de novos atributos e agregação de outros, necessários ao alcance do objetivo estabelecido. Foram também indispensáveis várias iterações nesta etapa, para realizar ajustes no conjunto de dados que estava sendo utilizado na etapa de mineração.

5.3.3 As etapas de mineração de dados e de análise dos resultados

Na etapa de mineração de dados, puderam ser constatados os resultados decorrentes da aplicação dos algoritmos C4.5 e CART em problemas reais. Os classificadores gerados pelos algoritmos enquadraram corretamente quase todos os casos do conjunto de teste, indicando que, quando aplicado a novos casos realizam uma classificação confiável. A qualidade dos classificadores também foi avaliada pelas especialistas.

As regras extraídas das árvores geradas para o problema "avaliar o bloqueio de AIHs", são simples, mas eficazes. As especialistas constataram que estas regras podem ser aplicadas corretamente à análise das AIHs, sendo utilizadas na prática.

Elas avaliaram que: se o paciente realiza exame de Tomografia Computadorizada de Crânio, é um bom indicativo que o procedimento realizado corresponde à realidade; se o paciente não realiza este exame e morre, é justificável a liberação da AIH pela gravidade do caso. Para o paciente que não realizou Tomografia Computadorizada de Crânio e sobreviveu a ocorrência do AVC, a AIH deve ser bloqueada, assim como apresentaram as regra, pois esta situação não justifica a cobrança pelo procedimento AVC Agudo.

As regras extraídas a partir da árvore de decisão gerada para o problema "avaliar tipo de internação", também foram consideradas válidas. As especialistas confirmaram que as internações provenientes de AVC devem ser sempre de urgência, não havendo internações eletivas (programadas), visto que se trata de um quadro clínico de surgimento súbito e imprevisto. Foi confirmado que AVC com internação eletiva, sugere casos de pacientes com seqüelas de AVC prévio e atualmente com uma intercorrência clínica, não devendo ser cobrado pelo procedimento AVC Agudo. O conhecimento referente aos hospitais que cobram erroneamente pelo tipo de internação, não era de conhecimento da SMSP. A partir deste conhecimento, constatou-se a necessidade de investigar as internações decorrentes de AVC.

O retorno à etapa de preparação de dados foi freqüente, a cada árvore gerada e avaliada nos vários testes efetuados, eram necessários novos ajustes nos dados. Também imprescindível a realização de novas entrevistas durante a realização das etapas, pois sempre surgiam dúvidas em relação aos dados e também para as especialistas avaliarem o conhecimento obtido a cada teste.

5.3.4 O ciclo do processo

Como mencionado, sem o auxílio dos especialistas, esta aplicação não teria como ser desenvolvida. Pelas peculiaridades das AIHs, em especial as originadas do procedimento AVC agudo, somente pessoas ligadas à área da saúde tem condições de compreendê-las.

Antes da extração do conhecimento considerado válido, foram necessárias várias iterações envolvendo as etapas de compreensão e preparação, abrangendo a seleção, o pré-processamento e a transformação dos dados, sendo realizado os ajustes necessários à conclusão do processo.

O ciclo realizado durante o processo seguiu a mesma ordem do abordado por [CAB97]. O processo somente foi finalizado quando cada modelo obtido foi considerado válido para o objetivo definido.

6 Conclusões e Trabalhos Futuros

Neste trabalho foi realizado um experimento com o uso de árvores de decisão, tendo em vista a descoberta de conhecimento em base de dados associadas à área da saúde. A base de dados utilizada referente-se às informações pertinentes as AIHs emitidas a partir dos serviços prestados pelos hospitais da cidade de Pelotas conveniados ao SUS. O trabalho não se deteve somente ao estudo de árvores de decisão, também foi explorado todo o processo de DCBD, tendo, como referencial, a abordagem de [CAB97]. Embora o trabalho contemple esta abordagem, ele também descreve uma visão das abordagens de [ADR96], [BRA96] e [FAY96].

O objetivo principal do trabalho é explorar os dados pertencentes a uma situação real, Autorizações de Internações Hospitalares, e buscar extrair padrões mediante a aplicação de árvores de decisão, auxiliando, assim, a Secretaria Municipal de Saúde de Pelotas na avaliação e controle das autorizações.

O processo de DCBD foi relatado na íntegra, visto que, em grande parte da literatura sobre o tema, não se encontra uma descrição mais detalhada de todo o processo. As situações apresentadas ao longo do desenvolvimento do experimento, foram enquadradas de acordo com o exposto no referencial tomado.

A proposta de [CAB97] tomada como referencial, dispõe o processo em etapas organizadas de forma seqüencial, cuja iteração entre as etapas e a interação das pessoas envolvidas no processo fazem-se necessárias. A situação foi comprovada na prática, uma vez que durante o desenvolvimento do processo de DCBD, foram necessárias várias iterações a etapas anteriores e, também, diversas consultas aos especialistas, com o intuito de compreender o domínio trabalhado e os dados disponíveis referentes ao sistema de AIHs.

Para a obtenção de classificadores baseados em árvores de decisão a partir das AIHs, muitas análises e estudos foram realizados para a compreensão da rotina de uma AIH, desde sua emissão até seu pagamento ou não. Igualmente, foram estudadas as situações diversas que podem ocorrer em relação a um único procedimento pois, na área da saúde, podemos consideram como nada sendo certo ou fixo para uma patologia, podendo os atos médicos realizados variarem muito de caso para caso. Em vista dessa situação e da falta de maiores detalhes sobre pacientes e patologias, a determinação de padrões entre os dados foi uma tarefa bastante trabalhosa.

Longo tempo foi gasto em testes, na tentativa de encontrar situações-problema que pudessem ser enquadradas dentro de padrões e, então, obter conhecimento por meio da aplicação de árvores de decisão. Devido à grande quantidade de dados disponíveis nas AIHs, foi necessário delimitar o problema, sendo selecionado, como foco de análise, a patologia AVC e as informações disponíveis associadas a ela.

Na bibliografia consultada não foi encontrado nenhum exemplo de aplicação de árvores de decisão à administração hospitalar. Comumente são relatadas aplicações de árvores de decisão que giram em torno da determinação do diagnóstico a partir de sintomas apresentados, da determinação da causas de óbitos, de prognósticos, etc.

Uma dificuldade sentida foi quanto aos dados não possuírem muitas informações relativas aos pacientes e aos sintomas apresentados por eles. Tais informações constam somente nos prontuários dos pacientes, que não puderam ser cedidos para análise por serem personalizados e sigilosos.

Com base nos dados disponíveis foram definidos os objetivos do processo de descoberta de conhecimento. Buscou-se padrões capazes de determinar situações que levam as AIHs, referentes à AVC, serem ou não separadas para revisão. A partir de uma conferência que ocorre manualmente, foram estabelecidos padrões que determinam quando uma AIH deve ser liberada ou bloqueada. Outro padrão que pode ser obtido a partir das informações disponíveis, foi quanto à identificação de internações "irregulares", caracterizando o tipo de internação de uma AIH como eletiva ou como de urgência.

Os padrões obtidos foram julgados válidos pelas especialistas da SMSP. Os padrões relativos as AIHs bloqueadas ou liberadas se aplicam corretamente à análise das AIHs, e os padrões relativos aos tipos de internações mostraram que hospitais realizam cobranças indevidas.

Verificou-se que os resultados apresentados pelos algoritmos C4.5 e CART puderam ser confrontados com a prática e apresentaram resultados válidos, e sua aplicação pode ser utilizada para novos casos obtendo resultados confiáveis.

A aplicação dos algoritmos aos dados, foi realizada através da ferramenta Sipina-W. A partir desta ferramenta, foram geradas as árvores de decisão e também extraídas regras correspondentes a cada caminho de classificação, sendo utilizadas na validação das árvores.

A conclusão do estudo sobre as AIHs é parcial, uma vez que não foi possível o acesso aos prontuários. Com as informações que constam neles, o leque de opções para realizar um trabalho mais completo seria, sem dúvida, muito maior, podendo-se descobrir, quem sabe, informações muito interessantes, que auxiliariam a SMSP a ter uma visão melhor dos pacientes atendidos pelos hospitais conveniados ao SUS.

Finalmente, buscou-se fazer um relato detalhado do processo de descoberta de conhecimento, descrevendo todas as etapas que o compõem, visando que pessoas interessadas em estudar o desenvolvimento do processo, assim como os classificadores baseados em árvores de decisão, tenham um referencial um pouco mais detalhado que o normalmente encontrado na literatura. As diversas vantagens do uso de árvores de decisão reforçam a importância deste método de extração de conhecimento e espera-se que este estudo seja um incentivo à elaboração de novos trabalhos. Sugere-se, que prontuários médicos sejam disponibilizados de forma não personalizada e digitalizados, possibilitando, assim, sua utilização para refinar a prática da administração hospitalar, proporcionando o desenvolvimento de novos trabalhos nesta área.

Anexo 1 Regras Estabelecidas pelo SUS para Bloqueio de AIHs

1. *AIH com mais 84 atos profissionais* - o limite de cobrança de atos profissionais por AIH é de 84, ultrapassado este total a AIH é rejeitada.
2. *AIH iniciais em AIH de continuação* - reapresentação da AIH inicial em vez de apresentação da AIH de continuação em caso de psiquiatria e/ou FPT.
3. *Procedimento autorizado não cadastrado* - lançamento de procedimento não cadastrado no campo médico auditor.
4. *AIH de continuação sem AIH inicial* - apresentação de AIH 5 sem a inicial (AIH-1) em FPT e/ou psiquiatria.
5. *Procedimento solicitado não consta da tabela de procedimentos* - lançamento do código do procedimento no campo "procedimento solicitado" inexistente na tabela.
6. *Procedimento solicitado errado* - lançamento de código com digito errado, apesar de constar na tabela de procedimentos.
7. *TOT UTI + Acompanhante > que o período de internação* - totalização de diárias de UTI e de acompanhante maior que os dias de internação do paciente.
8. *Falta CGC/CPF do profissional* - AIH sem lançamento de CGC/CPF no campo serviços profissionais.
9. *CPF/CGC errado* - lançamento de CPF/CGC com dígito errado no campo serviços profissionais.
10. *Falta ato profissional* - campo ato profissional sem lançamento.
11. *Ato profissional errado* - no campo ato profissional o código do procedimento deve ser preenchido com 8 dígitos, neste caso há lançamento errado de digito.
12. *Tipo de ato errado* - o código de ato profissional é composto de 02 dígitos, conforme tabela constante no manual AIH, neste caso há lançamento de código ou n.º de dígitos errados. (ex. cirurgião ou obstetra 01; primeiro auxiliar 02; anestesista 06; fisioterapia 11).
13. *Quantidade de ato errado* - "campo quantidade de ato" incompatível com os atos executados.
14. *Falta CPF Diretor Clínico* - AIH sem CPF do diretor clínico do hospital.
15. *Procedimento realizado diferente do solicitado* - o código do procedimento realizado tem que ser o mesmo do procedimento solicitado ou da 1ª linha do campo médico auditor, neste caso o lançamento não coincide com nenhum dos dois códigos possíveis.
16. *Falta procedimento realizado* - campo procedimento realizado sem preenchimento.

17. *Procedimento realizado incompatível com especialidade* - lançamento do procedimento realizado não condiz com especialidade informada.
18. *Procedimento realizado incompatível com idade/sexo* - procedimento realizado incompatível com a idade ou sexo do paciente.
19. *Data da internação errada* - data de internação incompatível com a data da alta.
20. *Data da saída > data da apresentação* - data da alta posterior a apresentação da AIH.
21. *Permanência meio de mês em especialidade 4 ou 5* - será motivo de rejeição o não lançamento da data de saída em AIH de psiquiatria e/ou FPT no último dia de cada mês quando o paciente permanecer internado.
22. *Ato profissional não cadastrado* - lançamento de ato profissional inexistente.
23. *Profissional bloqueado e/ou com conta corrente 1.9* - lançamento de CGC/CPF bloqueado ou com conta corrente inexistente (conta vala).
24. *Procedimento realizado não consta na tabela de procedimentos* - lançamento de código do procedimento no campo "procedimento realizado" inexistente na tabela.
25. *Hospital não cadastrado na especialidade* - hospital não apresenta leitos cadastrados na especialidade lançada na AI H.
26. *Longa permanência para AIH não apresentada* - cobrança de longa permanência (AIH-5) sem AIH inicial.
27. *Cobrança UTI indevida hospital sem leitos* - cobrança de diária de UTI em hospitais que não apresentam leitos de UTI cadastrados.
28. *AIH paga em outro processamento* - reapresentação de AIH paga em processamento anterior.
29. *AIH paga neste processamento* - apresentação de DCIH (Documento de Cobrança de Internação Hospitalar) em duplicidade.
30. *Permanência a maior superior ao permitido* - cobrança de diária de permanência a maior superior aos dias permitidos.
31. *Cobrança indevida de permanência à maior* - cobrança de permanência à maior em procedimentos nos quais não é permitida sua cobrança.
32. *Código não pode ser ato profissional* - lançamento de código incompatível com os códigos da tabela de atos profissionais.
33. *Somente hospital pode ser tipo 3* - lançamento de CPF como tipo 3 (exclusivo para hospitais)
34. *Psiquiatria só pode ato grupo 63* - os procedimentos psiquiátricos são todos dos grupos 63, neste caso houve cobrança na especialidade psiquiatria com códigos diferente de 63.
35. *Este código não pode mudar procedimento* - procedimento solicitado não admite mudança de procedimento.

36. *AIH extraviada, enviar AIH para FNS/DATASUS/RIO* - AIH em que foi solicitado cancelamento por motivo de extravio e/ou inutilização.

37. *CGC hospital da AIH-1 diferente CGC da AIH-5* - lançamento errado de CGC do hospital em AIH de continuação.12- Especialidade da AIH-5 diferente da AIH-1 - cobrança de especialidade diferente de psiquiatria e/ou FPT para AIH de continuação.

38. *Data da internação da AIH-5 diferente da AIH-1* - a data da internação na AIH tem que ser a mesma da internação constante na AIH inicial.

39. *Número de AIH fora do limite* - série de AIH apresentada maior que a liberada para processamento.

40. *Falta de procedimento aut. p/ 31.000.00.2, 39.000.00.1 e 70.000.00.0* - não lançamento no campo médico auditor dos procedimentos autorizados em cirurgia múltipla, politraumatizados e AIDS.

41. *Profissional não cadastrado* - lançamento de CPF de profissional não cadastrado no SIH-SUS.

42. *AIH com série numérica bloqueada* - AIH em que foi solicitado bloqueio/cancelamento.

43. *Hospital PUB/HUE não permite tipo 6, 7 e 8* - hospitais públicos e universitários não permitem cobrança de profissionais ou SADT sem vinculo.

44. *Transplante renal p/hospital não cadastrado SIPAC-RIM* - cobrança de transplante renal para hospital não autorizado para realização de procedimento de alta complexidade - rim.

45. *Tipo incompatível com tipo de ato* - lançamento no campo tipo incompatível com o tipo de ato preenchido (ex.: tipo 1 (OPM) tipo de ato 12 (Hemoterapia)).

46. *Número de aplicação de nutrição parenteral > dias internação* - quantidade de nutrição parenteral cobrada superior ao permitido para os dias de internação.

47. *Código ROPM não cadastrado* - código de ROPM não cadastrado no SIH-SUS, ou seja inexistente.

48. *Data internação anterior a 01.08.90 para hospitais públicos* - os hospitais públicos foram cadastrados no SIH-SUS a partir de 01.08.90, portanto não podem apresentar AIH anteriores a essa data.

49. *Material incompatível c/procedimento realizado* - material de OPM cobrado incompatível com procedimento realizado.

50. *Quantidade material superior ao permitido* - quantidade de material de OPM cobrado, superior ao permitido conforme o constante na tabela de OPM.

51. *Ato Profissional incompatível c/tipo* - lançamento no campo ato profissional incompatível com o tipo de ato preenchido.

52. *CGC não é banco de sangue* - CGC lançado não cadastrado como SADT - Banco de Sangue.

53. *Hosp. não cadastrado no SIPAC-CÂNCER* - lançamento de procedimento de alta complexidade em câncer para hospital não autorizado.

54. *AIH fora da faixa* - apresentação de AIH com faixa numérica de competência subsequente.

55. *Serv. Prof. com campo zerado* - não lançamento de serviços profissionais na AIH.

56. *AIH PSQ//FPT com prazo vencido* - AIH de psiq./fpt com mais de 107 dias.

57. *TMO não permite cobrança de OPM/hemoterapia* - no procedimento TMO, já estão incluídos os valores de OPM e Hemoterapia, portanto seu lançamento caracteriza cobrança em duplicidade, rejeitando a AIH.

58. *Hospital não cadastrado no SIPAC-AIDS* - cobrança de tratamento de AIDS em hospital não autorizado.

59. *Taxa de ocupação UTI > limite* - cobrança de diárias UTI maior que 100% de ocupação dos leitos cadastrados.

60. *Data internação anterior a 01/07/92 AIDS* - cobrança anterior a introdução do procedimento no SIH-SUS.

61. *Ano/mês alta maior que competência do Pagamento* - data de alta lançada na AIH superior a competência da apresentação para pagamento.

62. *Procedimento especial excede limite (Quadro Procedimentos Especiais)* - cobrança de procedimentos superior ao limite fixado por AIH.

63. *Cirurgia múltipla - cobrança indevida* - cobrança de atos cirúrgicos em duplicidade ou incompatíveis na mesma AIH.

64. *Estudo Eletrofisiológico não autorizado* - cobrança de estudo eletrofisiológico por hospital não autorizado para realização do procedimento.

65. *Hospital não é tipo IX (psiquiatria)* - cobrança de psiquiatria III para hospital não autorizado com tipo IX (conforme PT 408/92).

66. *Implante dentário não autorizado* - cobrança de implante dentário por hospital não autorizado pela alta complexidade para realização do procedimento.

67. *Hospital não cadastrado p/ transplante de fígado* - cobrança de transplante de fígado por hospital não autorizado pela alta complexidade para realização do procedimento.

68. *Hospital não cadastrado p/transplante de pulmão* - cobrança de transplante de pulmão por hospital não autorizado pela alta complexidade para realização do procedimento.

69. *Uso indevido de material de OPM* - cobrança de material em procedimento em que não é compatível o uso de OPM.

70. *Hospital não cad. psiquiatria IV* - cobrança do procedimento tratamento em psiquiatria em hosp. psiquiátrico B (psiq. IV) por hospital não autorizado.

71. *Hospital não cad. procedimento lábio palatais* - cobrança de procedimentos constantes da PT/S.A.S./MS 126/93 (lábio palatais), por hospital não autorizado pela alta complexidade.

72. *Tratamento psiquiatria hospital dia/geral não permite AIH-5* - hospital dia e hospital geral de psiquiatria não permite cobrança em AIH de continuação.

73. *PSQ/FPT não admite permanência à maior* - psiq./FPT permite somente cobrança em AIH de longa permanência, não permitindo cobrança de permanência à maior.

74. *Proc. incompatível com tipo do SIPAC-CV* - a alta complexidade em cardiologia é subdividida em tipos 1,2,3 (um implante de marca-passo; dois - implante de marca-passo e cirurgia cardíaca; três - implante de marca-passo, cirurgia cardíaco e estudo eletrofisiológico nestes casos, o hospital apresentou cobrança no tipo em que não está habilitado.

75. *Trat. epilepsia só permite tipo ¾* - somente hospitais com profissionais e SADT vinculados ao hospital poderão efetuar cobrança dos procedimento de tratamento da epilepsia pertencentes a alta complexidade.

76. *Epilepsia não permite UTI/auditor/sangue/OPM* - nos procedimentos de tratamento de epilepsia pertencentes a alta complexidade estão incluídos no valor as diárias de UTI, sangue, OPM, não sendo também permitida nenhuma cobrança no campo médico auditor.

77. *Hospital não cadastrado no SIPAC-Epilepsia* - cobrança de procedimento de alta complexidade em epilepsia por hospital não autorizado.

78. *Hospital não cadastrado no SIPAC-ORTO* - cobrança de procedimento de alta complexidade em ortopedia por hospital não autorizado.

79. *Procedimento solicitado não admite mudança de procedimento* - procedimento solicitado não permite mudança de procedimento (ex.: primeiro atendimento).

80. *RN sala de parto p/ procedimento diferente de parto* - cobrança de neonatologista em atendimento ao recém nato em sala de parto em procedimento diferente de parto.

81. *SIPAC-Câncer incompatível com material* - utilização indevida de material de OPM em procedimentos de alta complexidade em câncer.

82. *Hospital pertencente municípios em gestão Plena* - apresentação de cobrança em hospital pertencentes aos municípios em gestão Plena na fita nacional.

83. *AIH bloqueada pelo gestor do SUS* - realização pelo gestor de bloqueio de AIH através do programa PGFAIH.

84. *AIH com valor igual a zero PT/38 de 17/05* - hospitais que apresentaram AIH sem passar pela função gerar valores do programa SISAIH01.

85. *Hospital com endereço incompleto no cadastro* - hospital com dados cadastrais incorretos (Conta Corrente, C.E.P.)

86. *Tempo de permanência incompatível com o procedimento realizado* - procedimento que não atingiu pelo menos 50% do tempo de permanência previsto na Tabela do SIH/SUS.

87. *Diagnóstico principal (CID) incompatível com sexo* - procedimento realizado incompatível com o sexo.

88. *Não é permitida alta diretamente da UTI* - somente é permitida alta diretamente da UTI nos casos de óbito e transferência. Nos demais a AIH é rejeitada.

89. *AIH suspensa Ofício COSAU 355/95* - AIH rejeitada por tempo de permanência e taxa de ocupação. Somente serão pagas com justificativa do gestor em processamento em separado.

90. *Procedimento realizado incompatível com faixa etária* - procedimento incompatível com a idade do paciente.

91. *Mesmo CPF para cirurgião/anestesista/auxiliar cirúrgico* - não poderá o mesmo profissional executar os atos de cirurgião, anestesista e auxilio cirúrgico na mesma AIH.

92. *Mesmo CPF para mais de um auxílio* - não poderá o mesmo profissional executar atos de 1°, 2°, e 3° auxilio cirúrgicos.

Anexo 2 Descrição dos Atributos Referentes à Atos Médicos e à Procedimentos Especiais

- 17018030 hemograma completo
- 17023041 creatinina
- 17042046 glicose
- 17055040 potássio
- 17064040 uréia
- 17059046 sódio
- 04001010 eletrocardiograma
- 17009065 rotina de urina
- 21005060 tórax: pa, lateral
- 21003068 tórax: pa
- 17041040 gasometria completa
- 17012015 urina - cultura
- 17019044 colesterol total
- 32028040 dissecção da veia com colocação de cateter
- 17027039 protombina, tempo ou consumo, outros de coagulação
- 17019036 determinação de hemossedimentação
- 17063043 triglicerídios
- 23005025 aerosol-nebulizaçãoo sem rppt duração 15 a 20 min.
- 17023033 contagem de plaquetas
- 17066042 antibióticos
- 17008018 cultura em geral
- 17061040 transaminase oxalacetica ou piruvica
- 23004029 aerosol-inaloterapia rppt 15-20 min
- 17007011 bacterioscopico(gram,ziehl,albert)
- 17018048 hdl colesterol
- 17011043 cálcio
- 17056047 proteínas totais
- 17009049 bilirrubina total e frações
- 42004039 traqueotomia
- 97004057 estudo hemodinâmico de artérias coronarias
- 97006009 arteriografia de carótida bilateral
- 97007005 arteriografia vertebral
- 97011002 aortografia
- 97013013 tomografia computadorizada (crânio e coluna)

Bibliografia

[ADR96] ADRIAANS, P.; ZANTINGE, D. **Data Mining**. Harlow: Addison-Wesley, 1996. 158p.

[AGR2000] AGRAWAL, R. **An Interval Classifier for Database Mining Applications**. Disponível em: <http://www.acm.org/sigmod/vldb/conf/1992/P560.pdf>. Acesso em: 27 mar. 2000.

[AGR93] AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A.: Database Mining: a Performance Perspective. **IEEE Transactions on Knowledge and Data Engineering**, New York, v.5, n.6, p.914-925, Dez.1993.

[AGR97] AGRAWAL, R. **Data Mining**. Tutorial apresentado no 12. Simpósio Brasileiro de Banco de Dados. Fortaleza: Ufc, 1997.

[ARU99] ARUNASALAM, M. **Data Mining**. Disponível em <http://www.rpi.edu/~arunmk/dm1.html>. Acesso em: 21 dez. 1999.

[BAR2002] BARBOSA, L. R. **Terapia Ocupacional e Adaptações em AVC**. Campo Grande: UCDB, 2002. 48p.

[BEM97] BERRY, M.; LINOFF, G. **Data Mining Techniques**: For Marketing, Sales and Customer Support. New York: John Wiley & Sons, 1997. 454p.

[BER97] BERSON, A.; SMITH, S. **Data Warehousing, Data Mining e OLAP**. New York: McGraw-Hill, 1997. 612p.

[BIO97] BIOCH, J.; MERR, O.; POTHARST, R. Bivariate Decision Trees In:.EUROPEAN SYMPOSIUM ON PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY, pkdd, 1., 1997. **Principles of Data Mining and Knowledge Discovery**: proceedings. Berlin: Springer-Verlag, 1997. p.232-242.

[BRA2000] BRAZDIL, P. **Construção de Modelos de Decisão a partir de Dados**. Disponível em: <http://www.ncc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html>. Acesso em: 17 jun. 2000.

[BRA96] BRACHMAN, R.; ANAND, T. The Process of Knowledge Discovery in Databases. In: FAYYAD, U. et al. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA: AAAI Press, 1996. p.37-57.

[CAB97] CABENA, Peter et al. **Discovering Data Mining from Concept to Implementation**. Upper Saddle River, New Jersey: Prentice Hall, 1997. 195p.

[CAM2000] CAMPANI, C.; OLIVEIRA, S.; RODRIGUES, A. **Entropia e Teoria da Informação**. Disponível em: http://www.ufpel.tche.br/~campani/grupo.htm>. Acesso em: 20 jul. 2000.

[DAN2000] DANKEL, D. **The ID3 Algorithm.** Disponível em: <http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>. Acesso em: 23 mar. 2000.

[FAY2000] FAYYAD, U. **Data Mining and Knowledge Discovery**. Disponível em: <http://www.research.microsoft.com/research/datamine/vol1-1/editorial3>. Acesso em: 06 jan. 2000.

[FAY96] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. From Data Mining to Knowledge Discovery: an overview. In: FAYYAD, U. et al. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA: AAAI Press, 1996. p.1-34.

[FON94] FONSECA, J. **Indução de Árvores de Decisão**: HistClass – proposta de um algoritmo não paramétrico. 1994. 140p. Dissertação (Mestrado em Engenharia Informática) – Universidade Nova Lisboa, Lisboa.

[FRE97] FREITAS, A. **Generic, Set-Oriented Primitives to Support Data-Parallel:** Knowledge Discovery in Relational Database Systems. Thesis, UK: University of Essex, 1997. 275 p.

[GOE99] GOEBEL, M.; GRUENWALD, L. **A Survey of Data Mining and Knowledge Discovery Software Tools**. Disponível em: <http://www.acm.org/sigkdd/explorations/issue1-1/survey.pdf>. Acesso em: 18 dez 1999.

[HOL2000] HOLSHEIMER, M.; SIEBES, A. **Data Mining:** the search for knowledge in databases**.** 1994. Disponível em: <ftp.cwi.nl, no arquivo /pub/CWIreports/AA/CS-R9406.ps.Z>. Acesso em: 22 jan. 2000.

[ING2000] INGARGIOLA, G. **Building Classification Models:** ID3 and C4.5. Disponível em: <http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html>. Acesso em: 22 jan. 2000.

[JAY2000] JAYNES, E. **Discrete Prior Probabilities - The Entropy Principle**. Disponível em: <http://omega.albany.edu:8008/JaynesBook.html>. Acesso em: 15 jul. 2000.

[JOH97] JOHN, G. **Enhancements to the Data Mining Process.** Dissertation, Stanford: Stanford University, 1997. 194p.

[LIU94] LIU, W.; WHITE, A. The Importance of Attribute Selection Measures in Decision Tree Induction. **Machine Learning**, Dordrecht, v. 15, n. 1, p. 25-41, 1994.

[MAR2000] MARTIN, J.; STUART, T. **Learning from Observations**. Disponível em: <http://sern.ucalgary.ca/courses/CP...nos/L2_18B_Martin_Stuart/main.html>. Acesso em: 18 jul. 2000.

[MAT96] MATHEUS, C. J.; PIATETSKY-SHAPIRO, G.; MCNEILL, D. Selecting and Reporting What is Interesting: the KEFIR application to healthcare data. In: FAYYAD, U. et al. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, CA: AAAI Press, 1996. 494-515.

[MIT2000] MITCHELL, T. **Machine Learning and Data Mining**. Disponível em: <http://www.cmu.edu/~tom/publications.html>. Acesso em: 12 fev. 2000.

[MIT97] MITCHELL, T. **Machine Learning**. New York: McGraw-Hill, 1997. 414p.

[MOU93] MOURA-PIRES, F.; MOURA-PIRES, J. Modelo para Conjuntos de Treino e Indução de Árvores de Decisão. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 10., 1993, Porto Alegre. **Anais...** Porto Alegre: SBC, 1993. p. 225-238.

[OSU93] O'SULLIVAN, Susan B.; SCHMITZ, Thomas J. **Fisioterapia**: avaliação e tratamento. 2 ed. São Paulo: Manole, 1993. 519p.

[QUI86] QUINLAN, J. Induction of Decision Trees. **Machine Learning**, Dordrecht, v. 1, n. 1, p. 81-106, 1986.

[QUI93] QUINLAN, J. **C4.5**: Programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302p.

[RUL2000] RULEQUEST. **Data Mining Tools See5 and C5.0**. Disponível em: <http://www.rulequest.com>. Acesso em: 13 jul. 2000.

[SBN2002] SOCIEDADE BRASILEIRA DE NEUROCIRURGIA. **Manuais de Orientação**: Derrame Cerebral. Disponível em: <http://www.sbn.com.br/programas/prev13.htm>. Acesso em: 10 nov. 2002.

[SOM98] SOUZA, M. **Mineração de Dados**: uma implementação fortemente acoplada a um sistema gerenciador de banco de dados paralelo. 1998. 75p. Dissertação, (Mestrado em Engenharia de Sistemas e Computação) – Instituto de Engenharia, UFRJ, Rio de Janeiro.

[SOU98] SOUZA, M.; MATTOSO, M.; EBECKEN, N. Data Mining: a database perspective. In: EBECKEN, N. **Data Mining**. Boston: WIT Press, 1998. p.413-431.

[SUS98] SISTEMA ÚNICO DE SAÚDE. **Manual da AIH**: módulo hospitalar. Rio de Janeiro, Ministério da Saúde, 1998. 87p.

[UEA99] UEA. **Knowledge Discovery In Databases and Data mining**. Disponível em: <http://www.sys.uea.ac.uk/Research/researchareas/MAG/projects/data_mining.html>. Acesso em: 15 dez. 1999.

[UTG97] UTGOFF, P.; BERKAMN, N., CLOUSE, J**.** Decision Tree Induction Based on Efficient Tree Restructuring. **Machine Learning**, Dordrecht, v. 29, n. 1, p. 5-44, 1997.

[WEI98] WEISS, S. **Predictive Data Mining**: a practical guide. San Francisco: Morgan Kaufmann, 1998. 228p.

[WHI94] WHITE, A.; LIU, W. Bias in Information-Based Measures in Decision Tree Induction. **Machine Learning**, Dordrecht, v. 15, n.1, p.321-329, 1994.

[WRI2000] WRIGHT, P. **Knowledge Discovery In Databases:** Tools and Techniques. Disponível em: <http://www.acm.org/crossroads/xrds5-2/kdd.html>. Acesso em: 12 fev. 2000.

[ZIG2000] ZIGHED, D. **Sipina**. France: Université Lumière de Lyon. Arquivo de ajuda da ferramenta Sipina, versão beta. Disponível em: <ftplyon2.univ-lyon2.fr no arquivo /pub/pc/Eric/SIPINA>. Acesso em: 10 mar. 2000.