PATRÍCIA NASCIMENTO MANFRÉ BRANCO

Concepção de um Sistema de Informação Integrado Inteligente para Apoio ao Profissional Fisioterapeuta: Construído sobre uma Base de Dados Simulados

Curitiba

2010

PATRÍCIA NASCIMENTO MANFRÉ BRANCO

Concepção de um Sistema de Informação Integrado Inteligente para Apoio ao Profissional Fisioterapeuta: Construído sobre uma Base de Dados Simulados

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná como requisito para obtenção do título de mestre em Tecnologia em Saúde.

Área de Concentração: Informática em Saúde.

Orientador: Prof. Dr. Edson Emílio Scalabrin.

Curitiba

2010

Branco, Patrícia Nascimento Manfré

B816c 2010 Concepção de um sistema de informação integrado inteligente para apoio ao profissional fisioterapeuta : construído sobre uma base de dados simulados / Patrícia Nascimento Manfré Branco ; orientador, Edson Emílio Scalabrin. -- 2010.

108 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2010

Bibliografia: f. 96-100

1. Sistemas de recuperação da informação – Reabilitação. 2. Banco de dados. 3. Fisioterapia. I. Scalabrin, Edson Emílio. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Tecnologia em Saúde. III. Título.

CDD 20. ed. – 615.820285

Pontifícia Universidade Católica do Paraná
Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Tecnologia em Saúde

ATA DA SESSÃO PÚBLICA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
DO PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE
DA PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

DEFESA DE DISSERTAÇÃO Nº 124

Aos 29 dias de julho de 2010 realizou-se a sessão pública de defesa da dissertação **"Concepção de um Sistema de Informação Integrado Inteligente para Apoio ao Profissional Fisioterapeuta: Construído Sobre uma Base de Dados Simulados"**, apresentada por **Patrícia Nascimento Manfré Branco** como requisito parcial para a obtenção do título de Mestre em Tecnologia em Saúde – Área de Concentração – **Informática em Saúde** perante uma Banca Examinadora composta pelos seguintes membros:

	assinatura	parecer (aprov/ reprov.)
Prof. Dr. Edson Emílio Scalabrin PUCPR (Orientador)	Ed Scal	Aprovado
Prof. Dr. Gerson Linck Bichinho, PUCPR		APROVADO
Prof. Dr. João da Silva Dias UP		Aprovado

Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado aprovado (*aprovado/reprovado*), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do Programa.

Prof. Dr. Munir Antonio Gariba,
Coordenador do PPGTS PUCPR

AGRADECIMENTOS

Aos meus pais, esposo e filhas, pelo exemplo, amor e paciência neste período.

Ao meu Orientador, Prof. Dr. Edson Emílio Scalabrin, pelo constante apoio, persistência e aposta nesta pesquisa, e sobre tudo obrigada pelos grandes ensinamentos científicos, intelectuais e pessoais.

Aos Prof. Gerson Linck Bichinho e João Dias pelas contribuições e debates no exame de qualificação, à Prof. Claudia Moro pelo auxílio nas publicações.

Aos colegas de turma e a Erly, secretária e companheira, que com muita paciência acompanharam todo este processo.

A PUC – PR pela oportunidade de realizar este sonho.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

RESUMO

O presente trabalho de pesquisa concerne à concepção de um sistema integrado de informação baseado em recursos computacionais para área de fisioterapia dermato-funcional. A abordagem de concepção é dupla: a) à utilização de técnicas de aprendizagem de máquina para a descoberta de padrões relevantes ao planejamento e acompanhamento de pacientes que se submeteram a diferentes cirurgias plásticas; b) à utilização dos conhecimentos descobertos, de forma interativa, no processo de planejamento e acompanhamento de um paciente, à medida que cada novo evento relativo a um paciente é registrado no sistema clínico, um conjunto de conhecimentos, na forma de conjuntos disjuntos de regras, são aplicados ao evento. A base de dados usada para a obtenção dos padrões foi obtida a partir de um processo de simulação; o valor de cada atributo de um dado exemplo foi obtido por meio de uma função de geração de valores aleatórios aplicada sobre o domínio do atributo em questão.

A metodologia aplicada para concepção do sistema baseou-se nas etapas clássicas do processo de descobrimento de conhecimento e modelagem de dados multidimensional. A descoberta de conhecimento baseou-se em três métodos de classificação: C4.5, BAGGING e BOOSTING. A validação dos conhecimentos descobertos foi realizada por meio do método de validação cruzada das taxas de acertos destes métodos.

A interação do usuário com o sistema é de fundamental importância, sendo esta interação ciclos que se repetem a cada novo evento gerado pelo usuário, permitindo que a aplicação dos conhecimentos descobertos também seja de forma interativa. Tal modo de operação permite que o usuário obtenha informações adicionais, fornecidas pelos classificadores, mesmo com um número pequeno de eventos para um paciente, e à medida que o histórico do paciente é incrementado as informações fornecidas pelos classificadores tendem ser de melhor qualidade.

Os resultados encontrados mostram que os métodos BAGGING e BOOSTING melhoram a taxa de acerto em comparação ao método JRIP isolado; havendo ainda, como esperado teoricamente, uma melhor classificação, com menor taxa de erro médio e desvio padrão, pelo método BOOSTING. Observou-se também, que a taxa de acerto dos três métodos para os classificadores gerados a partir do conjunto de treinamento completo e reduzidos, por filtragem de atributos, mostrou pequena diferença nas taxas de erro e desvios padrão, o que nos permite concluir que a técnica de combinação de classificadores pode ser eficaz em situações onde o número de atributos é pequeno.

É importante salientar que os resultados referentes aos padrões descobertos devem ser vistos a luz de uma base de dados gerada por meio de um processo de simulação.Não se trata de uma base de dados real.

Palavras-chave: 1. Classificação. 2. Descoberta de Conhecimento. 3. Clínica de Estética.

ABSTRACT

The present research work relates to the development of an integrate system of information based on computational resources for aesthetic physical therapy area. The conception boarding is twofold and concerns: a) the use of machine learning techniques to discovery relevant standards to the planning and accompaniment of patients submitted to different plastic surgeries; b) the use of the discovered knowledge, interactively, in the process of planning and accompaniment of a patient, as each new event related to this patient is registered in the clinical system, a set of knowledge, in set rules form, is applied to the event. The used data base for the attainment of standards was gotten by a simulation process; the value of each attribute of data example was gotten by means of a function of generation of random values applied on the domain of the attribute.

The methodology applied for the system development is based on the classic stages of the process of knowledge discovery and multidimensional data modeling. The knowledge discovery was based on three methods of classification: C4.5, BAGGING, and BOOSTING. The validation of the knowledge discovery was done by the crossed validation method of rightness rates of these methods.

The interaction of the user with the system is very important, this interaction is: cycles that is repeated to each new event generated by the user, allowing the discovered knowledge application to be interactive. Such operation way allows the user to get additional information, provided by the classifiers, even with a small number of events per patient, and as soon as the patient's history is increased, the information provided by the classifiers tend to be of better quality.

The results show that the BAGGING and BOOSTING methods improve the rightness rates in comparison to the method C4.5 (JRIP) isolated; also, as hypothesized, a better classification, with the smallest mean error rate and standard deviation, on the BOOSTING method. It was also observed that the rightness rate of the three methods for classifiers created from the set of complete and reduced training, by filtering the attributes, showed a small difference on the error rate and standard deviations, what allow us to conclude that the classifiers combination technique can be efficient in situations where the number of attributes is small.

It is important to point out that the referring results to the standards discovered must be seen as a database created by a simulation process, is not about a real database.

Key words: 1. Classification. 2. Knowledge Discovery. 3. Aesthetic Clinic.

1 INTRODUÇÃO

Esta dissertação enquadra-se no contexto do Projeto A2M (Aplicação e Avaliação de Metodologias). O objetivo geral do A2M é aplicar e avaliar metodologias de apoio à tomada de decisão na área de saúde. O principal foco encontra-se nos processos de descoberta, validação e disseminação de conhecimentos relevantes ou estratégicos. O projeto objeto desta dissertação é um subprojeto do Projeto A2M. A abordagem é dupla, e diz respeito à: (i) utilização de técnicas de aprendizagem de máquina para a descoberta de padrões ao planejamento e acompanhamento de pacientes hipotéticos que se submeteram a cirurgias estéticas em uma clinica hipotética; e (ii) utilização dos conhecimentos descobertos de forma interativa no processo de planejamento e acompanhamento de um paciente, à medida que cada evento (*EV*) relativo a um paciente é registrado no sistema clínico, um conjunto de conhecimentos na forma de modelos/classificadores são aplicados ao *EV*. Desta forma, mesmo que incompleto o histórico do paciente, o usuário/profissional de fisioterapia, pode obter informações adicionais por meio das respostas dos classificadores ao *EV*. Em outras palavras, quando o número de *EV* sobre um paciente é pequeno (e.g., um), a maioria das repostas dos classificadores corresponderão aos conseqüentes das regras *default*. E à medida que o número de *EV* sobre o mesmo paciente for incrementado, as repostas dos classificadores não serão mais necessariamente aquelas das regras *default*. Os principais *EV* pós-cirúrgicos, foco deste trabalho, concernem aqueles da área de fisioterapia.

Fisioterapia é a ciência da reabilitação funcional do indivíduo como um todo, e é dividida em várias áreas: respiratória, neurológica, dermato-funcional, cardiopulmonar, do trabalho, ortopédica, entre outras, que por sua vez podem ser divididas em subáreas como a fisioterapia dermato-funcional no pós-operatório de cirurgia plástica que tem como objetivo principal acelerar o processo de recuperação pós-operatória, a prevenção e o controle das possíveis complicações associadas ao pós-operatório (GUIRRO & GUIRRO, 2004). Nestes termos, o tratamento fisioterapêutico requer conhecimentos especializados, para avaliação e seleção da conduta e recursos fisioterapêuticos que serão utilizados, como: massoterapia, eletroterapia e/ou cinésio-mecanoterapia (SILVA, 2006); com isto, um grande volume de dados relativos aos procedimentos executados e suas avaliações são registrados e armazenados, muitas vezes de forma isolada, ou

ainda, não registrados e perdidos. Na área da saúde, em geral, ainda é comum esta perda de dados e registros, em função do grande volume de atendimentos e pela falta de padronização dos prontuários, fazendo com que o tratamento muitas vezes seja estabelecido conforme o relato e observação momentânea do quadro clínico do paciente, sem a coleta e conservação adequada destes dados (FRONZA & OSÓRIO, 2006).

Os dados, quando podem ser consultados, cruzados, analisados ou simplesmente compartilhados, são grandes geradores de informação, justificando a necessidade da preservação e padronização dos mesmos (KIMBALL & MERS, 2000). A informática em saúde, atualmente, disponibiliza ferramentas para o processamento de dados relevantes de cada paciente, afim de, facilitar a tomada de decisão nos procedimentos e orientações terapêuticas, como também no gerenciamento administrativo, contábil e organizacional de uma clínica (SIGULEM, 1998).

É importante, entretanto, observar e identificar a maneira como estes dados são armazenados ou apresentados para o profissional da saúde, pois, muitas vezes ao invés de auxiliá-lo, pode induzi-lo ao erro (CANESTRARO, et al.,2006). Geralmente, este problema se dá pelo fato de que dados importantes do paciente, armazenados ou não em bases de dados, digitais ou analógicos, são pouco aproveitados para dar suporte à decisão, pois, como citado anteriormente, são em sua grande maioria dados heterogêneos e subjetivos, que podem ser irrelevantes ou redundantes, até mesmo pelo fato de estarem mal distribuídos em bases muito individualizadas sem qualquer integração.

Neste contexto, pensando em uma melhor organização e distribuição destes registros e dados de pacientes submetidos à cirurgia plástica, e na busca por conhecimentos, a tarefa de aprendizagem de máquina e descoberta de padrões foi desenvolvida ao decorrer desta dissertação.

Desta forma, algumas etapas precisaram ser realizadas: (i) o estudo dos modelos dos dados sobre os procedimentos fisioterapêutico e suas avaliações; (ii) enriquecimento da base dados, por meio de técnicas de filtragem, seleção de atributos e inserção de novos atributos relevantes ao problema, que foi estruturada, particularmente, em uma base de dados multidimensional, como um d/w (*data warehouse)*, que pode resolver tais problemas de tal forma que o profissional possa utilizar e acessar seus registros de maneira eficaz e efetiva, auxiliando na tomada de decisões e tornando viável a transformação dos dados em informação, bem como obtenção de conhecimento útil (INMON, 2002 & VREEMAN, et al., 2006).

A aprendizagem de máquina foi realizada por métodos supervisionados simples, denominado C4.5, e compostos, denominados BAGGING e BOOSTING. Esses últimos são ditos compostos porque eles combinam classificadores simples para realizar uma classificação, onde o processo de decisão é similar ao de um comitê de especialistas. O principal interesse nos métodos C4.5, BAGGING e BOOSTING concerne, essencialmente, a geração de regras de fácil compreensão ou leitura.

1.1 Motivação

O escopo de aplicação concerne os tratamentos fisioterápicos pós-operatórios em cirurgia estética, no intuito de dinamizar o atendimento, auxiliar nas decisões e condutas do fisioterapeuta. Tais procedimentos, em geral, são traçados segundo os achados clínicos da avaliação fisioterapêutica, baseando-se nos dados pessoais, da história pregressa, antecedentes patológicos, história atual, anamnese, inspeção física e evolução do paciente. No entanto, a prática diária nos mostra uma grande perda de dados e até mesmo de registros dado o volume de atendimentos e falta de padronização dos prontuários ou preenchimento destes.

A análise dos registros na avaliação fisioterápica em clínicas de estética mostrou inconsistência dos dados não apenas por apresentar dados ausentes ou incorretos, mas também pelo fato de serem coletados por profissionais diferentes, de uma mesma área ou não, como: diversos fisioterapeutas, ou auxiliares de enfermagem, médico, instrumentadora ou secretárias. Onde dados com mesmo significado podem ser encontrados em registros diferentes representados de forma heterogênea para o sistema, como exemplo podemos citar: a secretária registra que o paciente realizou cirurgia de pálpebra e o médico registra blefaroplastia, os dois termos estão corretos e podem representar a mesma cirurgia, porém para o meio computacional são coisas distintas; ou ainda, um fisioterapeuta registra que o paciente encontra-se no 4º dia pós-operatório, outro identifica a data em dia, mês e ano. O que torna estes dados totalmente "independentes", porém sem padronização e, ainda, inconsistentes, e esta inconsistência torna praticamente impossível o cruzamento dos dados para geração de informação.

Neste cenário, o desafio é minimizar problemas de distribuição e heterogeneidade dos registros e dados, para tornar possível, por exemplo, a transformação de dados em informações integradas e confiáveis, que possam ser compartilhadas e disseminadas entre todos os profissionais envolvidos com determinado paciente, seja este profissional médico, fisioterapeuta,

secretária ou administrador financeiro da clínica, facilitando a dinâmica do trabalho e criando certa homogeneidade no atendimento.

A concepção de um sistema que facilitasse a dinâmica de uma clínica de estética auxiliou na construção de soluções para problemas da rotina dos profissionais, bem como para problemas oculto, que aparecem com o co-relacionamento dos dados e das informações ainda não conhecidas. Para o sucesso do sistema, é necessário o conhecimento das ações e rotinas dos profissionais na clínica. Também é fundamental que as informações sejam baseadas especificamente nas exigências e necessidades dos profissionais que as utilizam. Para levar em conta tais necessidades, pensou-se então, em questões que o sistema pudesse responder como: Quais são as idades dos pacientes que mais operaram nos últimos três anos? Qual é a relação existente entre queixa principal e tipo de cirurgia? Qual é a relação entre alterações e o tipo de cirurgia? Qual é o percentual de satisfação dos pacientes? Entre outras. Uma vez identificada estas questões, iniciou-se a primeira etapa para construção do d/w.

1.2 Objetivo

O objetivo geral deste trabalho consiste na concepção de um sistema de informação integrado inteligente, a inteligência ocorre à medida que as fontes de dados operacionais, informacionais e de conhecimento operam em um ciclo produtivo de apoio à decisão do cenário.

Desta forma este trabalho compreende em estudar os dados históricos de várias especialidades de uma clínica de estética, afim de, obter padrões apartir destes dados e empregar técnicas computacionais de aprendizagem de máquina propostas por Quinlan (1993), Freund e Shapire (1996) e Breiman (1996).

Os objetivos específicos são:

- compreender e enriquecer os dados históricos de uma clínica de cirurgia estética e tratamentos fisioterapêutico pós-operatórios;
- extrair, a partir dos dados estudados e enriquecidos, conhecimentos/regras úteis à elaboração de planos de tratamentos em termos de fisioterapia dermato-funcional;
- validar os conhecimentos obtidos;
- definir um módulo de software que mostre à aplicabilidade potencial das regras descobertas vis-à-vis a elaboração de um sistema de apoio a decisão.

1.3 Hipótese

A presente pesquisa propõe a extração de modelos/classificadores confiáveis e compreensíveis que evidenciem a evolução antecipada de um paciente que se submete a uma ou mais cirurgias estética. Nós acreditamos que a aplicação de técnicas de descoberta de conhecimento, sobretudo métodos de aprendizagem de máquina gerem informações úteis à medida que o histórico de um paciente é construído. A aplicação de métodos de aprendizagem de máquina tem se mostrado eficaz na descoberta de regras em bases de dados de diferentes áreas de aplicação. As regras descobertas devem permitir a construção de um módulo de software interativo, fornecendo informações ao profissional de fisioterapia já a partir do primeiro evento registrado sobre um determinado do paciente; espera-se que a taxa de acerto sobre tais informações melhorem a partir que o histórico é incrementado.

1.4 Contribuições

As contribuições científicas da presente pesquisa são: (i) concerne à obtenção e a validação de padrões relevantes ao planejamento e acompanhamento de pacientes que se submeteram a cirurgias estéticas; e (ii) concerne à utilização dos conhecimentos descobertos de forma interativa no processo de planejamento e acompanhamento de um paciente, à medida que cada evento *EV* relativo a um paciente é registrado no sistema clínico, um conjunto de conhecimentos na forma de modelos/classificadores são aplicados ao *EV*. Desta forma, mesmo que ainda incompleto o histórico do paciente, o usuário/profissional de fisioterapia, pode obter informações adicionais por meio das respostas dos classificadores ao *EV*.

1.5 Estrutura do Documento

O restante do documento foi estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica em face ao objeto principal do estudo, a descoberta de conhecimentos por meio de métodos computacionais. São descritos em detalhes o método C4.5 para a obtenção de classificadores simples e os métodos BAGGING e BOOSTING para a obtenção de classificadores compostos, bem como o método de validação cruzada. O Capítulo 3 apresenta em detalhes a metodologia. O Capítulo 4 apresenta os resultados da pesquisa e uma discussão sobre os mesmos, e por fim, as nossas considerações finais e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA – DESCOBERTA DE CONHECIMENTOS

2.1 Introdução

A descoberta automatizada de conhecimentos, a partir de base de dados, é uma necessidade imperativa vis-à-vis ao reconhecimento de padrões de comportamento dos dados para apoiar processos decisórios em um determinado domínio (SALVADOR, et al., 2006). De forma pragmática, assumiu-se que: (i) dado é evento; (ii) informação é a sumarização de um conjunto de eventos (e.g., número total de cirurgias de lipoaspiração realizadas no ano de 2008); (iii) conhecimento é um conjunto não vazio de regras ordenadas induzidas a partir de um conjunto de dados (e.g., SE QUEIXA = NÃO e ATIVIDADE FÍSICA = SIM ENTÃO ALTERAÇÃO = SIM). A obtenção de tal conhecimento é a tarefa mais desafiadora, dado o elevado nível de abstração necessário para realizar o processo de aprendizagem. Este último, em Ciência da Computação, assume a forma de um algoritmo de indução de regras a partir de um conjunto de dados.

A fundamentação apresentada, na seqüência, aborda de forma detalhada alguns algoritmos para a obtenção de classificadores simbólicos e técnicas de validação. Cada classificador pode assumir a forma de uma árvore de decisão ou de um conjunto ordenado de regras. A complexidade do processo de obtenção de um classificador está fortemente ligada à ordem do domínio de cada atributo e também do número destes na base de dados; assim, a redução da dimensionalidade dos dados facilita sobremaneira a extração dos padrões. Tal redução pode ser feita por meio da técnica de seleção de atributos. Nestes termos, a fundamentação teórica porta essencialmente sobre os seguintes temas:

- seleção de atributos como uma forma de redução da complexidade dos dados por meio da filtragem de atributos irrelevantes;
- aprendizagem indutiva como método básico à obtenção de padrões a partir de um conjunto de dados/eventos;
- aprendizagem de máquina simbólica mono-classificador e multiclassificador; e
- avaliação de classificadores realizada por meio da técnica de validação cruzada.

A abordagem de apresentação da fundamentação dar-se-á na seguinte ordem: seleção de atributos, aprendizagem indutiva, aprendizagem de máquina simbólica e validação cruzada.

2.2 Seleção de Atributos

O objetivo da seleção de atributos é encontrar o melhor subconjunto de dados a partir de um determinado critério. A aplicação de tal critério pode ser vista como um processo de filtragem de dados. Desta forma, alguns atributos irrelevantes podem ser removidos da base de dados, o que potencialmente aumentará o desempenho do classificador e reduzirá a taxa de erro do mesmo (HUAN, et al., 1998) (LIU, et al., 2005).

A obtenção do melhor subconjunto *T* de dados pode ser assegurada por meio de uma busca completa. Uma busca é completa quando ela abrange todas as combinações de atributos possíveis. Ela pode ser realizada de forma seqüencial inserindo ou removendo atributos ao subconjunto *T*, ou gerando um conjunto inicial aleatório e, a partir dele, construir o conjunto *T* de modo seqüencial (LIU, 1998).

O primeiro passo no processo de seleção de atributos é gerar, a partir de um conjunto de dados e por meio da adição ou remoção atributos, um subconjunto *S* que atenda certa medida de qualidade. Caso a qualidade de *S* seja satisfatória, atendendo o critério de parada, então *S* é considerado o melhor subconjunto gerado. Do contrário, um novo subconjunto é gerado e o processo é repetido. A qualidade de um subconjunto pode ser medida por três métodos distintos (BORGES, 2009):

- *Wrapper*: a avaliação do subconjunto é realizada por meio de um algoritmo de aprendizagem selecionado como critério de parada e aplicado sobre um conjunto de atributos;
- *Filtro*: a avaliação é feita de acordo com um critério de parada que é uma medida independente do algoritmo de aprendizagem e aplicada em um único atributo por vez. Em casos que utilizam bases de dados reais e com grande quantidade de atributos, o método de filtragem mostra-se computacionalmente menos custoso em relação, por exemplo, ao modelo Wrapper;
- *Híbrido*: neste modelo, o subconjunto é avaliado primeiramente por meio de uma medida independente e na seqüência por um algoritmo de aprendizagem (SEBBAN, et al., 2001).

Em resumo, o critério de parada depende de quando o método de seleção de atributos utilizado encontra-se satisfeito: (i) quando a busca pelo melhor subconjunto foi alcançada; (ii) quando um limite – número de atributos ou número de interações – for atingido; (iii) quando a adição ou remoção de um certo atributo ocasiona uma melhora no subconjunto; ou (iv) quando um subconjunto suficientemente bom foi gerado – produzindo uma taxa de erro aceitável para determinada tarefa. A seleção de atributos é uma etapa importante porque, a partir dela, é possível obter o conjunto reduzido de atributos e, conseqüentemente, a quantidade dos dados antes da execução da tarefa de descoberta automática de conhecimentos.

2.3 Descoberta Automática de Conhecimentos

Classicamente, a descoberta automatizada de conhecimentos é vista como um processo complexo. E tal processo busca encontrar padrões válidos e interessantes a partir de conjunto de dados de um determinado domínio (CIOS, et al., 2007) (HAN, et al., 2006). Estes padrões são interessantes se forem de fácil compreensão, potencialmente utilizáveis e novos, e válidos se forem aplicáveis para classificar novos dados/eventos.

De forma objetiva, uma medida indispensável na análise do grau de interesse de uma regra é o suporte, que representa a porcentagem de transações da base de dados que satisfaz a regra, o cálculo do suporte está representado na Fórmula (1) da Tabela 1. Outra medida indispensável é o grau de confiança, que expressa à probabilidade de que uma regra contendo o atributo X contenha outro atributo Y (HAN, et al., 2006), mostrado na Fórmula (2) da Tabela 1. A aceitação ou não de uma regra baseada em tais medidas pode ser controlada pelo usuário, que pode estabelecer, por exemplo, a aceitação de regras com suporte maior que 50%. Mais a frente será mostrado um exemplo de cálculo para suporte e confiança.

Tabela 1. Fórmulas de Suporte e Confiança.

$Suporte = \frac{N°de\ registros\ com\ X\ e\ Y}{N°total\ de\ registros}$	(1)
$Confiança = \frac{N°de\ registros\ com\ X\ e\ Y}{N°de\ registros\ com\ X}$	(2)

O interesse é aplicar os algoritmos de descoberta automática de conhecimentos em bases de dados históricos da área de Dermato-Funcional, subárea da Fisioterapia, visando à descoberta de regras de caracterização de alterações pós-operatórias. O uso da mineração de dados (FAYYAD, 1996; FAYYAD et al., 1996) tem se mostrado eficiente em diversas áreas do conhecimento humano, como na química (KALOS, et al., 2005), classificação de músicas (CHEN, et al., 2009), exploração de conhecimento médico (RODDICK, et al., 2003), extração de regras de condução de trens de carga (BORGES, 2009), entre outras. E acreditando na facillidade que a aprendizagem de máquina possa oferecer, na agilidade de identificação das alterações pós-operatórias, ao profissional de fisioterapia. Este estudo buscou combinar as contribuições da descoberta de conhecimento aos procedimentos e condutas da fisioterapia dermato-funcional.

2.3.1 WEKA

Os experimentos foram realizados com o pacote de software WEKA. Ele compreende uma coleção de algoritmos de aprendizagem para tarefas de descoberta automática de conhecimentos a partir de bases de dados (WITTEN, et al., 2005).

2.4 Aprendizagem de Máquina

A aprendizagem de máquina é uma área da inteligência artificial com o objetivo de desenvolver algoritmos e técnicas que permitam ao computador aprender, ou seja, adquirir conhecimento de forma automática. Os sistemas de aprendizagem de máquina tomam decisões baseados em experiências acumuladas por meio de soluções bem sucedidas de eventos anteriores. Dado um conjunto de exemplos *T* rotulados, a aprendizagem de máquina pode ser vista como a inferência automática de conceitos a partir de *T* (MITCHELL, 1997).

Existem algumas diferenças entre a aprendizagem de máquina e a descoberta automática de conhecimentos a partir de base de dados. Segundo Prati, em (PRATI, 2006), a descoberta automática de conhecimentos a partir de base de dados é uma fonte de aplicação de algoritmos de aprendizagem de máquina, por prover dados reais e em grande volume. Na origem, os algoritmos de aprendizagem de máquina operavam apenas sobre pequenas bases de dados, com atributos previamente selecionados para facilitar o aprendizado. Como a descoberta automática de conhecimentos a partir de base de dados é parte de um processo maior, que vai desde a

preparação dos dados até a utilização dos resultados obtidos, ela permitiu que a aprendizagem de máquina explorasse maiores conjuntos de dados.

A aplicação prática da aprendizagem de máquina inclui diferentes formas de processamento de aprendizado, de linguagem de descrição e paradigmas; como aprendizagem de máquina: simbólica, que aprende construindo representações simbólicas de um conceito ex.: árvore de decisão, regras de produção; estática, que representa o aprendizado através de modelos estatísticos ex.: probabilístico bayesiano; baseada em exemplos, classifica exemplos nunca vistos como similares conhecidos, ex.: raciocínio baseado em casos; conexionista: aprendizado através da interconexão de unidades simples, ex.: redes neurais artificiais; evolutiva: onde os elementos fracos são descartados e os fortes predominam, ex.: genética.

2.4.1 Aprendizagem Simbólica de Máquina

A aprendizagem simbólica de máquina é fundamental à construção de um sistema de previsão. Um sistema de previsão é um programa de computador capaz de tomar decisões aplicando conhecimentos obtidos a partir de experiências ou de soluções de problemas, geralmente armazenadas em grandes bases de dados ao longo do tempo. Ou seja, um sistema de aprendizagem de máquina simbólica realiza a tarefa de aprender e construir representações simbólicas de um conceito a partir da análise de exemplos e contra-exemplos disponíveis na forma de um conjunto de treinamento. Tais representações assumem, em geral, a forma de árvores de decisão ou regras de produção e podem ser consideradas muito eficientes quando comparadas a outras abordagens de aprendizagem, como por exemplo, as redes neurais.

A aprendizagem de máquina simbólica é utilizada nas situações em que o modelo obtido assume uma forma compreensível. Tal compreensibilidade é, em alguns casos, fundamental. O sistema ID3 desenvolvido por Quinlan (1986) é um exemplo de sistema de aprendizagem simbólica, onde o conhecimento é identificado através da modelagem do problema e representado através de uma linguagem formal simples e de fácil compreensão por meio da indução de uma árvore de decisão. Em Quinlan (1987) tem-se também a realização de um sistema voltado à geração de regras de produção a partir de uma árvore de decisão.

Além destas importantes contribuições para a área de aprendizagem simbólica de máquina, o leitor encontra em Quinlan (1993), versões mais recentes e mais eficientes desses

algoritmos, denominadas de C4.5 e C5.0. O algoritmo C4.5 será detalhado na seqüência deste documento, bem como os métodos de combinação de classificadores BAGGING e BOOSTING.

Em resumo, um sistema de aprendizagem de máquina simbólica corresponde à automatização de um processo de aprendizagem, enquanto que a aprendizagem indutiva equivale à obtenção de regras baseada em observações de estados ambientais e transições.

2.4.2 Aprendizagem Indutiva

Na aprendizagem indutiva, o sistema de aprendizagem deduz o conhecimento pela observação do seu ambiente. Existem duas estratégias principais para realizar tal tarefa de dedução. A estratégia de *aprendizagem supervisionada* que corresponde à aprendizagem por meio de exemplos previamente classificados por um supervisor; o sistema reúne as propriedades comuns dos exemplos de cada classe, definindo uma descrição simbólica para cada classe. A estratégica de *aprendizagem não supervisionada* é regida pela aprendizagem por meio de observações não previamente classificadas. A busca da descrição de cada classe é realizada através do reconhecimento de padrões por si só, examinando os exemplos.

A aprendizagem indutiva consiste na extração de padrões a partir de exemplos. Um método de aprendizagem indutiva pode produzir um modelo cuja qualidade é tal que o mesmo poderia ser usado para predizer o resultado de situações futuras. A indução é uma forma de inferência lógica que permite a utilização de premissas para obter conclusões genéricas a partir de exemplos particulares. A indução pode ser caracterizada como uma forma de raciocínio que parte de um conceito específico e o generaliza (MALOOF, et al., 2000).

Um sistema de aprendizagem simbólica pode construir vários modelos a partir dos dados de um ambiente, onde alguns destes modelos são mais simples que outros. Neste caso, opta-se normalmente pelos mais simples. Tal decisão adere à teoria de Ockham, segundo a qual a pluralidade não deve ser posta sem necessidade ou se existem inúmeras explicações igualmente válidas para um fato, então, deve-se escolher a explicação mais simples (OCKHAM, 1999). Trata-se de uma diretriz lógica reducionista em ciência. Porém, ela nos ajuda a escolher entre várias hipóteses a serem verificadas, aquela que contém o menor número de afirmações não demonstradas.

2.5 Métodos de Classificação

Primeiramente, deve-se observar que a classificação é uma das tarefas mais populares da aprendizagem de máquina e ela visa encontrar uma função que mapeia um conjunto de registros em um conjunto de rótulos pré-definidos. Estes rótulos são denominados de classes. Uma vez obtida esta função, ela pode ser aplicada a novos registros para prever a classe em que estes se enquadrariam ou que se enquadram (QUINLAN, 1996). Em outras palavras, um método de classificação equivale à obtenção de regras baseado em observações de estados ambientais e transições. Por outro lado, um sistema de aprendizagem simbólica de máquina corresponde à automatização de um processo de aprendizagem. O algoritmo C4.5 é um exemplo de extrator de padrões de bases de dados.

2.5.1 Algoritmo C4.5

Este algoritmo realiza a tarefa de aprender de forma supervisionada (MITCHELL, 1997). Isto é, ele gera, a partir de um conjunto de exemplos previamente rotulados, um classificador no formato de árvore de decisão ou no formato regras de produção ordenadas. Tal classificador tem como objetivo representar os conhecimentos sobre determinado assunto (MITCHELL, 1997) (HAN & KAMBER, 2006). A aprendizagem baseada em árvore de decisão desperta o interesse por ser robusta quanto há ruídos nos dados (QUINLAN, 1996). A robusteza do algoritmo caracteriza-se também por operar tanto sobre valores discretos (e.g., o atributo *faixa etária* pode assumir um dos seguintes valores: *0-15, 16-30, 31-45, 45-60, 60-75*), quanto sobre valores contínuos (e.g., o atributo *volume de líquido* pode assumir valores reais como: 1.0, 2.5, 3.3, 4.2). Deve-se observar que os valores contínuos precisam ser discretizados para possibilitar a geração de regras de classificação.

A árvore de decisão é formada por um nó raiz e vários nós-folha, onde o nó raiz é o atributo que melhor separa por si só os exemplos a serem classificados. Cada nó da árvore representa um atributo. Os ramos são formados pelos valores dos atributos e as folhas são as classificações dos exemplos de acordo com os nós e os ramos. A árvore de decisão pode ser também representada por um conjunto de regras no formato *se-então*. (MITCHELL, 1997) (QUINLAN, 1993) (KOHAVI & JOHN, 1996). A classificação de uma determinada instância é feita à medida que a árvore é percorrida, sempre de modo descendente e *guloso*, o que significa que

uma árvore nunca retorna ao nível superior para testar novamente um determinado atributo. A Figura 1 ilustra um exemplo de classificação para prática de esporte através de árvore de decisão e o Quadro 1 mostra esta classificação no formato de regras (MITCHELL, 1997).

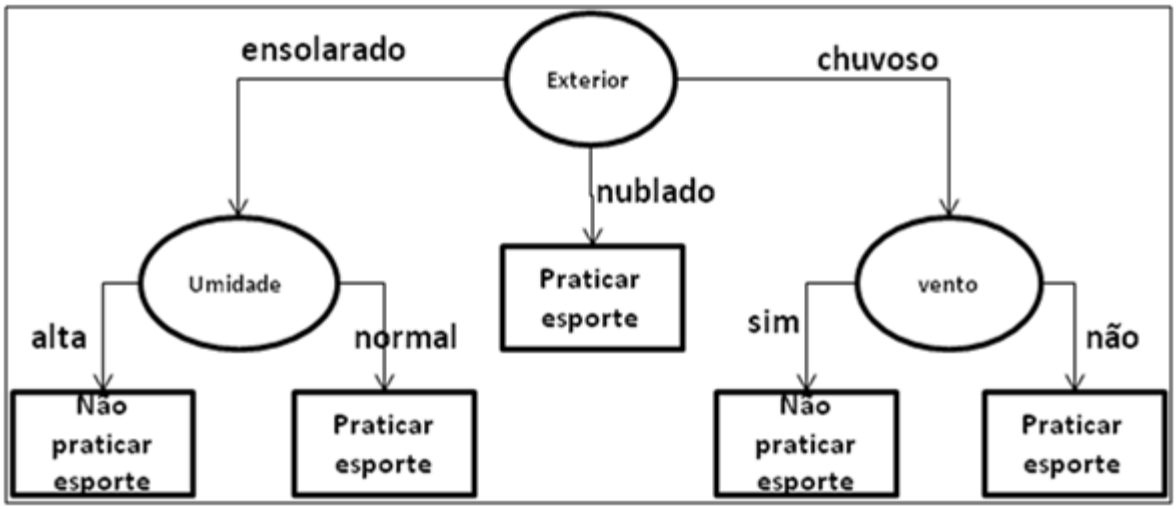

Figura 1. Exemplo de árvore de decisão gerada pelo C4.5.
Fonte: Mitchel, 1997.

Regra 1	**SE** o exterior está ensolarado **E** a umidade está alta **ENTÃO** não praticar esporte
Regra 2	**SE** o exterior está ensolarado **E** a umidade está normal **ENTÃO** praticar esporte
Regra 3	**SE** o exterior está nublado **ENTÃO** praticar esporte
Regra 4	**SE** o exterior está chuvoso **E** há vento **ENTÃO** não praticar esporte
Regra 5	**SE** o exterior está chuvoso **E** não há vento **ENTÃO** praticar esporte

Quadro 1. Conjunto de Regras geradas pelo C4.5.
Fonte: Mitchel, 1997.

A Tabela 2 ilustra uma base de dados, criados hipoteticamente para ilustrar os dados de uma clínica de estética, onde se descrevem três atributos previsores e um atributo meta. Tal base de dados apresenta as seguintes características: 14 exemplos, 4 atributos e 2 classes, onde *n* é o número de exemplos, *q* é o número de atributos e *m* é o numero de classes, lendo n=14, q= 4 e m=2.

Tabela 2. Conjunto de treinamento previamente ordenado.

Instâncias	ATRIBUTOS			
	PREVISORES			META
	Tipo de Queixa	Tipo de Cirurgia	Atividade Física	Alteração
1	Nula	Prótese Mamária	Não	Sim
2	Nula	Prótese Mamária	Não	Não
3	Nula	Prótese Mamária	Sim	Sim
4	Nula	Lipoaspiração	Não	Não
5	Nula	Lipoaspiração	Sim	Não
6	Constipação	Prótese Mamária	Não	Sim
7	Constipação	Prótese Mamária	Sim	Não
8	Constipação	Lipoaspiração	Sim	Sim
9	Constipação	Lipoaspiração	Não	Sim
10	Dor	Prótese Mamária	Não	Não
11	Dor	Prótese Mamária	Não	Sim
12	Dor	Prótese Mamária	Sim	Não
13	Dor	Lipoaspiração	Sim	Não
14	Dor	Lipoaspiração	Não	Não
LEGENDA	N=Nula C=Constipação D=Dor	PM=Prótese Mamária L=Lipoaspiração	S=Sim N=Não	S=Sim N=Não

O processo de criação da árvore de decisão, a partir de um conjunto contendo *n* exemplos de treinamento $T = \{t_1, ..., t_n\}$, onde cada *t* corresponde a um exemplo de treinamento, um conjunto contendo *q* atributos $A = \{a_1, ..., a_q\}$, onde cada *a* é um atributo, um conjunto contendo *m* classes $C = \{c_1, ..., c_m\}$, onde cada *c* é classe que rotula um exemplo *t*. Todo este processo ocorre de acordo com o algoritmo descrito a seguir na Figura 2:

```
Entradas: uma base de dados T
          um conjunto de atributos A
          um conjunto de classes C
Saída:    uma árvore de decisão D

1  procedimento C4.5( T, A, D, C )
3      cria um nó α de decisão e adiciona-o à árvore D
4      se todas as instâncias de T pertencem a uma classe c ∈ C
5      então
6          atribui ao nó α uma única folha com classe c ∈ C
7      senão
8          se A é um conjunto unitário
9          então
10             atribui ao nó α uma única folha identificando o valor
               mais comum do atributo preditor.
11         senão
12             calcula o ganho de informação de cada um dos atributos de A
13             se um dos atributos de A possui ganho de informação médio
               maior que os demais
14             então
15                 define at o atributo com maior ganho
16                 para cada valor v do atributo at faça
17                     adiciona uma sub-árvore d ao nó apontado por D
18                     define Tv como subconjunto de T, onde at = v
19                     C4.5( Tv, A-{at}, d, C )
20             senão
21                 atribui ao nó apontado por D uma única folha
           ligando o valor mais comum do atributo preditor.
22 fim.
```

Figura 2. Algoritmo C4.5.

Na seqüência será mostrado passa a passo os diferentes cálculos necessários para obter um classificador na forma de uma árvore de decisão (C4.5). O processo começa pela busca da informação que melhor codifica uma classificação de um determinado caso pertencente a um conjunto de treinamento T. Em termos matemáticos, esta informação é determinada pelo cálculo da entropia.

A entropia é calculada por meio da freqüência dos exemplos que resultam em casos positivos p_i presentes no conjunto de treinamento, multiplicando tal freqüência pelo logaritmo na base dois destes exemplos positivos, somando com os exemplos negativos. Estes últimos são obtidos da mesma forma. O resultado é a quantidade de informação necessária para codificar a classificação de um determinado caso pertencente ao conjunto de treinamento *T* (MITCHELL, 1997).

Na Tabela 3 estão ordenadas as fórmulas utilizadas para classificação pelo método C4.5 na forma de árvore de decisão.

Tabela 3. Fórmulas para Classificar no Formato de Árvore de Decisão.

$$Entropia(T) = -\sum_{c}^{C} p_c \log_2 p_c$$	(3)
$$Entropia(T_{a=v}) = -\sum_{c}^{C} p_{c,v} \log_2 p_{c,v}$$	(4)
$$\text{info}(\text{T}_\text{a}) = \sum_{\text{v} \in \text{dom(a)}} \frac{\lvert T_{a=v} \rvert}{T} \times Entropia(T_{a=v})$$	(5)
$$ganho(T_a) = entropia(T) - info(T_a)$$	(6)
$$\text{infoDiv}(T_{\text{ai}}) = -\sum_{i=1}^{n} \frac{\lvert T_{\text{ai}} \rvert}{\lvert T \rvert} \times \log_2 \left(\frac{\lvert T_{\text{ai}} \rvert}{\lvert T \rvert} \right)$$	(7)
$$\text{ganhoMedio}(T_{\text{a}}) = \frac{ganho(T_{\text{a}})}{\text{infoDiv}(T_{\text{a}})}$$	(8)

Para o conjunto de atributos *T,* o cálculo da entropia é dado pela Fórmula (3) da Tabela 3, onde p_c é a razão definida pelo número de instâncias que pertencem a classe *c* e número total de instâncias da base dados.

Tomando a Tabela 2 como o conjunto de treinamento T, tem-se a seguir o cálculo do valor da entropia; as classes são definidas por C={*sim, não*}. O valor 14 nos diferentes denominadores corresponde ao número de exemplos de T. Os valores 6 e 8 correspondem respectivamente aos números de exemplos que possuem como classe *sim* e *não*.

$$Entropia(T) = -\frac{6}{14} \times \log_2 \left(\frac{6}{14} \right) - \frac{8}{14} \times \log_2 \left(\frac{8}{14} \right) = 0.983$$

O cálculo da entropia para um dado atributo de predição *a* para o valor *v* é dado pela Fórmula (4) da Tabela 3, onde $p_{c,v}$ é a razão entre o número de instâncias que pertencem a cada classe $c \in$ C para um dado valor v do atributo de predição *a*. Caso uma dada razão seja nula, ou seja, quando nenhuma instância possuir o valor *v*, assume-se a entropia *nula*.

$$Entropia(T_{queixa=nula}) = \left(-\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times \log_2\left(\frac{3}{5}\right) \right) = 0{,}971$$

$$Entropia(T_{queixa=constipação}) = \left(-\frac{3}{4} \times \log_2\left(\frac{3}{4}\right) - \frac{1}{4} \times \log_2\left(\frac{1}{4}\right) \right) = 0{,}811$$

$$Entropia(T_{queixa=dor}) = \left(-\frac{4}{5} \times \log_2\left(\frac{4}{5}\right) - \frac{1}{5} \times \log_2\left(\frac{1}{5}\right) \right) = 0{,}722$$

$$Entropia(T_{cirurgia=lipoaspiração}) = \left(-\frac{2}{6} \times \log_2\left(\frac{2}{6}\right) - \frac{4}{6} \times \log_2\left(\frac{4}{6}\right) \right) = 0{,}918$$

$$Entropia(T_{cirurgia=prótesemamária}) = \left(-\frac{4}{8} \times \log_2\left(\frac{4}{8}\right) - \frac{4}{8} \times \log_2\left(\frac{4}{8}\right) \right) = 1$$

$$Entropia(T_{atividadefísica=sim}) = \left(-\frac{2}{6} \times \log_2\left(\frac{2}{6}\right) - \frac{4}{6} \times \log_2\left(\frac{4}{6}\right) \right) = 0{,}918$$

$$Entropia(T_{atividadefísica=não}) = \left(-\frac{4}{8} \times \log_2\left(\frac{4}{8}\right) - \frac{4}{8} \times \log_2\left(\frac{4}{8}\right) \right) = 1$$

O critério de ganho de informação é a maneira mais eficaz de dividir o conjunto de treinamento T e selecionar o atributo que irá dividir a informação do conjunto de treinamento. Tal critério consiste em medir a freqüência *freq(Cj, S)* na qual um conjunto qualquer de exemplos *S* pertence à mesma classe C_j, sendo que |S| é o número de exemplos em *S*. Tomando como exemplo a base de dados da Tabela 2, a freqüência de exemplos positivos (cujo valor da classe é "sim") é 6/14 ou 0,428. Já a freqüência de exemplos negativos (valor da classe "não") é 8/14 ou 0,571.

O critério de ganho de informação, mostrado na Fórmula (6) da Tabela 3, utilizado na seleção de um teste que maximiza o ganho de informação, é medido pela entropia menos a informação de *T* do atributo *a*. A informação do atributo *a* é medida de acordo com a Fórmula (5) da Tabela 3, onde o conjunto de dados de treinamento T é particionado em subconjuntos, de acordo com o conjunto de valores $v \in dom(a)$ que é domínio do atributo *a*, sendo $T_{a=v}$ o subconjunto formado por exemplos que possuem valor *v* para o atributo *a*.

Pode-se, desta forma, calcular o ganho para cada atributo da base de dados, escolhendo assim o atributo que melhor separa por si só as informações. O exemplo a seguir mostra o ganho calculado em detalhes para o atributo *Queixa* e também os valores do ganho de informação dos atributos *Cirurgia*, e *Atividade Física*.

$$info(T_{queixa}) = \frac{5}{14} \times Entropia(T_{queixa=nula}) + \frac{4}{14} \times Entropia(T_{queixa=constipação}) + \frac{5}{14} \times Entropia(T_{queixa=dor}) = 0.836$$

$$info(T_{cirurgia}) = \frac{6}{14} \times Entropia(_{cirurgia=lipoaspiração}) + \frac{8}{14} \times Entropia(_{cirurgia=prótesemamária}) = 0{,}964$$

$$info(T_{atividadefísica}) = \frac{6}{14} \times Entropia(T_{atividadefísica=sim}) + \frac{8}{14} \times Entropia(T_{atividadefísica=não}) = 0{,}964$$

Após calcular a informação para todos os atributos, o próximo passo é calcular o ganho de informação para os mesmos.

$$ganho(T_{queixa}) = entropia(T) - info(T_{queixa})$$
$$ganho(T, queixa) = 0{,}983 - 0{,}835 = 0{,}147$$

$$ganho(T_{cirurgia}) = entropia(T) - info(T_{cirurgia})$$
$$ganho(T_{cirurgia}) = 0{,}983 - 0{,}964 = 0{,}018$$

$$ganho(T_{\text{atividadefísica}}) = entropia(T) - info(T_{\text{atividadefísica}})$$
$$ganho(T_{\text{atividadefísica}}) = 0{,}983 - 0{,}964 = 0{,}018$$

O atributo *X* que possuir o valor mais alto de *ganho(X)* é então escolhido como sendo a raiz da árvore, neste caso o atributo *queixa*. O critério de ganho possui um *bias* que favorece atributos com muitos valores possíveis. Quando há diferentes funções que possuem valores em comum para um determinado exemplo, deve haver a seleção de uma função que se encaixe para

ambos os pontos *a priori*, tal seleção do valor *a priori* é chamada *bias*. Sem o valor do *bias,* um aprendizado útil se torna impossível (NILSSON, 1996). Tal fato ocorre quando há um atributo, que possui valor único para cada instância do conjunto de dados, ocorrendo uma divisão que gera inúmeros subconjuntos com apenas um valor possível para cada um deles, onde o valor de *info(S)* é igual a zero, portanto, mínimo. A correção para este problema consiste em ajustar os resultados possíveis, sendo que a potencial divisão da informação utiliza em seu cálculo a subtração do conjunto de treinamento *T* e de cada subconjunto *n*, conforme Fórmula (7) da Tabela 3. A divisão da informação resulta em informações úteis na classificação, eliminando assim atributos que podem por si só classificar amostras (i.e. o código de um paciente, que é único e possuí ganho de informação máximo. Apesar disto, não é recomendável sua utilização por não tornar genérica a árvore de decisão).

$$\text{infoDiv}(\text{T}_{\text{queixa}}) = \underbrace{-\frac{5}{14}\times\log_2\left(\frac{|5|}{|14|}\right)}_{nula} + \underbrace{\frac{4}{14}\times\log_2\left(\frac{|4|}{|14|}\right)}_{constipação} + \underbrace{\frac{5}{14}\times\log_2\left(\frac{|5|}{|14|}\right)}_{dor} = 1{,}577$$

$$\text{infoDiv}(T_{\text{atividadefisica}}) = \underbrace{-\frac{6}{14}\times\log_2\left(\frac{|6|}{|14|}\right)}_{sim} + \underbrace{\frac{8}{14}\times\log_2\left(\frac{|8|}{|14|}\right)}_{não} = 0{,}985$$

$$\text{infoDiv}(T_{cirurgia}) = \underbrace{-\frac{6}{14}\times\log_2\left(\frac{|6|}{|14|}\right)}_{lipoaspiração} + \underbrace{\frac{8}{14}\times\log_2\left(\frac{|8|}{|14|}\right)}_{prótese-mamária} = 0{,}985$$

A proporção de ganho de um atributo a_i é medida com base na informação relevante, Fórmula (8) da Tabela 3:

$$\text{ganhoMedio}(\text{T}_{\text{queixa}}) = \frac{\text{ganho}(\text{T}_{\text{queixa}})}{\text{infoDiv}(\text{T}_{\text{queixa}})} = \frac{0{,}147}{1{,}577} = 0{,}093$$

$$\text{ganhoMedio}(\text{T}_{\text{cirurgia}}) = \frac{\text{ganho}(\text{T}_{\text{cirurgia}})}{\text{infoDiv}(\text{T}_{\text{cirurgia}})} = \frac{0{,}018}{0{,}985} = 0{,}018$$

$$\text{ganhoMedio}(\text{T}_{\text{atividadefisica}}) = \frac{\text{ganho}(\text{T}_{\text{atividadefisica}})}{\text{infoDiv}(\text{T}_{\text{atividadefisica}})} = \frac{0{,}018}{0{,}985} = 0{,}018$$

A estrutura gerada a partir da execução do algoritmo C4.5 (implementação JRIP) sobre os dados da Tabela 2 pode ser encontrada na Figura 3. Cada valor entre parênteses corresponde ao número de exemplos bem classificados / número de exemplos mal classificados. O critério de ganho de informação que apresentar o maior valor é o atributo do nó raiz da árvore, neste caso o atributo Queixa. Após a identificação do nó raiz, este atributo recebe seus valores correspondentes, no caso, Nula, Constipação e Dor, e para cada valor será, novamente, calculado a entropia e o ganho de informação dos atributos Cirurgia e Atividade Física.

Onde para Queixa com o valor Nula, ou seja, para o subconjunto de instâncias 1, 2, 3, 4 e 5, obtem-se os resultados:

$$Entropia(\text{T}_{queia=>nula}) = -\frac{2}{5}\times\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\times\log_2\left(\frac{3}{5}\right) = 0{,}971$$

$$Entropia(\text{T}_{\text{q=>n, cirurgia=lipoaspiração}}) = \left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right) = 0$$

$$Entropia(\text{T}_{\text{q=>n, cirurgia=prótesemamária}}) = \left(-\frac{2}{3}\times\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\times\log_2\left(\frac{1}{3}\right)\right) = 0{,}918$$

$$Entropia(\text{T}_{\text{q=>n, atividadefísica=sim}}) = \left(-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\times\log_2\left(\frac{1}{2}\right)\right) = 1$$

$$Entropia(\text{T}_{\text{q=>n, atividadefísica=não}}) = -\left(\frac{1}{3}\times\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\times\log_2\left(\frac{2}{3}\right)\right) = 0{,}918$$

$$info(T_{q=>n,cirurgia}) = \frac{2}{5}\times Entropia(T_{q=>n,cirurgia=\text{lipoaspiração}}) +$$
$$\frac{3}{5}\times Entropia(T_{q=>ncirurgia=\text{prótesemamária}}) = 0{,}550$$

$$info(T_{\text{q=>n, atividadefísica}}) = \frac{2}{5}\times Entropia(T_{\text{q=>n,atividadefísica=sim}}) +$$
$$\frac{3}{5}\times Entropia(T_{\text{q=>n,atividadefísica=não}}) = 0{,}950$$

$$ganho(T_{q=>n,queixa=cirurgia}) = entropia(T_{queixa=>n}) - info(T_{q=>n,queixa=cirurgia})$$

$$ganho(T_{q=>n,queixa=cirurgia}) = 0{,}971 - 0{,}550 = 0{,}421$$

$$ganho(T_{q=>n,queixa=\text{atividadefísica}}) = entropia(T_{queixa=>n}) - info(T_{q=>n,queixa=\text{atividadefísica}})$$

$$ganho(T_{q=>n,queixa=\text{atividadefísica}}) = 0{,}971 - 0{,}950 = 0{,}021$$

$$\text{infoDiv}(T_{\text{q=>n,atividadefísica}}) = \underbrace{-\frac{2}{5} \times \log_2\left(\frac{|2|}{|5|}\right)}_{sim} + \underbrace{\frac{3}{5} \times \log_2\left(\frac{|3|}{|5|}\right)}_{não} = 0{,}970$$

$$\text{infoDiv}(T_{\text{q=>n,cirurgia}}) = \underbrace{-\frac{2}{5} \times \log_2\left(\frac{|2|}{|5|}\right)}_{lipoaspiração} + \underbrace{\frac{2}{5} \times \log_2\left(\frac{|2|}{|5|}\right)}_{prótese-mamária} = 0{,}970$$

$$\text{ganhoMedio}(\text{T}_{\text{q=>n,cirurgia}}) = \frac{\text{ganho}(\text{T}_{\text{q=>n,queixa=cirurgia}})}{\text{infoDiv}(\text{T}_{\text{q=>n,cirurgia}})} = \frac{0{,}421}{0{,}970} = 0{,}434$$

$$\text{ganhoMedio}(\text{T}_{\text{q=>n,atividadefisica}}) = \frac{\text{ganho}(\text{T}_{\text{q=>n,queixa=atividadefisica}})}{\text{infoDiv}(\text{T}_{\text{q=>n,atividadefisica}})} = \frac{0{,}021}{0{,}970} = 0{,}022$$

De acordo com os resultados dos cálculos o atributo *Cirurgia* apresentou um maior ganho de informação, portanto, este atributo será o nó de decisão para o valor de queixa Nula. Em seguida, será continuada a seqüência do atributo *Cirurgia*, para depois retornar aos outros valores do atributo Queixa.

Para o atributo Cirurgia com valor de Prótese Mamária, que é o subconjunto de instâncias 1, 2 e 3; será incluído uma folha com o valor mais comum do atributo preditor, que no caso será uma folha com valor Sim.

Enquanto para o atributo Cirurgia com o valor de Lipoaspiração, referentes as instâncias 4 e 5, será inserida uma folha com o valor comum a todas as instâncias deste subconjunto, o valor Não.

Concluído a seqüência para o atributo Cirurgia, retorna-se ao atributo Queixa e calcula-se a entropia e o ganho de informação para os atributos Cirurgia e Atividade Física.

Onde para o atributo Queixa com valor de Constipação, ou subconjunto 6, 7, 8 e 9, encontramos:

$$Entropia(T_{queixa=constipação}) = -\frac{3}{4}\times\log_2\left(\frac{3}{4}\right)-\frac{1}{4}\times\log_2\left(\frac{1}{4}\right)=0{,}811$$

$$Entropia(T_{\text{q=>c,cirurgia=lipoaspiração}}) = \left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right)=0$$

$$Entropia(T_{\text{q=>c,cirurgia=prótesemamária}}) = \left(-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)\right)=1$$

$$Entropia(T_{\text{q=>c,atividadefísica=sim}}) = \left(-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\times\log_2\left(\frac{1}{2}\right)\right)=1$$

$$Entropia(T_{\text{q=>c,atividadefísica=}não}) = \left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right)=0$$

$$info(T_{q=>c,cirurgia}) = \frac{2}{4}\times Entropia(T_{q=c,cirurgia=lipoaspiração})+$$
$$\frac{2}{4}\times Entropia(T_{q=>c,cirurgia=\text{prótesemamária}}) = 0{,}500$$

$$info(T_{\text{q=>c,atividadefísica}}) = \frac{2}{4}\times Entropia(T_{\text{q=>c,atividadefísica=sim}})+$$
$$\frac{2}{4}\times Entropia(T_{\text{q=>c,atividadefísica=}não}) = 0{,}500$$

$$ganho(T_{q=>c,queixa=cirurgia}) = entropia(T_{queixa=constipação}) - info(T_{q=>c,cirurgia})$$
$$ganho(T_{q=>c,queixa=cirurgia}) = 0{,}811 - 0.500 = 0{,}311$$

$$ganho(T_{q=>c,queixa=\text{atividadefísica}}) = entropia(T_{queixa=constipação}) - info(T_{q=>c,\text{atividadefísica}})$$

$$ganho(T_{q=>c,queixa=\text{atividadefísica}}) = 0{,}811 - 0.500 = 0{,}311$$

$$\text{infoDiv}(T_{\text{q=>c,atividadefísica}}) = \underbrace{-\frac{2}{4}\times\log_2\left(\frac{|2|}{|4|}\right)}_{sim} \underbrace{+\frac{2}{4}\times\log_2\left(\frac{|2|}{|4|}\right)}_{não} = 1$$

$$\text{infoDiv}(T_{q=>c,\text{cirurgia}}) = \underbrace{-\frac{2}{4}\times\log_2\left(\frac{|2|}{|4|}\right)}_{lipoaspiração} + \underbrace{\frac{2}{4}\times\log_2\left(\frac{|2|}{|4|}\right)}_{prótese-mamária} = 1$$

$$\text{ganhoMedio}(T_{\text{q=>c,cirurgia}}) = \frac{\text{ganho}(T_{\text{q=>c,queixa=cirurgia}})}{\text{infoDiv}(T_{\text{q=>c,cirurgia}})} = \frac{0{,}311}{1} = 0{,}311$$

$$\text{ganhoMedio}(T_{\text{q=>c,atividadefisica}}) = \frac{\text{ganho}(T_{\text{q=>c,queixa=atividadefisica}})}{\text{infoDiv}(T_{\text{q=>c,atividadefisica}})} = \frac{0{,}311}{1} = 0{,}311$$

Como podemos observar o valor do ganho de informação para os atributos cirurgia e atividade física são os mesmos, impossibilitando a escolha de um atributo por maior ganho de informação, será, portanto, realizado um poda, a fim de simplificar a árvore, incluindo uma folha com o valor mais freqüente para o atributo Queixa com valor Constipação, o valor Sim.

E para o atributo Queixa com valor de Dor os valores de entropia e ganho de informação, são os seguintes:

$$Entropia(T_{queixa=dor}) = -\frac{1}{5}\times\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\times\log_2\left(\frac{4}{5}\right) = 0{,}721$$

$$Entropia(T_{\text{q=>d,cirurgia=lipoaspiração}}) = \left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right) = 0$$

$$Entropia(T_{\text{q=>d,cirurgia= prótesemamária}}) = \left(-\frac{2}{3}\times\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\times\log_2\left(\frac{1}{3}\right)\right) = 0{,}918$$

$$Entropia(T_{\text{q=>d,atividadefísica=sim}}) = \left(-\frac{2}{2}\times\log_2\left(\frac{2}{2}\right)\right) = 0$$

$$Entropia(T_{\text{q=>d,atividadefísica= não}}) = \left(-\frac{1}{3} \times \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \times \log_2\left(\frac{2}{3}\right)\right) = 0{,}918$$

$$info(T_{q=>d,cirurgia}) = \frac{2}{5} \times Entropia(T_{q=>d,cirurgia=\text{lipoaspiração}}) + \frac{3}{5} \times Entropia(T_{\text{q=>d,cirurgia= prótesemamária}}) = 0{,}551$$

$$info(T_{\text{q=>d,atividadefísica}}) = \frac{2}{5} \times Entropia(T_{\text{q=>d,atividadefísica=sim}}) + \frac{3}{5} \times Entropia(T_{\text{q=>d,atividadefísica=não}}) = 0{,}551$$

$$ganho(T_{q=>d,queixa=cirurgia}) = entropia(T_{queixa=dor}) - info(T_{q=>d,queixa=cirurgia})$$

$$ganho(T_{q=>d,queixa=cirurgia}) = 0{,}721 - 0{,}550 = 0{,}171$$

$$ganho(T_{q=>d,queixa=\text{atividadefísica}}) = entropia(T_{queixa=dor}) - info(T_{q=>d,queixa=\text{atividadefísica}})$$

$$ganho(T_{q=d,queixa=\text{atividadefísica}}) = 0{,}721 - 0{,}550 = 0{,}171$$

$$\text{infoDiv}(T_{q=d,\text{atividadefisica}}) = \underbrace{-\frac{2}{5} \times \log_2\left(\frac{|2|}{|5|}\right)}_{sim} \underbrace{+\frac{3}{5} \times \log_2\left(\frac{|3|}{|5|}\right)}_{não} = 0{,}971$$

$$\text{infoDiv}(T_{q=>d,\text{cirurgia}}) = \underbrace{-\frac{2}{5} \times \log_2\left(\frac{|2|}{|5|}\right)}_{lipoaspiração} + \underbrace{\frac{3}{5} \times \log_2\left(\frac{|3|}{|5|}\right)}_{prótese-mamária} = 0{,}971$$

$$\text{ganhoMedio}(\text{T}_{\text{q=>d,cirurgia}}) = \frac{\text{ganho}(\text{T}_{\text{q=>d,queixa=cirurgia}})}{\text{infoDiv}(\text{T}_{\text{q=>d,cirurgia}})} = \frac{0{,}171}{0{,}971} = 0{,}176$$

$$\text{ganhoMedio}(\text{T}_{\text{q=>d,atividadefisica}}) = \frac{\text{ganho}(\text{T}_{\text{q=>d,queixa=atividadefisica}})}{\text{infoDiv}(\text{T}_{\text{q=d,atividadefisica}})} = \frac{0{,}171}{0{,}971} = 0{,}176$$

Como no exemplo anterior, os atributos *cirurgia* e *atividade física* apresentaram o mesmo ganho de informação para o subconjunto de *queixa* com valor *dor*, optando-se por incluir uma

folha com o valor mais comum nestas instâncias: 10, 11, 12, 13 e 14, que é o valor *não*, ilustrado na Figura 3.

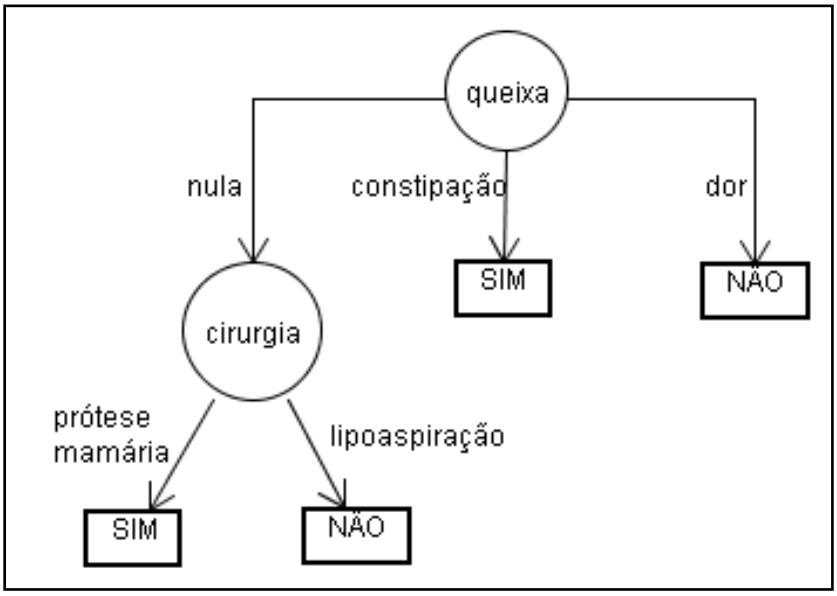

Figura 3. Exemplo de árvore de decisão.

A mesma árvore pode ser vista no formato de regras (Figura 4). Por exemplo, a Regra 2, pode ser lida da seguinte forma: **SE** não houve queixa **e a** cirurgia realizada foi lipoaspiração **ENTÃO** tem-se a indicação que não haverá alteração.

```
REGRA 1:
   SE QUEIXA = NULA E CIRURGIA = PRÓTESE MAMÁRIA
   ENTAO ALTERAÇÃO = SIM
REGRA 2:
   SE QUEIXA = NULA E CIRURGIA = LIPOASPIRAÇÃO
   ENTAO ALTERAÇÃO = NÃO
REGRA 3:
   SE QUEIXA = CONSTIPAÇÃO
   ENTAO ALTERAÇÃO = SIM
REGRA 4:
   SE QUEIXA = DOR
   ENTAO ALTERAÇÃO = NÃO
```

Figura 4. Árvore de decisão no formato de regras.

Para exemplificar o processo de classificação no método C4.5, na Tabela 4 será assumido o recebimento de três instâncias a serem classificadas.

Tabela 4. Instâncias a serem classificadas pelo método C4.5.

Instâncias	ATRIBUTOS			
	PREVISORES			META
	Tipo de Queixa	Tipo de Cirurgia	Atividade Física	Alteração
a1	Nula	Prótese Mamária	Não	?
a2	Constipação	Prótese Mamária	Não	?
a3	Dor	Lipoaspiração	Sim	?

A classificação da instância a_1 resultaria na indicação do valor *sim* para o atributo meta alteração, enquanto que para a instância a_2 resultaria na indicação do valor *sim* para o atributo meta alteração, e a instância a_3 resultaria na indicação do valor *não* para o atributo meta alteração.

A Figura 5 mostra por meio de uma matriz de confusão, os acertos e erros de classificação, tomando como conjunto de teste os mesmos usados para gerar o classificador, a saber: Tabela 2.

Nas linhas da matriz estão representados os valores previstos pelo classificador e as colunas os valores atuais de cada classe. Os valores previstos são representados na matriz conforme seus valores reais sejam estes corretos ou incorretos.

Na Figura 5, cinco exemplos do tipo *a* foram classificadas corretamente como *a*, e seis exemplos do tipo *b* foram classificadas corretamente, valores ilustrados na diagonal de previsões corretas. Um exemplo do tipo *a* foi classificado como sendo do tipo *b*, caracterizando assim erro na classificação, dois exemplos do tipo *b* foram classificados como sendo do tipo *a*, também caracterizando assim erro na classificação, valores estes representados na diagonal de previsões incorretas (ROMÃO, 2002).

```
a b ← classificado como
5 1 | a = sim
2 6 | b = não
```

Figura 5. Matriz de confusão.

A taxa de previsão correta do classificador é calculada através do percentual da soma dos valores da diagonal de previsões corretas dividido pelo total dos valores das previsões. E a taxa de previsão incorreta é 100% menos a taxa de previsão correta. Portanto, na matriz de confusão da Figura 5 a taxa de exemplos classificados corretamente foi de 78.57% e incorretamente foi de 21.42%.

Para cada regra gerada pelo classificador, um fator de suporte e um fator de confiança são atribuídos (HAN, et al., 2006).

> O fator de suporte é igual ao número de registros contendo a característica X e Y dividido pelo número total de registros. A confiança é igual ao número de registros com X e Y dividido pelo número de registros com X. Por exemplo, houve 200 atendimentos em uma clínica de estética. Em 30 destes 200 atendimentos a queixa foi sobrepeso e destes 30, 12 deles culminaram em cirurgias de redução de estomago. Assim a regra "SE queixa foi sobrepeso ENTÃO cirurgia de redução de estomago" teria um suporte de 30/200 = 15% e confiança de 12/30 = 40%.

Até aqui, o método apresentado permite a geração de classificadores simbólicos simples a partir de uma base de dados. Ou seja, ele não gera vários classificadores para diferentes amostras de uma mesma base de dados.

Teoricamente, a eficiência de classificadores simples poderia melhorar, por meio, da combinação de diferentes classificadores obtidos a partir de diferentes amostras de uma mesma base de dados. Há dois métodos básicos de combinação de classificadores: BAGGING e BOOSTING. A combinação de diferentes classificadores visa obter uma melhor taxa de acerto do que a obtida pela aplicação de um único classificador (WITTEN & FRANK, 2000).

2.5.2 Método BAGGING

O método BAGGING combina k classificadores a partir de k amostras da base original. As amostras devem ter o mesmo tamanho do conjunto original e para cada uma das k amostras um classificador é obtido (BREIMAN, 1996). Cabe salientar que estas amostras são sorteadas com reposição e distribuídas de forma uniforme, pois, elas deverão ter o mesmo número de instâncias que o conjunto original.

A Figura 6 descreve a implementação do algoritmo adaptado de Breiman que toma como entrada um conjunto de treinamento T e devolve um conjunto C de classificadores, onde cada membro tem igual peso. As linhas cinco e seis correspondem respectivamente a duas chamadas de funções, sendo a primeira para a obtenção de uma amostra A_i, com reposição, de T e a segunda para a obtenção de um classificador a partir da amostra A_i. A obtenção de tal classificador pode ser feita por meio da invocação do C4.5; este último retorna uma árvore de decisão. O processo de

amostragem e treinamento repete-se por k vezes. Em outras palavras, cada classificador de C é obtido a partir de uma amostra diferente A_i de T (GRANDVALET, 2004).

```
   Entradas: uma base de dados T com n instâncias
             um número de classificadores k a serem gerados
   Saída:    um conjunto de k classificadores C

1  função f(T, n, k): conjunto de classificadores
2  inicio
3     C ← Ø;                 {C é ajustado inicialmente como vazio}
4     para i de 1 até k faça
5        A_i ← amostrar(T, n);   {obtenção da amostra A_i com reposição}
6        C_i ← treinar(A_i, n);   {obtenção do classificador C_i a partir de A_i }
7        C ← C ∪ C_i;         {adição do classificador C_i ao conjunto C}
8     fimpara
9     retorne C;          {retorna o conjunto dos k classificadores}
10 fim.
```

Figura 6. Método BAGGING. Adaptado de BREIMAN (1996).

O método BAGGING é particularmente interessante, pois quando o algoritmo de aprendizagem de máquina aplicado em uma base de dado possui um comportamento instável; uma pequena mudança na base de dados gera classificadores substancialmente diferentes. Logo, um único classificador não é capaz de oferecer uma resposta confiável para todas as situações; um classificador composto pode ter maior chance de acerto. Em outras palavras, o método BAGGING permite a obtenção de modelos que melhoram a taxa de acerto se comparado a método de obtenção de modelos simples (e.g., C4.5), mas, perde-se em termos de facilidade de interpretação (TAN; STEINBACH; KUMAR, 2006).

A Figura 7 mostra esquematicamente as etapas do método BAGGING, a saber:

- gerar diferentes amostras a partir da mesma base de dados de treinamento, onde as amostras são geradas com reposição, distribuição uniforme e tamanhos idênticos;
- obter um classificador (e.g., uma árvore de decisão) para cada amostra;
- agrupar classificadores obtidos individualmente na etapa anterior em uma unidade de decisão por consenso; e
- classificar por consenso, onde elege-se a classificação mais popular dentre os classificadores individuais por meio de uma votação simples.

É fundamental salientar que o método BAGGING, assim como os demais métodos de combinação de classificadores, gera uma unidade de decisão composta formada por um conjunto de classificadores simples. Nestes termos, o produto final do método BAGGING não é um único

classificador, assim como não é a fusão de k árvores de decisão em uma árvore única. E cada nova instância a ser classificada é avaliada pela unidade de decisão composta, cuja classificação da instância é a escolha realizada pela maioria dos k classificadores.

Em termos práticos, o processo de classificação do método BAGGING equivale à situação, onde um gestor humano que se cerca de k consultores e toma a decisão baseada na votação feita por estes consultores (WITTEN; FRANK, 2000). Teoricamente, o resultado de um processo decisório consensual obtido de k consultores ou de k classificadores tende a apresentar uma taxa de acerto maior do que a taxa de acerto de cada um individualmente.

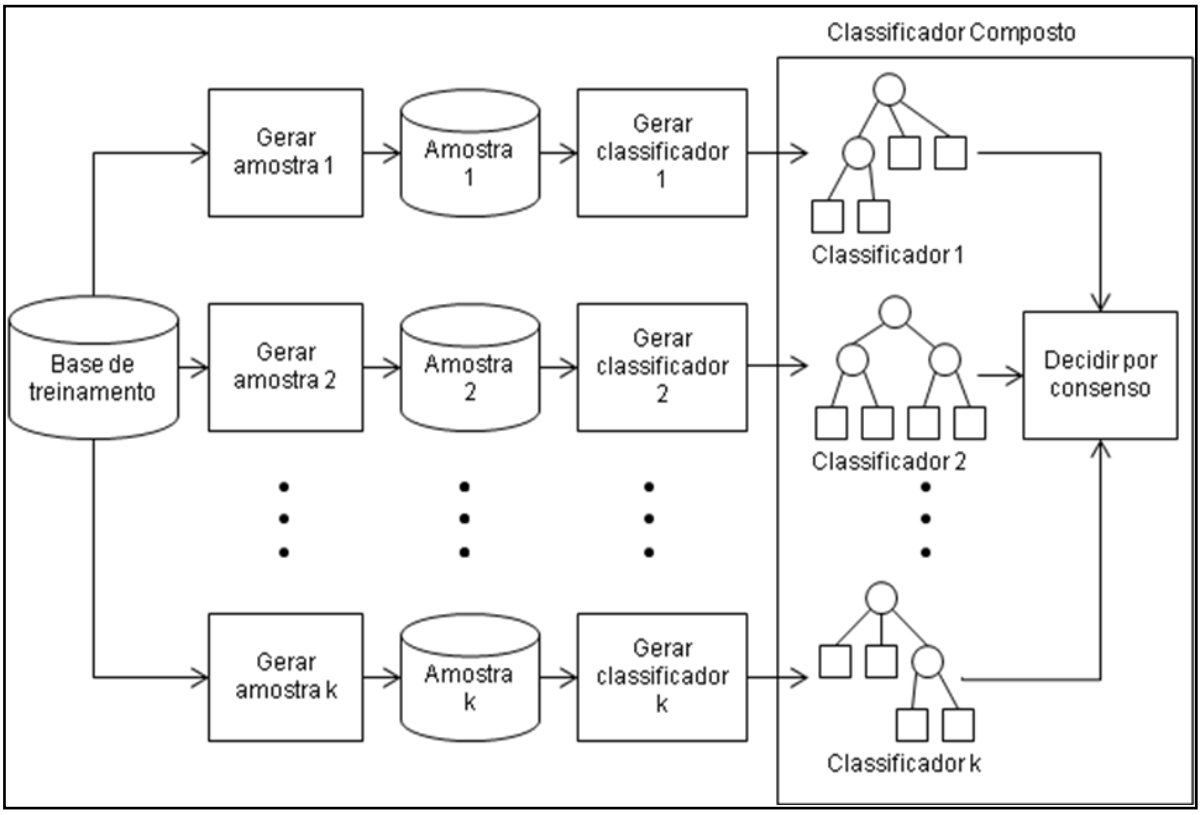

Figura 7. Esquema geral de aplicação do resultado do método BAGGING.

A utilização do resultado do método BAGGING sobre a base de treinamento representada pela Tabela 2 dar-se-á no seguinte cenário:

- a base de treinamento T possui 14 exemplos;
- três amostras geradas com reposição, conforme esquema Tabela 2; e
- a unidade de decisão composta possui 3 classificadores.

Para facilitar, a visualização das amostras da Tabela 5 as instâncias foram ordenadas; a coluna P indica a posição da instância na tabela de treinamento original. Como na base original, o atributo meta corresponde a coluna mais a direita de cada da amostra, neste caso o atributo meta é alteração. A geração destas amostras corresponde à execução da linha cinco do algoritmo da Figura 6.

Tabela 5. Amostras com reposição e exemplos com pesos idênticos.

AMOSTRA I					AMOSTRA II					AMOSTRA III				
P	**TQ**	**TC**	**AF**	**A**	**P**	**TQ**	**TC**	**AF**	**A**	**P**	**TQ**	**TC**	**AF**	**A**
1	N	PM	N	S	1	N	PM	N	S	2	N	PM	N	N
2	N	PM	N	N	3	N	PM	S	S	2	N	PM	N	N
3	N	PM	S	S	3	N	PM	S	S	2	N	PM	N	N
4	N	L	N	N	6	C	PM	N	S	3	N	PM	S	S
5	N	L	S	N	6	C	PM	N	S	3	N	PM	S	S
6	C	PM	N	S	7	C	PM	S	N	3	N	PM	S	S
6	C	PM	N	S	8	C	L	S	S	4	N	L	N	N
7	C	PM	S	N	9	C	L	N	S	5	N	L	S	N
7	C	PM	S	N	10	D	PM	N	N	6	C	PM	N	S
9	C	L	N	S	10	D	PM	N	N	8	C	L	S	S
10	D	PM	N	N	11	D	PM	N	S	12	D	PM	S	N
10	D	PM	N	N	13	D	L	S	N	14	D	L	N	N
10	D	PM	N	N	13	D	L	S	N	14	D	L	N	N
14	D	L	N	N	14	D	L	N	N	14	D	L	N	N

LEGENDA:
N:Nula, C:Constipação, D:Dor
TQ:Tipo de Queixa, TC:Tipo de Cirurgia, AF:Atividade Física, A:Alteração

A Figura 8 mostra um conjunto de três classificadores, sendo que os classificadores C1, C2 e C3 foram obtidos, respectivamente, a partir da Amostra I, II e III da Tabela 5. O software usado para gerar os classificadores foi o J48, implementado no WEKA (2008) Para facilitar a leitura, os classificadores C1, C2 e C3 geradas pelo J48 no formato de árvores de decisão foram re-escritas na forma de três conjuntos de regras ordenadas.

```
CLASSIFICADOR C1
   SE TQ = N E TC = PM ENTÃO A = S (3.0/1.0)
   SENÃO SE TQ = N E TC = L ENTÃO A = N (2.0)
   SENÃO SE TQ = C E AF = N ENTÃO A = S (3.0)
   SENÃO SE TQ = C E AF = S ENTÃO A = N (2.0)
   SENÃO SE TQ = D ENTÃO A = N (4.0)
PESO DE C1: 1

CLASSIFICADOR C2
   SE TQ = N ENTÃO A = S (3.0)
   SENÃO SE TQ = C ENTÃO A = S (5.0/1.0)
   SENÃO SE TQ = D ENTÃO A = N (6.0/1.0)
PESO DE C2: 1

CLASSIFICADOR C3
   SE TQ = N E AF = N ENTÃO A = N (4.0)
   SENÃO SE TQ = N E AF = S ENTÃO A = S (4.0/1.0)
   SENÃO SE TQ = C ENTÃO A = S (2.0)
   SENÃO SE TQ = D ENTÃO A = N (4.0)
PESO DE C3: 1
```

Figura 8. Classificador composto – exemplo BAGGING.

A Tabela 6 possui três instâncias cujas predições são desconhecidas. Para conhecê-las, o processo consiste em submeter cada uma delas aos três classificadores da Figura 8. Cada classificador informa como resposta um valor de predição. O valor de predição mais freqüente é a classificação.

Tabela 6. Instâncias a serem classificadas pelo método BAGGING.

Instância	Tipo de Queixa	Tipo de Cirurgia	Atividade Física	Alteração
a_1	Nula	Prótese Mamária	Não	?
a_2	Constipação	Prótese Mamária	Não	?
a_3	Dor	Lipoaspiração	Sim	?

A classificação da instância a_1 resulta na resposta **Sim**, para o Classificador *C1*; **Sim**, para o Classificador *C2*; e **Não**, para o Classificador *C3*. Conseqüentemente, a resposta **Sim** será considerada, vista que a mesma obteve dois votos contra um. A classificação da instância a_2 resulta na resposta **Sim** para todos os classificadores; assim o resultado da votação foi por unanimidade. A mesma situação acontece para a instância a_3, cuja resposta é **Não** por unanimidade.

Deve-se observar que o peso/importância do voto de cada classificador, na abordagem BAGGING, é um. Entretanto, no método BOOSTING, a seguir, pode-se assumir que o voto de cada classificador tenha peso/importância diferenciada, de forma análoga, ao que pode ocorrer entre diferentes consultores opinando sobre um tema.

2.5.3 Método BOOSTING

O método BOOSTING combina k classificadores a partir de k amostras da base original. Ele foi desenvolvido com o intuito de oferecer classificadores mais eficientes. O AdaBoost (Adaptative BOOSTING) é o representante mais comum do método BOOSTING (FREUND; SCHAPIRE, 1996). Deve-se, entretanto, observar que duas distinções entre os métodos BAGGING e BOOSTING, a primeira concerne à forma com que as amostras são geradas e a segunda à maneira com que os resultados dos classificadores são combinados (LOPES, 2007).

Para uma base de treinamento T com *n* instâncias, ao invés de gerar amostras de tamanho *n,* assumindo sempre uma distribuição uniforme (probabilidade *1/n*) sobre as instâncias de T, levam-se em conta as amostras já geradas de forma a alterar a distribuição das próximas amostras (BAUER; KOHAVI, 1999). Este processo pode ser dividido em diferentes passos, sendo que cada passo compreende: (i) a geração de uma amostra; (ii) a geração do classificador associado à amostra; (iii) a aplicação deste classificador a base de treinamento original; (iv) a análise de eficiência geral deste classificador; (v) a eficiência do classificador frente a cada instância da base de treinamento individualmente; e, finalmente, (vi) a alteração das probabilidades para a geração da amostra no próximo passo (FREUND; SCHAPIRE, 1996).

De forma mais específica, os passos do processo para a obtenção de um classificador composto BOOSTING, consistem em:

- gerar a primeira amostra A_1 assumindo uma distribuição uniforme e com reposição, ou seja, todas as instâncias da base de treinamento T têm a mesma probabilidade (1/n) de serem incluídas na primeira amostra gerada A_1; em seguida,
- gerar um classificador C_1 para a amostra A_1 e aplicar o classificador C_1 a base de treinamento original T;
- de acordo com o resultado apresentado pelo classificador gerado C_1, diminuir a probabilidade de serem incluídas, na próxima amostra, as instâncias de T que foram corretamente classificadas e aumentar a probabilidade das instâncias que foram incorretamente classificadas;
- gerar a segunda amostra A_2 a partir da base de treinamento original T, levando em conta as novas probabilidades de cada instância ser escolhida;
- gerar um classificador C_2 para a segunda amostra A_2 e aplicar o classificador C_2 a base de treinamento T, e mais uma vez;

- diminuir a probabilidade das instâncias em T classificadas corretamente e aumentar a probabilidade das instâncias em T classificadas incorretamente;
- repetir este processo de geração de amostras, classificadores e alteração de pesos até serem gerados as *k*-ésimas amostras e os *k*-ésimos classificadores.

É importante frisar que as instâncias classificadas incorretamente por um classificador têm os seus pesos incrementados. Esta lógica faz com que as instâncias com os maiores pesos sejam preferencialmente incluídas nas amostras subseqüentes. Desta forma, tende-se a gerar amostras mais especializadas vis-à-vis a obtenção de modelos mais eficientes para classificar determinadas instâncias.

Analogamente ao método BAGGING, o método BOOSTING é um classificador composto. Entretanto, a importância de cada classificador pode ser significativamente diferente. O método BOOSTING usa um mecanismo de votação ponderada, segundo um índice de importância, para determinar uma classificação. Para isto é necessário que o processo de geração de cada classificador memorize a eficiência do seu produto frente à base de treinamento (BARLETT; FREUND; LEE; SCHAPIRE, 1997). Este processo não adiciona custos ao método. Os classificadores são testados a cada passo para gerar amostras com probabilidades distintas.

A Figura 9, de forma análoga a Figura7, mostra os passos do método BOOSTING. A principal particularidade é a atualização dos pesos das instâncias na base de treinamento classificadas incorretamente. Por exemplo, cada instância de T é submetida ao Classificador 1; toda instância mal classificada terá seu peso incrementado; a base de treinamento T' é mesma base de treinamento T, exceto que as instâncias classificadas incorretamente pelo Classificador 1 terão seus pesos incrementados. Este processo se repete até a geração do Classificador k. Nestes termos, os classificadores possuem importâncias diferentes; logo, a classificação de uma nova instância é ponderada de acordo com a importância de cada classificador, cujo detalhamento do cálculo será mostrado na seqüência; ou seja, a decisão é dada por meio de uma votação ponderada para as respostas dos classificadores para obter o resultado do classificador composto.

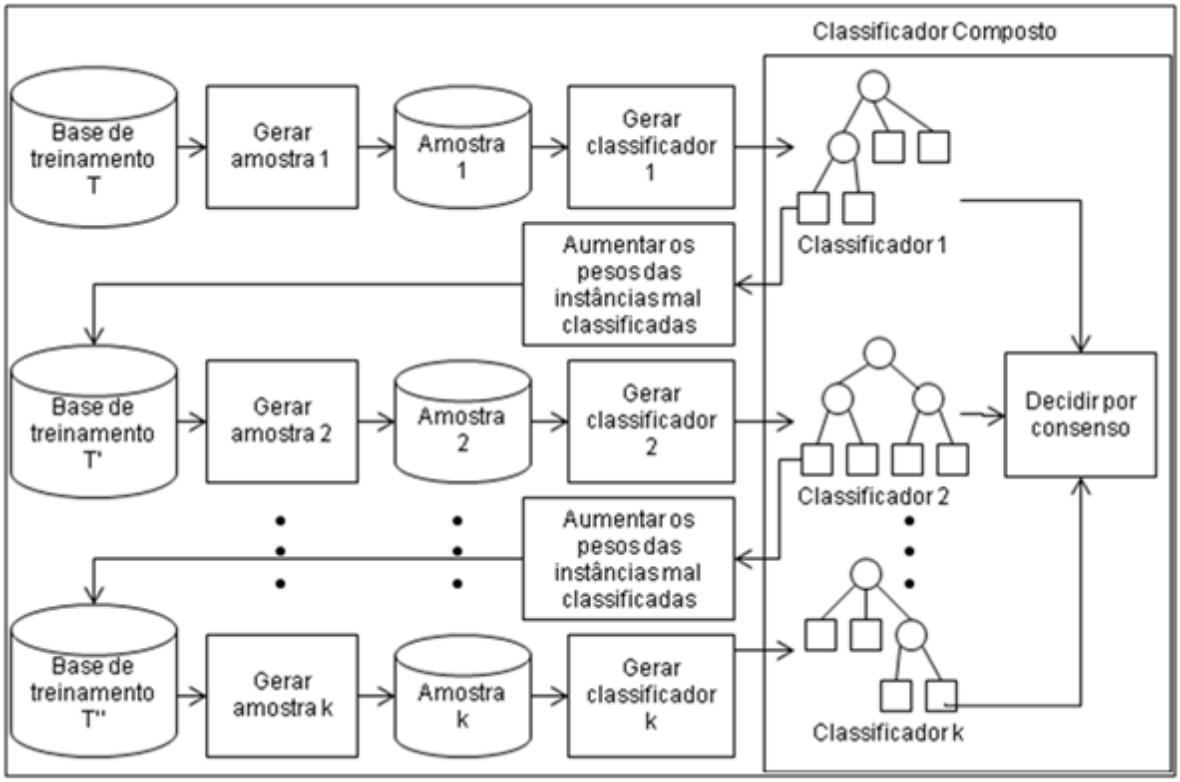

Figura 9. Esquema geral de funcionamento do método BOOSTING.

A realização prática desta política de variação nas probabilidades das instâncias para as amostras é feita por meio da definição de um vetor de pesos (*w*). Este último associa um valor real $w_j \in [0, 1]$ para cada instância da base de treinamento T. Este vetor é modificado após a geração de cada novo classificador. O elemento w_j deste vetor está associado a *j*-ésima instância da base de treinamento T. Além da probabilidade da *j*-ésima instância ser incluída na próxima amostra, este peso w_j também significa o quanto a *j*-ésima instância é importante na obtenção dos próximos classificadores. Note que ambos os significados tem uma semântica muito semelhante, mas como será mostrado a seguir, cada um destes significados representa um uso distinto dos pesos w_j na mecânica do método BOOSTING.

O processo básico é o seguinte: uma amostra *Ai* é gerada, levando em conta os atuais pesos das instâncias em T; um classificador *Ci* é gerado a partir de *Ai*; as instâncias de T são submetidas ao classificador *Ci;* calcula-se a sua taxa de erro ($\varepsilon^{(i)}$) do classificador *Ci* e sua importância ($\alpha^{(i)}$); e recalcula-se o peso w_j de cada instância da base de treinamento T. Tal processo é ilustrado na forma de um algoritmo (Figura10). Este algoritmo é uma ligeira adaptação da versão original de Freund; Schapire (1996).

```
   Entradas: uma base de dados T com n instâncias
             um número de classificadores k a serem gerados
   Saída:    um conjunto de k classificadores C e suas importâncias α

1  função g(T, n, k): conjunto de classificadores e suas importâncias
2  inicio
3     para i de 1 até n faça  {ajustar os valores dos pesos a 1/n}
4        w[ i ] ← 1/n;
5     C ← Ø;                  {C é ajustado inicialmente como vazio}
6     para i de 1 até k faça
7        A_i ← amostrar(T,n,w);  {obter amostra A_i com reposição e com base em w}
8        C_i ← treinar(A_i,n); {obter classificador C_i a partir de A_i }
9        B ← testar(C_i,T,n);    {obter os índices das instâncias mal classificadas}
10       ε^(i) ← erro(C_i,T,n);  {calcular a taxa de erro do classificador Ci}
11       α^(i) ← importância(ε^(i)); {calcular a importância do classificador Ci}
12       se ε^(i) > 0,5 então    {ajustar os pesos das instâncias aos valores iniciais}
13          para i de 1 até n faça
14             w[ i ] ← 1/n;
15       senão                   {ajustar os pesos das instâncias mal classificadas}
16          para i de 1 até | B | faça
17             w[ B[ i ] ] ← w[ B[ i ] ] * α^(i);
18          C ← C ∪ <C_i, α^(i)> {adição do classificador <C_i, α^(i)> ao conjunto C}
19    fimpara
20    retorne C;           {retorna o conjunto dos k classificadores}
21 fim.
```

Figura 10. Método BOOSTING – implementação AdaBOOSTING.

Das linhas 3 a 4 são realizadas os ajustes das probabilidades de cada instância de T. Para tal cada entrada do vetor auxiliar *w* é ajustado em 1/*n*. Pode-se destacar que em função da probabilidade de cada instância de T ser a mesma, a geração da primeira amostra (linha 7) é igual a geração feita pelo método BAGGING; as instâncias a serem incluídas em tal amostra possuem uma distribuição de probabilidade uniforme. Na linha 5, o conjunto de classificadores C e suas respectivas importâncias ajustado para vazio, conjunto este que será preenchido conforme o algoritmo é percorrido.

A Fórmula (9) da Tabela 7 expressa o cálculo da taxa de erro do classificador C_i (linha 10). Ele é realizado, levando em consideração: (i) o número de instâncias da amostra *n*; (ii) o peso de cada uma das instâncias w_j; e (iii) a verificação da efetividade do classificador C_i em classificar corretamente cada uma das instâncias.

A Fórmula (10) da Tabela 7 expressa o cálculo da importância do classificador, correspondente a execução da linha 11 da Figura 10. Como escrito anteriormente, o método BOOSTING usa um mecanismo de votação ponderada, segundo um índice de importância, para determinar uma classificação. E para isto é necessário que o processo de geração de cada

classificador determine a eficiência do seu produto frente à base de treinamento. Tal eficiência é determinada pela importância do classificador ($\alpha^{(i)}$) gerado e ela é calculada com base na sua taxa de erro.

Tabela 7. Fórmulas da Taxa de Erro e da Importância do Classificador.

$\varepsilon^{(i)} = \sum_{j=1}^{n} w_j \delta(C_i(x_j) \neq y_j)$	(9)
$\alpha^{(i)} = \frac{1}{2}\ln\left(\frac{1-\varepsilon^i}{\varepsilon^i}\right)$	(10)

Onde:

- $C_i(x_j)$ representa a classe prevista pelo classificador C_i para a *j*-ésima instância da base de treinamento T;
- y_j representa a classe da *j*-ésima instância expressa pelo valor do atributo preditor da base de treinamento T;
- w_j é a probabilidade da *j*-ésima instância da base de treinamento T;
- δ (*x*) é um operador, cujos valores são definidos por $\delta(x) = \begin{cases} 0, se C_i(x_j) = y_j \\ 1, se C_i(x_j) \neq y_j \end{cases}$
- $\varepsilon^{(i)}$ é a taxa de erro do classificador C_i.

A determinação da importância do classificador C_i permite decidir se a amostra e seu classificador serão considerados ou descartados. Caso a taxa de erro seja superior ao limiar de 0,5 ($\varepsilon^{(i)} > 0,5$), descarta-se a amostra e outra amostra é gerada com os pesos reajustados para uma distribuição uniforme (linhas 13 e 14 do algoritmo da Figura 10).

Caso a importância do classificador seja satisfatória, os pesos são ajustados de acordo com a importância do classificador (linhas 16 e 17 do algoritmo da Figura 10). Tal ajuste de pesos é feita em duas etapas:

- primeiramente, aplica-se um fator multiplicativo aos pesos existentes e depois normalizam-se os pesos obtidos. O peso w_j da instância que tiver sua classe prevista corretamente é decrescido pela divisão pelo fator $e^{\alpha(i)}$, (e=2,71828; importância do

classificador $\alpha^{(i)}$). De forma análoga, o peso de cada instância, que for prevista incorretamente, é acrescido por meio da multiplicação pelo mesmo fator $e^{\alpha(i)}$.

- finalmente, o vetor de pesos *w* é normalizado, de tal forma que $\sum_{j=1}^{n} w_j$ seja igual a 1.

Ilustração do Método BOOSTING

A ilustração da aplicação do método BOOSTING toma como base de treinamento o conjunto de instâncias da Tabela 2. O número de classificadores é 3. O resultado da primeira execução do algoritmo da Figura 10, das linhas 7 e 8, encontra-se na Figura 11. A primeira amostra $A_{i=1}$ foi obtida a partir de uma distribuição uniforme. O Classificar C_i=1 é o resultado da execução da linha 8 para a amostra $A_{i=1}$.

AMOSTRA I

P	TQ	TC	AF	A
1	N	PM	N	S
2	N	PM	N	N
3	N	PM	S	S
4	N	L	N	N
5	N	L	S	N
6	C	PM	N	S
6	C	PM	N	S
7	C	PM	S	N
7	C	PM	S	N
9	C	L	N	S
10	D	PM	N	N
10	D	PM	N	N
10	D	PM	N	N
14	D	L	N	N

CLASSIFICADOR *C1*

SE TQ = N **E** TC = PM **ENTÃO** A = S (3.0/1.0)
SENÃO SE TQ = N **E** TC = L **ENTÃO** A = N (2.0)
SENÃO SE TQ = C **E** AF = N **ENTÃO** A = S (3.0)
SENÃO SE TQ = C **E** AF = S **ENTÃO** A = N (2.0)
SENÃO SE TQ = D **ENTÃO** A = N (4.0)

PESO DE C1: 0,650

Figura 11. Distribuição de probabilidade uniforme: Amostra A_1 e Classificador C_1

O próximo passo consiste em testar o Classificador C_1, visando conhecer as instâncias de T mal classificadas por C_1. No algoritmo da Figura 10, este passo é realizado pela chamada da linha 8, cujos índices das instâncias mal classificadas são armazenados no vetor B. Para o exemplo, as instâncias 2, 8 e 11 de T foram incorretamente classificadas. Na seqüência, as chamadas das linhas 10 e 11 realizam, respectivamente, as tarefas que resultam na taxa de erro e no valor de importância do Classificador C_1. Têm-se os seguintes valores para o exemplo em questão:

$\varepsilon^{(1)}$ = 0,357 $\alpha^{(1)}$ = 0,650

Como o teste da linha 12 não é verdade, ou seja, a taxa de erro não é superior a 0,5, deve-se então ajustar os pesos das instâncias da base de treinamento T conforme exibido no Quadro 2.

J	w_j inicial	w_j iteração 1
1	0,71	0,453
2	0,71	0,167
3	0,71	0,453
4	0,71	0,453
5	0,71	0,453
6	0,71	0,453
7	0,71	0,453
8	0,71	0,167
9	0,71	0,453
10	0,71	0,453
11	0,71	0,167
12	0,71	0,453
13	0,71	0,453
14	0,71	0,453

Quadro 2. Ajuste dos pesos das instâncias de T — primeira interação.

A segunda interação iniciará com a obtenção da segunda amostra $A_{i=2}$. Está última foi obtida a partir da base de treinamento T. Deve-se, entretanto, observar as instâncias de T tiveram seus pesos atualizados e as instâncias que tiveram seus pesos aumentados possuem maior probabilidade de aparecer na amostra $A_{i=2}$. A Figura 12 mostra as instâncias da amostra $A_{i=2}$ e o Classificador $C_{i=2}$, obtido a partir de $A_{i=2}$.

Amostra A_2				
P	**TQ**	**TC**	**AF**	**A**
1	**N**	**PM**	**N**	**S**
2	**N**	**PM**	**N**	**N**
2	**N**	**PM**	**N**	**N**
6	**C**	**PM**	**N**	**S**
6	**C**	**PM**	**N**	**S**
7	**C**	**PM**	**S**	**N**
8	**C**	**L**	**S**	**S**
9	**C**	**L**	**N**	**S**
10	**D**	**PM**	**N**	**N**
10	**D**	**PM**	**N**	**N**
11	**D**	**PM**	**N**	**S**
13	**D**	**L**	**S**	**N**
13	**D**	**L**	**S**	**N**
14	**D**	**L**	**N**	**N**

CLASSIFICADOR C_2

SE TQ = N **ENTÃO** A = N (3.0/1.0)
SENÃO SE TQ = C **ENTÃO** A = S (5.0/1.0)
SENÃO SE TQ = D EN**T**ÃO A= N (6.0/1.0)

PESO DE C2: 0,416

Figura 12, Distribuição de probabilidade não uniforme: Amostra A_2 e Classificador C_2.

A primeira observação a ser feita sobre a amostra gerada é que mesmo sendo aleatória, percebe-se que todas as instâncias classificadas incorretamente pelo Classificador C_1 estão

presentes no Classificador C_2. Outro fator marcante é a grande diferença entre o Classificador C_1 e o Classificador C_2. Submetendo os exemplos à base de treinamento original (Tabela 2), verificou-se que as instâncias 1, 3, 7 e 11 foram incorretamente classificadas. Sendo assim, têm-se os seguintes valores para o classificador em questão:

$$\varepsilon^{(2)} = 0{,}303 \qquad \alpha^{(2)} = 0{,}416$$

Como o valor da taxa de erro inferior a 0,5, então os pesos das instâncias são ajustados conforme os valores do Quadro . Observa-se neste quadro o aumento da probabilidade das instâncias 2 e 8 que foram erroneamente classificadas nos dois classificadores: C_1 e C_2. Por outro lado, observa-se que as instâncias 1, 3 e 7 possuem pesos baixos. Estas observações permitem ilustrar o procedimento do método BOOSTING que privilegia a geração de classificadores distintos a cada novo passo.

J	***w_j* inicial**	***w_j* iteração 1**	***w_j* iteração 2**
1	0,71	0,453	0,013
2	0,71	0,167	0,220
3	0,71	0,453	0,013
4	0,71	0,453	0,060
5	0,71	0,453	0,060
6	0,71	0,453	0,060
7	0,71	0,453	0,013
8	0,71	0,167	0,220
9	0,71	0,453	0,060
10	0,71	0,453	0,060
11	0,71	0,167	0,046
12	0,71	0,453	0,060
13	0,71	0,453	0,060
14	0,71	0,453	0,060

Quadro 3. Ajuste dos pesos das instâncias de T — segunda iteração.

Na Figura 13 são apresentados a terceira amostra e seu classificador, terminando a geração de classificadores para este exemplo (k = 3). Submetendo os exemplos à base de treinamento original (Tabela 2), verificou-se que as instâncias 2, 3, 10, 13 e 14 foram incorretamente classificadas. Sendo assim, têm-se os seguintes valores para o classificador em questão:

$$\varepsilon^{(3)} = 0{,}411 \qquad \alpha^{(3)} = 0{,}179$$

Amostra A_3				
P	**TQ**	**TC**	**AF**	**A**
1	**N**	**PM**	**N**	**S**
4	**N**	**L**	**N**	**N**
4	**N**	**L**	**N**	**N**
7	**C**	**PM**	**S**	**N**
7	**C**	**PM**	**S**	**N**
8	**C**	**L**	**S**	**S**
8	**C**	**L**	**S**	**S**
8	**C**	**L**	**S**	**S**
9	**C**	**L**	**N**	**S**
10	**D**	**PM**	**N**	**N**
11	**D**	**PM**	**N**	**S**
12	**D**	**PM**	**S**	**N**
12	**D**	**PM**	**S**	**N**
12	**D**	**PM**	**S**	**N**

```
CLASSIFICADOR C3

SE TC = PM E AF = N ENTÃO A= S (3.0/1.0)
SENÃO SE TC = PM E AF = S ENTÃO A = N (5.0)
SENÃO SE TC = L E TQ = N ENTÃO A = N (2.0)
SENÃO SE TC = L E TQ = C ENTÃO A=S (4.0)
SENÃO SE TC = L E TQ = D ENTÃO A=S (0.0)

PESO DE C3: 0,179
```

Figura 13. Distribuição de probabilidade não uniforme: Amostra A3 e Classificador C3..

Daqui em diante não é mais necessário recalcular os pesos, por que C_3 é o último classificador que deveria ser gerado (k=3). No entanto, é necessário refazer o cálculo tanto para a taxa de erro, que pode implicar no descarte do classificador gerado, quanto para a importância de C_3. Esta última será utilizada para montagem da votação ponderada.

Para ilustrar o processo de classificação, para o método BOOSTING, um conjunto de três novas instâncias serão classificadas (Tabela 8).

Tabela 8. Instâncias a serem classificadas pelo método BOOSTING.

Instância	Tipo de Queixa	Tipo de Cirurgia	Atividade Física	Alteração
a_1	Nula	Prótese Mamária	Não	?
a_2	Constipação	Prótese Mamária	Não	?
a_3	Dor	Lipoaspiração	Sim	?

As classificações da instância a_1 são as seguintes:

- **Sim**, para o Classificador C_1 que tem importância 0,650;
- **Não**, para o Classificador C_2 que tem importância 0,416;
- **Sim**, para o Classificador C_3 que tem importância 0,179.

Como a soma das importâncias para a classe **Sim** (0,650 + 0,179 = 0,829) é superior a soma das importâncias para a classe **Não** (0,416), então classe resposta é **Sim**.

As classificações da instância a_2 são as seguintes:

- **Sim**, para o Classificador C_1 que tem importância 0,650;
- **Sim**, para o Classificador C_2 que tem importância 0,416;
- **Sim**, para o Classificador C_3 que tem importância 0,179.

Como a votação foi unânime para a classe **Sim** (0,650 + 0,416 + 0,179 = 1,245), então classe resposta é **Sim**.

As classificações da instância a_3 são as seguintes:

- **Não**, para o Classificador C_1 que tem importância 0,650;
- **Não**, para o Classificador C_2 que tem importância 0,416;
- **Sim**, para o Classificador C_3 que tem importância 0,179.

Como a soma das importâncias para a classe **Não** (0,650 + 0,416 = 1,166) é superior a soma das importâncias para a classe **Sim** (0,179), então classe resposta é **Não**.

Finalmente, o funcionamento do método BOOSTING é similar ao caso do gestor ser humano que se cerca de diversos conselheiros com diferentes especialidades. Ainda, no método BOOSTING a obtenção de amostras tende a ser ortogonal, por que se busca incluir instâncias classificadas incorretamente e excluir instâncias classificadas corretamente. Isto permite não desprezar as relações entre algumas instâncias da base de treinamento. Entretanto, a geração de amostras para o método BAGGING é feita sempre sorteando instâncias com distribuição uniforme. Em resumo, a forma de geração das amostras é a grande diferença entre os métodos.

Teoricamente, a eficiência dos classificadores, em termos de taxa de acerto, segue a seguinte ordem: JRIP, BAGGING+JRIP, BOOSTING+JRIP. A taxa de acerto está ligada a forma de validação de cada classificador, seja ele simples ou composto. Nesta dissertação, a técnica de validação escolhida foi a cruzada; ela produz bons resultados e é uma das formas mais utilizadas.

2.5.4 Validação Cruzada

Em aprendizagem de máquina, a obtenção do conhecimento, por exemplo, na forma de uma árvore de decisão é uma primeira tarefa fundamental. A segunda tarefa é estimar a taxa de acerto do conhecimento descoberto. Nestes termos, a escolha de uma boa técnica de validação é essencial, assim sendo, a técnica de validação cruzada é uma das formas mais completas e utilizadas para estimar a taxa de acerto de classificadores (DIAMANTIDIS; KARLIS; GIAKOUMAKIS, 2000).

É importante salientar que a técnica de validação empregada deva ser simples e não tendenciosa. Em termos de validação simples, uma opção seria verificar a taxa de acerto, de um classificador *C* obtido por meio de um método qualquer *M* e de uma base de dados de treinamento *T* com *n* instâncias, e gerar o classificador *C* utilizando as n instâncias de *T* como

conjunto de treinamento e testar o classificador C utilizando a mesma base de dados *T*. Esta técnica é simples por que ela gera um valor de taxa de acerto a partir de uma única taxa de acerto; ou seja, a taxa de acerto final não é uma composição/média de várias outras. Todavia, tal valor de taxa de acerto pode ser tendencioso. Isto pode ocorrer por que o mesmo conjunto de instâncias foi utilizado para gerar o classificador e para testá-lo. Uma alternativa é fazer com que a técnica combine vários valores de taxa de acerto. Nestes termos, a validação cruzada aparece como uma opção menos tendenciosa; ela busca evitar os problemas da técnica simplista supracitada.

O princípio da técnica de validação cruzada é simples. Aqui, a base de dados *T* com *n* instâncias é subdividida em *f* amostras geradas aleatoriamente. O tamanho de cada amostra é igual ao número de instâncias *n* dividido pelo número de amostras *f*; tal fração pode gerar um valor aproximado para o caso *n* de uma divisão não exata (TAN; STEINBACH; KUMAR, 2006).

A operacionalização do método de validação cruzada é a seguinte: repete *f* vezes um processo de treinamento e teste. Sendo que, cada interação utiliza uma das *f* amostras para ser o conjunto de teste/validação *tf* e as demais *f-1* amostras para ser o conjunto de treinamento *tt*. Cada uma destas *f* amostras de validação retorna uma taxa de acerto tx_i. A taxa de acerto final *txf* é expressa pela média aritmética simples das tx_i, conforme fórmula a seguir: $txf = \frac{1}{f}\sum_{i=1}^{f} tx_i$

Ilustração do Método Validação Cruzada

Seja T uma base de dados com *n* instâncias, onde *n* é igual a duzentos. Seja o número de *folds f* igual cinco. Para estes valores têm-se as seguintes amostras: A_1, A_2, A_3, A_4, A_5. Cada amostra A_i tem quarenta instâncias. As configurações para as iterações são as seguintes:

```
Interação 1:
Entradas: conjunto de treinamento T é {A2, A3, A4, A5}
          conjunto de teste tt é {A1}
Saídas:   classificador C1
          taxa de acerto tx1 é 80%

Interação 2:
Entradas: conjunto de treinamento T é {A1, A3, A4, A5}
          conjunto de teste tt é {A2}
Saídas:   classificador C2
          taxa de acerto tx2 é 92%

Interação 3:
Entradas: conjunto de treinamento T é {A1, A2, A4, A5}
          conjunto de teste tt é {A3}
Saídas:   classificador C3
          taxa de acerto tx3 é 98%

Interação 4:
Entradas: conjunto de treinamento T é {A1, A2, A3, A5}
          conjunto de teste tt é {A4}
Saídas:   classificador C4
          taxa de acerto tx4 é 78%

Interação 5:
Entradas: conjunto de treinamento T é {A1, A2, A3, A4}
          conjunto de teste tt é {A5}
Saídas:   classificador C5
          taxa de acerto tx5 é 91%

Taxa de acerto final é 87,8%
```

Como descrito anteriormente, a técnica de validação cruzada, pode ser aplicada a qualquer método de aprendizagem de máquina.

2.6 Considerações Finais

Ao longo deste capítulo, foram examinados três métodos de geração de classificadores, que serão utilizados nos experimentos, a saber: C4.5, BAGGING+C4.5, BOOSTING+C4.5 e também uma técnica de validação, conhecida como validação cruzada.

Sobre os métodos examinados, o C4.5, além de possuir a capacidade de gerar modelos de classificação a partir de exemplos com valores faltantes, apresenta bons resultados em bases de dados ruidosas. Frente estas características, o C4.5 tende a apresentar resultados bastante satisfatórios, principalmente quando combinado com métodos BAGGING e BOOSTING.

Na literatura, o método BAGGING, mostrou-se eficiente para reduzir a taxa de erro de classificadores gerados pelo algoritmo C4.5, quando aplicado em bases de dados com características diferentes (BREIMAN, 1996), bem como mostrou eficiência satisfatória em

amostras com uma quantidade pequena de instâncias (LOPES, 2007). O método BOOSTING mostrou-se eficiente para operar com bases de dados que possuíam uma grande quantidade de atributos, como as bases de dados na área de saúde. Ele também apresentou bons resultados em bases de dados reduzidas dimensionalmente por meio da filtragem de atributos (LOPES, 2007).

A validação cruzada é particularmente interessante por reduzir a possibilidade de taxas de acerto tendenciosas. A técnica também tem méritos pela simplicidade de realização e compreensão.

3 METODOLOGIA

O trabalho proposto visa descobrir padrões a partir de uma base de dados. Entretanto, é importante observar que a base de dados utilizada foi obtida por meio de um processo de simulação. Neste último, o valor de cada atributo de um dado registro/exemplo foi obtido por meio de uma função de geração de valores aleatórios aplicada ao domínio do atributo em questão.

Em termos gerais, o processo de geração de dados simulou a obtenção dos dados de diferentes fontes de uma clínica de cirurgia plástica hipotética. Tal simulação foi dirigida para obter dados que permitissem obter padrões para ilustrar um conjunto de informações úteis para o planejamento e execução de um bom tratamento fisioterapêutico de um paciente hipotético. A consecução deste objetivo incluiu diferentes tarefas não triviais, a saber: (i) um estudo sobre quais dados deveriam ser coletados de diferentes bases de diferentes profissionais atuantes em uma clínica de cirurgia plástica; (ii) a execução de inúmeros experimentos visando extrair conhecimentos para o processo de planejamento e execução de um bom tratamento fisioterapêutico; e (iii) uma análise comparativa das taxas de acerto dos diferentes classificadores obtidos por diferentes métodos.

Os resultados obtidos dos experimentos foram estruturados por meio de um conjunto de módulos de software, que executam tarefas bem especificas: (i) registro de eventos clínicos e de avaliação de cada paciente (sistema operacional); (ii) registro histórico dos eventos clínicos e de avaliação de cada paciente em esquema multidimensional (sistema informacional); (iii) obtenção de padrões de comportamento, na forma de regras, a partir dos históricos dos pacientes; e, finalmente, (iv) interpretação e aplicação, de forma interativa, dos conhecimentos descobertos por meio da execução das tarefas descritas no item anterior. Deve-se, entretanto, destacar que o foco do trabalho foi fixado, em termos de fundamentação teórica e de desenvolvimento, nos itens (iii) e (iv). Finalmente, é importante reafirmar que os resultados referentes aos padrões descobertos devem ser interpretados a luz de uma base de dados gerada por meio de um processo de simulação, i.e., não se trata de uma base de dados real.

3.1 Modelagem e Geração dos Dados

A base de dados utilizada foi montada por meio de um processo de geração de dados, que simulou os dados de diferentes fontes de uma clínica de cirurgia plástica. Tais fontes seriam os profissionais de diversas áreas, entre elas, médicos, fisioterapeutas, secretárias e administrador. A simulação tenta se aproximar de uma situação real, onde cada profissional gera seus registros de forma independente para pacientes em comum ou não, o que dificulta a análise dos dados, que em sua maioria são inconsistentes, surgindo, assim, a necessidade de integração dos dados destas diferentes fontes. Tal integração é a primeira tarefa para permitir que tomadores de decisão possam fazer uso dos dados de forma confiável, eficiente e flexível.

Inicialmente, foram determinados alguns assuntos canonicamente presentes em qualquer clínica de cirurgia estética, como por exemplo: paciente, cirurgia, alteração, queixa principal e tempo. Na seqüência, o processo de detalhamento de cada assunto foi realizado em duas etapas, uma mais geral e outra mais específica. A primeira etapa consistiu em definir para cada assunto um conjunto de atributos e seus respectivos domínios. A segunda etapa consistiu em restringir os domínios de certos atributos quando tomados em conjunto para gerar um determinado registro ou exemplo. Por exemplo, o domínio do atributo idade sem restrição é de 15 até 70. Entretanto, quando se trata de paciente cuja ocupação é estudante, o domínio considerado para atributo idade limita-se ao intervalo fechado [15, 21]. Esta abordagem foi generalizada para todos os atributos.

Os detalhes dos assuntos supracitados, em termos dos atributos, dos domínios destes atributos e das restrições sobre tais domínios foram sumarizados nas tabelas A1, A2, A3 e A4 do anexo A. Em termos mais específicos, a Tabela A1 descreve os dados usados para a geração de exemplos de “paciente”. A Tabela A2 descreve os dados usados para a geração de exemplos de “cirurgia”. Tabela A3 descreve os dados usados para a geração de exemplos de “alteração”. Tabela A4 descreve os dados usados para a geração de exemplos de “queixa principal”. Finalmente, a Tabela R do respectivo anexo descreve o esquema da tabela de dados que conterá os exemplos gerados a partir das tabelas A1, A2, A3 e A4.

A função aplicada sobre as tabelas A1, A2, A3 e A4 para gerar a Tabela R está representado na Figura 14.

```
função  geraExemplo(A, var R : TABELA, α, k : inteiro ) : inteiro
{A é tabela que descreve os dados para a geração de exemplos de um dado assunto.}
{R é referência da tabela que conterá os eventos/exemplos gerados a partir de A}
{α é número máximo de exemplos que serão  gerados aleatoriamente para cada valor da coluna 1 da
Tabela A.}
{k é o apontador para a próxima linha vazia da Tabela R, onde será inserido o próximo exemplo gerado}
inicio
      inteiro: i;     {linha corrente da Tabela A}
      inteiro: n;    {número de linhas da Tabela A}
      inteiro: j ;            {coluna/atributo corrente da Tabela A}
      inteiro: m ;  {número de colunas/atributos da Tabela A}
      String: X ;             {valores possíveis para o atributo j correspondente ao atributo 1 da
            linha i}
      inteiro: q,h; {variáveis auxiliares}
      para cada i de 1 até n passo 1 faça
      inicio
            h = randon( α );        {retorna um número inteiro aleatório ≤ α}
            para cada q de 1 até h passo 1 faça    {para cada i será gerado h exemplos}
            inicio
                  R[ k ][ 1 ] = A[i][1];
                  para cada j de 2 até m passo 1 faça    {para cada q será gerado m-1 valores}
                  inicio
                        R[k][j] = pegaValor(A[i][j]);
                        k = k + 1;
                  fimpara
            fimpara
      fimpara
      retorne k;    {o valor de k indica o número de exemplos +1 constante na Tabela R}
fim.
```

```
Função pegaValor(Vetor<String> V) : String

{V é um vetor de cadeias de caracteres (String)}

{retorna um dos elementos de V selecionados aleatoriamente}

inicio

      inteiro: u = V.lenth();                  {retorna o número de elementos de V}

      inteiro: h = randon( u );     {retorna um número aleatório ≤ u}

      retorne V[ h ];

fim.
```

Figura 14. Construção da Base de Dados Simulados, Tabela R.

A geração propriamente dita pode ser ilustrada por meio da invocação da função geraExemplo da seguinte forma:

```
Inteiro k;
k  = geraExemplo(A1, ref R, 16, 1);          {geração de exemplos para anamnese}
k = geraExemplo(A2, ref R, 16, k);           {geração de exemplos para cirurgia}
k = geraExemplo(A3, ref R, 16, k);           {geração de exemplos para alteração}
k = geraExemplo(A4, ref R, 16, k);           {geração de exemplos para queixa principal}
```

ou simplesmente

```
k = geraExemplo(A4, ref R, 16,
                    geraExemplo(A3, ref R, 16,
                                geraExemplo(A2, ref R, 16,
                                            geraExemplo(A1, ref R, 16, 1))));
```

Deve-se salientar que a cada execução das chamadas anterior, um conjunto com um número diferente de exemplos é obtido. Para o experimento considerado neste trabalho, o número total de registros gerados foi 4826. Tal número de registros é fornecido pelo valor da variável k.

A Figura 15 mostra esquematicamente os fluxos e as etapas de manipulação dos dados. Tal manipulação inclui a extração, transformação e carga dos dados de diferentes fontes para uma única fonte de dados, canonicamente chamada de d/w (*data warehouse*). Esta fonte única de dados, tanto em termos de armazenamento quanto de representação, facilita a geração de informações e também a comunicação entre os profissionais. Classicamente, tais informações são tendências que dão suporte ao processo de tomada de decisão de maneira eficiente e rápida.

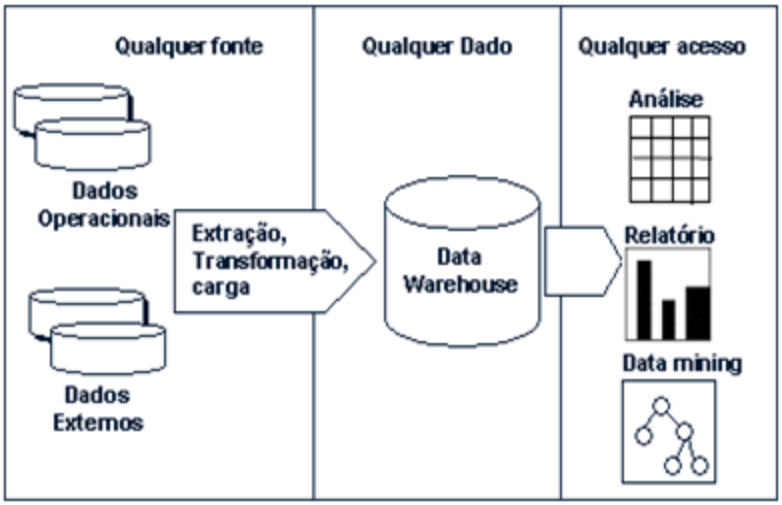

Figura 15. Esquema simplificado para a montagem e exploração de um d/w.

A construção desta base de dados do d/w foi dirigida por meio de um conjunto de questionamento que o mesmo deveria estar habilitado a responder, tais como: Qual é a relação entre alterações e o tipo de cirurgia? Qual foi o tipo de cirurgia mais realizada em determinado ano? Quais são as idades dos pacientes que mais operaram nos últimos três anos? Qual é a relação existente entre queixa principal e tipo de cirurgia? Qual é o perfil, com relação a sexo e idade, dos pacientes que mais realizaram cirurgia de lipoaspiração em um ano específico? Qual foi o número de pacientes do sexo feminino usuários de antidepressivo, que realizaram cirurgia de abdominoplastia e obtiveram um bom grau de satisfação pós-operatória? Qual é o percentual de satisfação dos pacientes? Qual é a relação existente entre profissão e tipo de cirurgia realizada? Qual é a relação da idade com o volume de prótese mamária implantada?

Identificado os dados empiricamente para os experimentos, estes foram extraídos/gerados simulando diversas fontes de registros utilizadas. Caso os dados não fossem gerados por um processo de simulação controlado, seria necessária filtragem ou limpeza dos dados. Isto poderia ser realizado para resolver problemas de consistência dos nomes, dos valores e das unidades que devem ser transformados em uma representação uniforme. Como por exemplo: um paciente pode ter realizado cirurgia nos braços e em um determinado registro estar descrito como Braquioplastia e em outro como Liffiting Braquial, sendo que os dois termos identificam a mesma cirurgia para a área de saúde, mas são termos distintos para linguagem de máquina, devendo ser padronizados, portanto, filtrados antes de carregar a base de dados central.

A Figura 16 mostra o modelo parcial dos dados que foram considerados para a montagem da base de dados para os experimentos. O modelo foi representado em UML (Grossman, et al. 2005), e ele traduz, de forma simplificada, as seguintes narrativas:

- cada paciente é definido pelas seguintes características: data de nascimento, ocupação, sexo, hábito alimentar, pratica ou não de atividade física, fumante ou não, medicamentos que eventualmente faz uso. Além disto, ele submete-se a um ou mais eventos;
- cada evento ocorre em uma data e pode ser do tipo simples ou composto;
- um evento composto compreende pelo menos um ou mais eventos simples;
- uma avaliação possui um parecer e é um evento composto;
- uma alteração possui um tamanho, uma cor, uma localização, uma data e um tipo de alteração, que pode ser, por exemplo: HEMATOMA, FIBROSE, SEROMA, etc.;
- uma queixa possui uma intensidade, uma localização, uma data e um tipo de queixa, que pode ser, por exemplo: DOR, INSÔNIA, etc.;
- uma *perimetria* possui uma medida, uma localização, uma data;
- uma cirurgia possui um volume, uma localização, uma data e um tipo de cirurgia, que pode ser, por exemplo: LIPOASPIRAÇÃO, ABDOMENOPLASTIA, etc., uma conduta, que pode ser, por exemplo: DRENAGEM LINFÁTICA MANUAL DE ABDOMEN, ULTRA-SOM, etc., uma técnica, que pode ser, por exemplo: CLÁSSICA, ILLOUZ, etc.;
- toda alteração, queixa, *perimetria* fazem parte de uma avaliação.

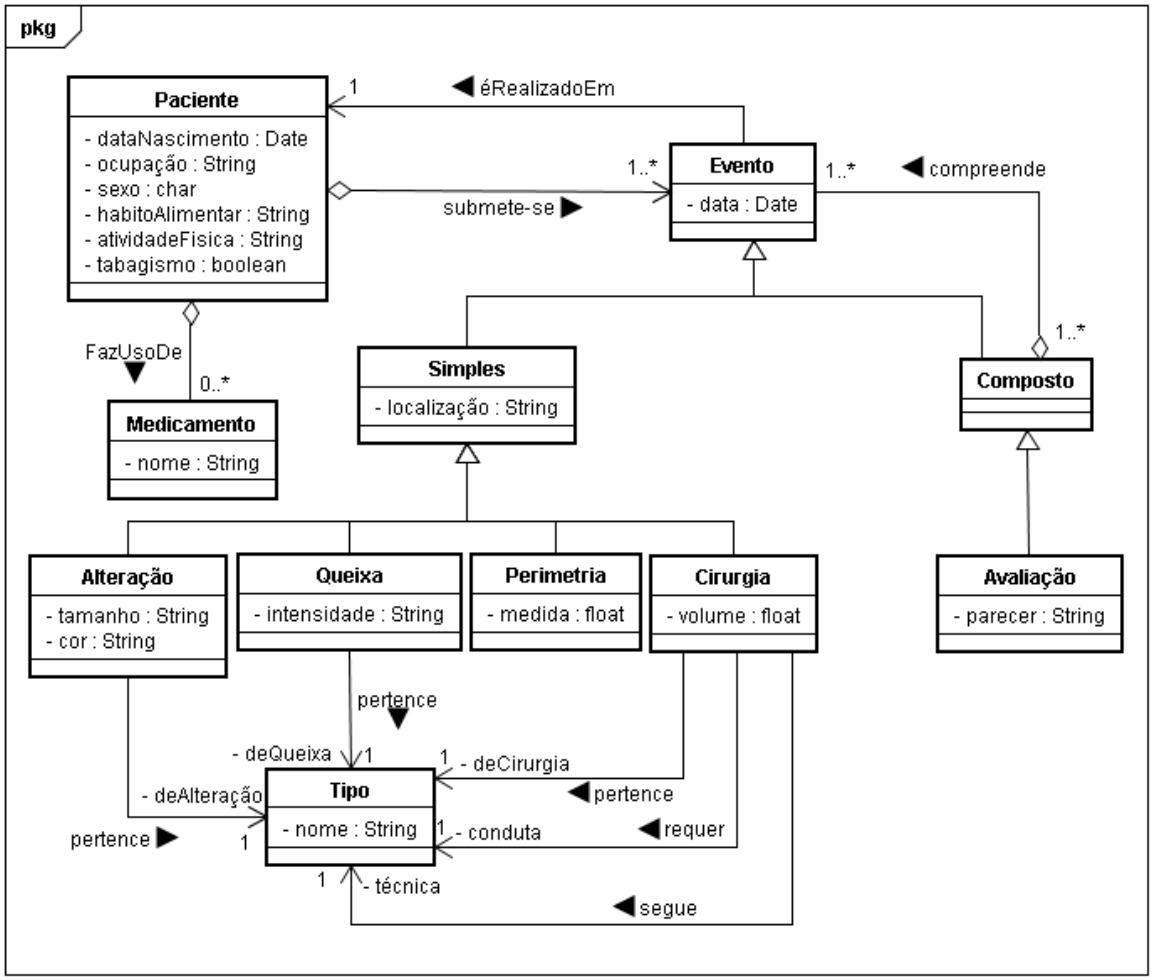

Figura 16. Modelo de dados parcial do d/w – esquema em UML.

Deve-se destacar que foi construído, neste trabalho, um sistema de informação para o registro dos dados operacionais. E a carga do *d/w* foi realizada por meio de programas específicos escritos na linguagem de programação Java. A base de dados foi construída sobre o SGBD Access da Microsoft™. Este último implementa o modelo de dados relacional, que permite facilmente a definição dos esquemas de um *d/w*.

O esquema da Figura 17 mostra um modelo multidimensional. Este último se apóia em três elementos básicos: os fatos, que estão representados pelos eventos/serviços realizados; as dimensões que é o paciente, tempo, cirurgia, avaliação fisioterapêutica. etc..; e as medidas que são as quantidades de cirurgias, de alterações, de queixas principais, além destes poder-se-ia ter também os valores despendidos.

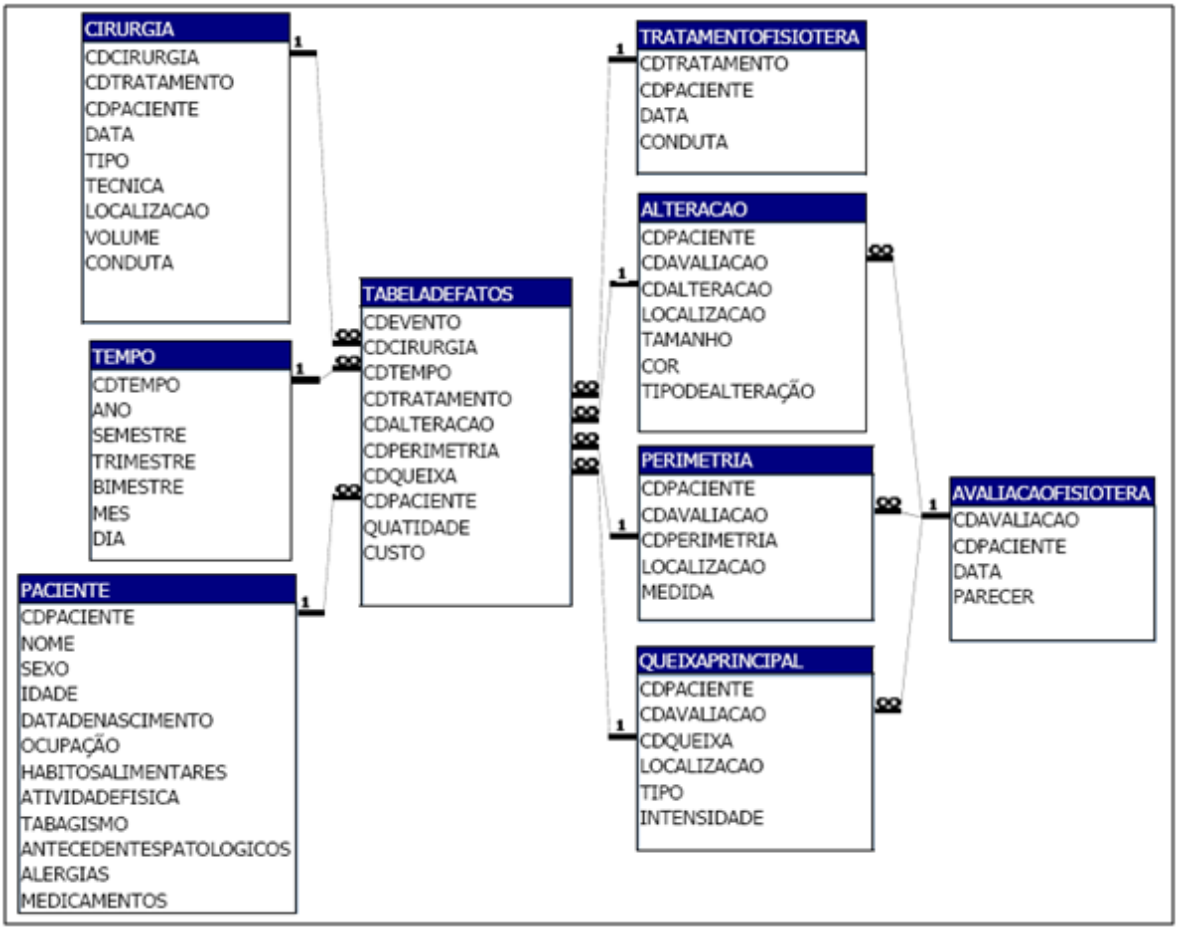

Figura 17. Modelo de dados parcial do d/w – esquema em E-R.

Visualmente, este modelo pode ser representado por meio de um cubo, onde cada uma das dimensões é uma face do cubo (Figura 18). Deve-se ler a informação da célula laranja do cubo da seguinte forma: no ano de 2008, há 183 pacientes que se submeteram a cirurgias e tomavam medicamentos antidepressivos.

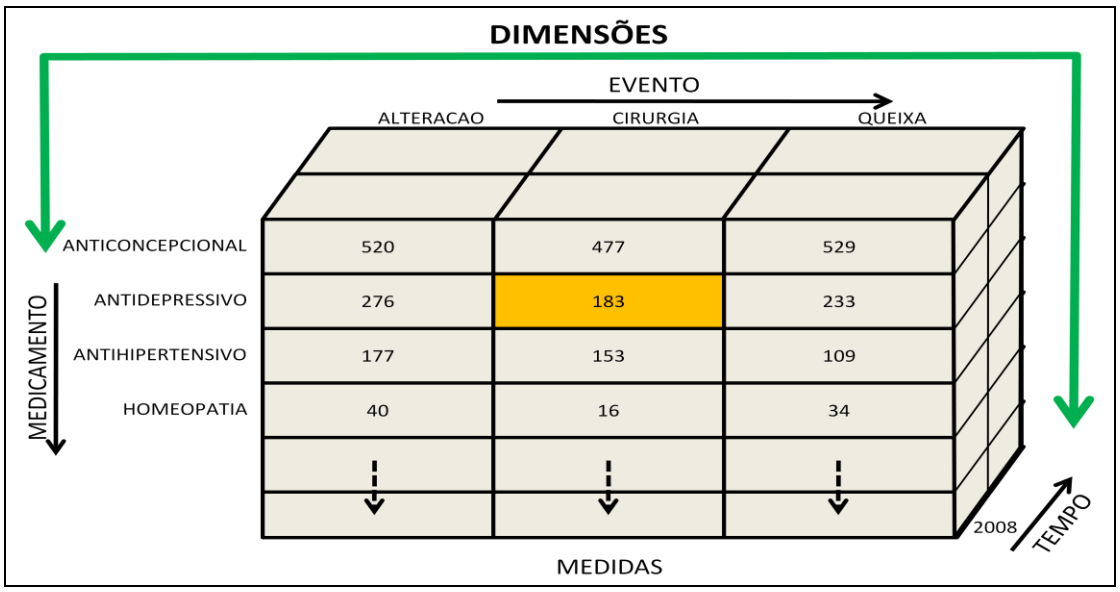

Figura 18. Cruzamento de três dimensões: MEDICAMENTO, EVENTO e TEMPO – cubo 3D.

A Figura 19 apresenta a mesma informação da Figura 18 na forma de um gráfico de barras. O eixo das ordenadas são as quantidades de medicamentos.

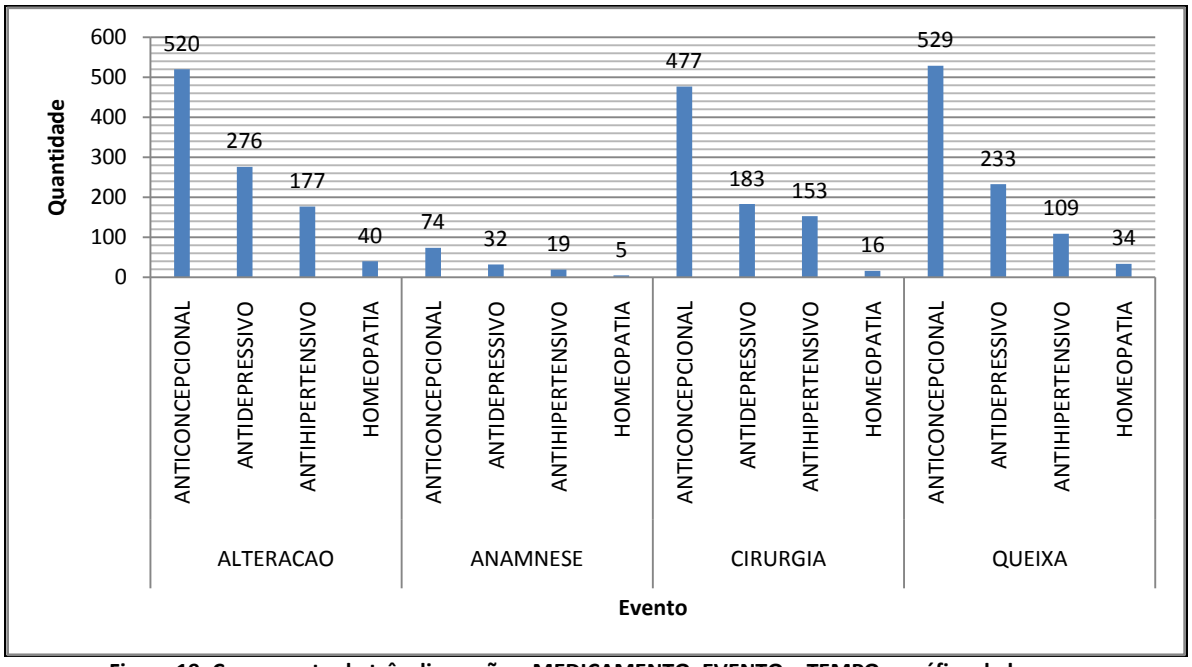

Figura 19. Cruzamento de três dimensões: MEDICAMENTO, EVENTO e TEMPO – gráfico de barras.

A Tabela 9 mostra a mesma informação Figura 19 na forma de um relatório formatado em linhas e colunas. Este formato é comumente usado em ferramentas clientes que implementam

técnicas de navegação nos dados por meio de tabelas dinâmicas. Estas últimas são ditas dinâmicas porque o usuário tem a liberdade alternar suas linhas e colunas ou compô-las de acordo com a sua necessidade.

Tabela 9. Cruzamento de três dimensões: MEDICAMENTO, EVENTO e TEMPO – relatório de linhas e colunas.

Soma de QUANTIDADE	Rótulos de Coluna	
Rótulos de Linha	2008	Total geral
ALTERACAO	**1013**	**1013**
ANTICONCEPCIONAL	520	520
ANTIDEPRESSIVO	276	276
ANTIHIPERTENSIVO	177	177
HOMEOPATIA	40	40
ANAMNESE	**130**	**130**
ANTICONCEPCIONAL	74	74
ANTIDEPRESSIVO	32	32
ANTIHIPERTENSIVO	19	19
HOMEOPATIA	5	5
CIRURGIA	**829**	**829**
ANTICONCEPCIONAL	477	477
ANTIDEPRESSIVO	183	183
ANTIHIPERTENSIVO	153	153
HOMEOPATIA	16	16
QUEIXA	**905**	**905**
ANTICONCEPCIONAL	529	529
ANTIDEPRESSIVO	233	233
ANTIHIPERTENSIVO	109	109
HOMEOPATIA	34	34
Total geral	**2877**	**2877**

O cubo oferece recursos que permitem realizar diferentes operações, sendo as principais: *drill-down*, ou seja, aumentar o nível de detalhamento de uma dimensão, ex: evento (cirurgia.técnica.tipo); *drill-up*, quando diminuiu o nível de detalhamento, ex: tempo:ano ao invés de tempo:mês. A Tabela10 ilustra tal situação, onde a dimensão cirurgia é mostrada em mais detalhes.

Tabela 10. Cruzamento de três dimensões: EVENTO (CIRURGIA.TÉCNICA.TIPO) e TEMPO – relatório de linhas e colunas.

Soma de QUANTIDADE	Rótulos de Coluna	
Rótulos de Linha	2008	Total geral
CIRURGIA	**1090**	**1090**
CLASSICA	**109**	**109**
ABDOMENOPLASTIA	109	109
ANCORA	**90**	**90**
ABDOMENOPLASTIA	90	90
CLASSICA	**759**	**759**
BLEFAROPLASTIA	181	181
CORRECAO-DE-CICATRIZ	168	168
DORSOPLASTIA	149	149
INTIMA-PEQ-LABIOS	136	136
LIFITING-DE-BRACO	125	125
ILLOUZ	**116**	**116**
LIPOASPIRACAO	116	116
T-INVERTIDO	**16**	**16**
MAMOPEXIA	16	16
Total geral	**1090**	**1090**

O acesso ao *d/w* fornece a geração de informações consistentes para apoiar as tomadas de decisões. Tal geração é intermediada por ferramentas de OLAP, sob um modelo multidimensional. Tais ferramentas facilitam a recuperação, a investigação, o resumo e a organização de dados. Estes últimos são representados por meios de cubos de dados, conforme exemplificado anteriormente na Figura 18.

As informações obtidas podem ser empregadas para dar suporte tanto no planejamento estratégico administrativo como nos tratamentos terapêuticos futuros como, por exemplo, a observação de que o maior número de queixas não está relacionada ao uso de antidepressivo e sim ao uso de anticoncepcionais, e proporcionalmente, os pacientes que fazem uso de anticoncepcionais têm maior tendência a queixas do que os pacientes que fazem uso de antidepressivo.

A informação de suporte a decisão obtida por meio de consultas OLAP têm seus limites. Ela pode ser incrementada substancialmente por meio da aplicação de técnica de descoberta de conhecimentos a partir dos dados constantes no *d/w*.

3.1.1 Origem e formato dos dados

Como já dito anteriormente, dados são elementos de tomadas de decisões de diferentes especialidades de uma clínica de cirurgia estética. Em geral, estes dados foram extraídos das diversas fontes individuais de registros utilizados pelos especialistas. Na seqüência, estes dados foram filtrados/depurados. Esta filtragem ou limpeza dos dados refere-se, por exemplo, à consistência dos nomes, dos valores e das unidades que devem ser transformados em uma representação uniforme. Deve-se, entretanto, observar como já dito anteriormente que os dados das fontes mencionadas foram obtidos por meio de um processo de simulação.

Deliberadamente, os dados referentes à parte de custos dos procedimentos assim como outros elementos de informação puramente administrativa não foram exibidos nos relatórios deste trabalho em geral.

A próxima fase dos estudos concerne à remoção e inclusão de novos dados.

3.2 Preparação da Base de Dados para os Modelos de Previsão

A preparação da base de treinamento para os modelos de previsão tomou como entrada os dados fictícios do *d/w,* que foram mostrados anteriormente. De forma pragmática, a etapa de pré-processamento limitou-se no enriquecimento e na filtragem de atributos da base de dados original (ou do *d/w*). Tal enriquecimento foi feito por meio da adição de novos atributos e a filtragem consistiu na remoção manual ou automática de atributos. Nesta linha, vários atributos ditos irrelevantes foram excluídos por um processo de filtragem manual, a saber:

- diferentes códigos (ou chaves primárias) das diferentes tabelas do d/w;
- diferentes datas nos formatos dd/mm/aaaa;
- nome do paciente;
- quantidade; e
- custos de cada procedimento.

Por outro lado, outros atributos foram incluídos por derivação, como exemplo, os atributos:

- semestre, trimestre, bimestre, mês e dia;

- assiduidade, número total de procedimentos que um paciente submeteu-se na clínica; e
- merge, corresponde a união dos valores dos atributos: deAlteração, deQueixa, deCirurgia.

Nos diferentes experimentos, tendo em vistas diferentes objetivos em termos de atributos metas, foram gerados diferentes conjuntos de treinamentos por intermédio de algoritmos de seleção de atributos.

3.2.1 Seleção de Atributos

Segundo Lee (1999), muitos algoritmos de mineração de dados não funcionam bem com grandes quantidades de atributos, deste modo, a seleção de atributos tende a melhorar o desempenho de tais algoritmos. Outro fator que justifica a seleção de atributos é a melhor compreensão do problema com uma menor quantidade de atributos, resultando em um conhecimento induzido na forma de classificadores com compreensão mais fácil.

Em nossos experimentos as motivações para a seleção de atributos foram para reduzir o número de atributos e, em conseqüência, reduzir o tempo de aprendizagem e também melhorar a qualidade/simplicidade dos classificadores obtidos. A redução foi realizada por meio da aplicação dos métodos *CfsSubSetEval,* utilizando o algoritmo *BestFirst* como mecanismo de busca, e *GainRatio,* utilizando o algoritmo *Ranker* como mecanismo de busca. Os métodos foram empregados da seguinte forma: quando o primeiro método filtrava um conjunto de atributos, no mínimo, superior a 10% do total, o resultado era considerado; mas, quando o primeiro método não atingia tal percentual mínimo, o segundo método era empregado. Em geral, redução média do número de atributos foi de 85%. Deve-se observar, entretanto, que quando foi considerado o segundo método, optou-se em tomar não mais que 20% dos atributos ranqueados; ou seja, apenas os mais significativos. Ambos os métodos estão disponíveis no software WEKA (WITTEN & FRANK, 2000).

Segundo Hall (2000) e Koller (1996), deve-se tratar a redundância dos atributos durante o processo de seleção; tal redundância tende a afetar a qualidade da classificação. O problema da redundância ocorreu, neste trabalho, após o enriquecimento dos dados com os seguintes

atributos: as datas e seus desmembramentos. Por exemplo, cada data foi desmembrada em ano, semestre, trimestre, bimestre, mês, dia.

Como visto anteriormente, existem vários objetivos a serem avaliados em termos de atributos metas, para tal foram definidas diferentes configurações para os conjuntos de treinamento. O primeiro destes conjuntos é descrito pelo dicionário de dados da Tabela 11; tal conjunto de treinamento é para gerar o modelo de previsão que **VISA CARACTERIZAR O PERFIL DAS ALTERAÇÕES PÓS-CIRÚRGICAS.**

Tabela 11. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção CfsSubSetEval e mecanismo de busca: BestFirst, meta tipo de alteração.

	Atributo	Tipo	Descrição
previsão	CONDUTA	Nominal	Conduta fisioterapêutica pós-cirúrgica, por exemplo, DRENAGEM LINFÁTICA MANUAL DE ABDOMEN após uma cirurgia de lipoaspiração abdominal.
	VOLUME	Contínuo	Volume de líquido retirado, por exemplo, em uma cirurgia de LIPOASPIRAÇÃO ABDOMINAL.
	COR	Nominal	Cores dos hematomas, por exemplo, ROXO.
	INTENSIDADE	Nominal	Indicação do nível de uma dor, por exemplo, LEVE,
	LOCALIZACAO	Nominal	Local onde um procedimento foi realizado, por exemplo, ABDOMEN.
	TAMANHO	Nominal	Tamanhos dos hematomas, por exemplo, GRANDE.
meta	deALTERAÇÃO	Nominal	Alterações pós-cirúrgicas observadas nos pacientes, por exemplo, HEMATOMA.

A Tabela 12 mostra a descrição do conjunto de treinamento para o modelo de previsão, que **VISA CARACTERIZAR O PERFIL DAS ALERGIAS DOS PACIENTES SUBMETIDOS AOS PROCEDIMENTOS EM UMA CLÍNICA DE CIRURGIA ESTÉTICA.** Neste caso, tomaram-se como classes as alergias relatadas pelos pacientes na anamnese.

Tabela 12. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção GainRatio e mecanismo de busca: Ranker, meta ALERGIA.

	Atributo	Tipo	Descrição
previsão	ANT_PATOL	Nominal	Antecedente patológico mais significativo para um dado paciente, por exemplo, MAMOPLASTIA.
	ATIVIDADEFISICA	Nominal	Atividade física realizada regularmente pelo paciente, por exemplo, CAMINHADA.
	OCUPAÇÃO	Nominal	Atividade principal exercida pelo paciente, por exemplo, ESTUDANTE.
	TABAGISMO	Nominal	Fumante ou não fumante.
meta	ALERGIA	Nominal	Alergia mais significativa de um dado paciente, por exemplo, PÓ.

A Tabela 13 mostra a descrição do conjunto de treinamento para o modelo de previsão, que **VISA CARACTERIZAR O PERFIL DOS PACIENTES QUE TOMAM (OU TOMARAM RECENTEMENTE) CERTOS MEDICAMENTOS**.

Tabela 13. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção GainRatio e mecanismo de busca: Ranker, meta medicamento em uso pelo paciente.

	Atributo	Tipo	Descrição
Previsão	ALTERACAO	Nominal	Alterações pós-cirúrgicas observadas nos pacientes, por exemplo, HEMATOMA.
	ANT_PATOL	Nominal	Antecedente patológico mais significativo para um dado paciente, por exemplo, MAMOPLASTIA.
	AT_FISICA	Nominal	Atividade física realizada regularmente pelo paciente, por exemplo, CAMINHADA.
	HAB_ALIMENT	Nominal	Habito alimentar informado pelo paciente, por exemplo, DIETA MÉDICA
	OCUPAÇÃO	Nominal	Atividade principal exercida pelo paciente, por exemplo, ESTUDANTE.
	FAIXAETARIA	Nominal	Intervalo de valores indicando a faixa etária de um dado paciente, por exemplo, (-inf; 21,5].
	TABAGISMO	Nominal	Fumante ou não fumante.
Meta	MEDICAMENTO	Nominal	Medicamento mais significativo em uso pelo paciente, por exemplo, ANTIDEPRESSIVO.

A Tabela 14 mostra a descrição do conjunto de treinamento para o modelo de previsão, que **VISA CARACTERIZAR O PERFIL DAS CONDUTAS SEGUIDAS/INDICADAS AOS PACIENTES QUE SE SUBMETERAM A CERTO NÚMERO DE CIRURGIAS ESTÉTICAS**.

Tabela 14. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção CfsSubSetEval e mecanismo de busca: BestFirst, meta conduta seguida pelo paciente após submissão a uma ou mais cirurgias estéticas.

	Atributo	Tipo	Descrição
Previsão	ASSIDUIDADE	Nominal	Faixa de valores indicando quantos possíveis procedimentos um dado paciente submeteu-se, por exemplo, (-inf; 3,8].
	VOLUME	Contínuo	Volume de líquido retirado, por exemplo, em uma cirurgia de LIPOASPIRAÇÃO ABDOMINAL.
	LOCALIZAÇÃO	Nominal	Local onde um determinado procedimento foi realizado, ou uma alteração foi encontrada, por exemplo,ABDOMEN.
	TRIMESTRE	Nominal	Trimestre que um dado paciente submeteu-se a um procedimento, por exemplo, TERCEIRO.
Meta	CONDUTA	Nominal	Conduta fisioterapêutica pós-cirúrgica, por exemplo, DRENAGEM LINFÁTICA MANUAL DE ABDOMEN após uma cirurgia de lipoaspiração abdominal.

A Tabela 15 mostra a descrição do conjunto de treinamento para o modelo de previsão, que **VISA CARACTERIZAR A ASSIDUIDADE DOS PACIENTES QUE SE SUBMETERAM A PROCEDIMENTOS REALIZADOS NA CLÍNICA DE ESTÉTICA**. Deve-se destacar que o atributo meta foi derivado do atributo TOTAL de procedimentos que um paciente já realizou na clínica em questão. Em outras palavras, foram criadas faixas de valores a partir do atributo TOTAL. Dar-se-á mais detalhes do processo de discretização mais a frente, em seção especifica.

Tabela 15. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção CfsSubSetEval e mecanismo de busca: BestFirst, meta Assiduidade dos pacientes na clínica de estética.

	Atributo	Tipo	Descrição
Previsão	ALERGIA	Nominal	Alergia mais significativa de um dado paciente, por exemplo, PÓ.
	ANT_PATOL	Nominal	Antecedente patológico mais significativo para um dado paciente, por exemplo, MAMOPLASTIA.
	ATIVIDADEFISICA	Nominal	Atividade física realizada regularmente pelo paciente, por exemplo, CAMINHADA.
	OCUPACAO	Nominal	Atividade principal exercida pelo paciente, por exemplo, ESTUDANTE.
	TABAGISMO	Nominal	Fumante ou não fumante.
Meta	ASSIDUIDADE	Nominal	Faixa de valores indicando quantos possíveis procedimentos um dado paciente submeteu-se, por exemplo, (-inf; 3,8].

A Tabela 16 mostra a descrição do conjunto de treinamento para o modelo de previsão, que visa caracterizar a OCUPAÇÃO dos pacientes que se submeteram aos principais procedimentos de uma clínica de estética.

Tabela 16. Dicionário dos dados da base de exemplos para obtenção dos modelos de previsão: método de seleção CfsSubSetEval e mecanismo de busca: BestFirst, meta OCUPAÇAO de certos pacientes.

	Atributo	Tipo	Descrição
Previsão	ALERGIA	Nominal	Alergia mais significativa de um dado paciente, por exemplo, PÓ.
	ANT_PATOL	Nominal	Antecedente patológico mais significativo para um dado paciente, por exemplo, MAMOPLASTIA.
	FAIXAETARIA	Nominal	Intervalo de valores indicando a faixa etária de um dado paciente, por exemplo, (-inf-21.5].
	ATIVIDADEFISICA	Nominal	Atividade física realizada regularmente pelo paciente, por exemplo, CAMINHADA.
	HABITOSALIMENTARE	Nominal	Situação alimentar de um dado paciente, por exemplo, REEDUCAÇÃO ALIMENTAR.
	MEDICAMENTO	Nominal	Medicamento mais significativo em uso pelo paciente, por exemplo, ANTIDEPRESSIVO.
	TABAGISMO	Nominal	Fumante ou não fumante.
Meta	OCUPAÇÃO	Nominal	Atividade principal exercida pelo paciente, por exemplo, ESTUDANTE.

Relembrando que as diferentes configurações do conjunto de treinamento supra-descritas visam descobrir/explicitar conhecimentos que caracterizam os seguintes perfis:

- das alterações pós-cirúrgicas;
- das alergias relatadas pelos pacientes que submeteram-se a procedimentos realizados em clínica de estética;
- dos pacientes que tomam (ou tomaram recentemente) certos medicamentos;
- das condutas seguidas/indicadas aos pacientes que se submeteram a certo número de cirurgias estéticas;
- da assiduidade dos pacientes se submetem a cirurgias e procedimentos em clínicas de estética; e
- das ocupações dos pacientes que submetem-se a cirurgias/procedimentos estéticos.

3.3 Descoberta de Modelos de Previsão

A partir dos conjuntos de treinamentos preparados de acordo com a metodologia descrita na seção anterior, estabeleceu-se que a compreensibilidade dos modelos gerados é uma característica importante. Optou-se, portanto, em utilizar como base um algoritmo de geração de regras de decisão, o algoritmo JRIP, baseado no C4.5 e disponível na plataforma WEKA (WITTEN &

FRANK, 2000). Este algoritmo gera conjuntos ordenados de regras. Além da escolha do algoritmo JRIP, empregou-se também técnicas de combinação de classificadores, a saber: BAGGING e BOOSTING. A simplicidade das regras facilitou a construção de um módulo de software para a interpretação e aplicação das regras; este software será mostrado mais a frente.

Em seguida os modelos produzidos com os algoritmos JRIP, BAGGING e BOOSTING foram comparados utilizando o procedimento de validação cruzada. Tanto os algoritmos BAGGING e BOOSTING fazem parte da plataforma WEKA; o algoritmo base para ambos os métodos foi o JRIP; em particular, o BAGGING foi configurado para gerar 10 classificadores. Deve-se destacar, que dado o pequeno tamanho da base de dados (4826 instâncias), não foi necessário trabalhar com amostradas de tamanhos reduzidos, como exemplo, 5% ou 10%. Entretanto, caso houvesse a necessidade o procedimento de amostragem utilizado seria o de estratificação com reposição (BOLFARINE & BUSSAB, 2005).

3.3.1 Resultados dos Modelos de Classificação Descobertos

Todos os experimentos, descritos nesta seção, foram realizados considerando todas as instâncias da base de dados; aproximadamente 4826. Como já dito na seção anterior, não foi necessário usar nenhum processo de amostragem para reduzir o número de instâncias do conjunto de treinamento. O leitor interessado em um estudo envolvendo aprendizagem de máquina e o impacto do procedimento de amostragem na precisão dos classificadores pode consultar os seguintes trabalhos Lopes (2007) e Kobus, et al.,(2009) realizados no PPGTS/PUCPR.

A Tabela 17 resume as taxas de acerto dos métodos JRIP, BAGGING e BOOSTING. Cada célula (exemplo célula em destaque na cor cinza) desta tabela deve ser lida da seguinte forma: (i) o algoritmo JRIP produziu uma taxa de classificações corretas de 90,83%, considerando 100% dos atributos previsores do conjunto de treinamento T para o atributo meta ALERGIA; e (ii) o algoritmo BAGGING produziu uma taxa de classificações corretas de 85,42%, considerando 21% dos atributos previsores do conjunto de treinamento T também para o atributo meta ALERGIA; e assim por diante para os demais valores.

Tabela 17. Iterações de validação cruzada para conjuntos de treinamentos (T) com números de atributos previsores e metas diferentes em percentuais (%).

ATRIBUTOS METAS	ALERGIA		TIPO DE ALTERAÇÃO		ASSIDUIDADE	
% DE ATRIBUTOS PREVISORES	T:100%	T:21%	T:100%	T:30%	T:100%	T:17%
% DE ACERTO-JRIP	90,83	85,90	86,56	83,00	98,50	85,90
% DE ACERTO-BAGGING	91,17	85,42	87,30	83,66	98,62	98,62
% DE ACERTO-BOOSTING	91,23	85,42	86,56	83,00	99,77	98,62

ATRIBUTOS METAS	CONDUTA		MEDICAMENTO		OCUPAÇÃO	
% DE ATRIBUTOS PREVISORES	T:100%	T:21%	T:100%	T:35%	T:100%	T:35%
% DE ACERTO-JRIP	91,33	91,09	90,45	89,99	94,99	93,81
% DE ACERTO-BAGGING	92,11	91,41	90,63	90,25	95,51	94,25
% DE ACERTO-BOOSTING	92,17	91,09	90,77	90,15	95,71	93,81

A Tabela 18 resume as taxas de erro médio e desvio padrão referente aos dados da Tabela 17. A coluna C1 representa as taxas de erro médio e desvio padrão dos classificadores gerados a partir dos conjuntos de treinamento completos (T: 100%). A coluna C2 representa as taxas de erro médio e desvio padrão dos classificadores gerados a partir dos conjuntos de treinamento reduzidos por filtragem de atributos. A coluna C3 representa as taxas de erro médio e desvio padrão por método. Finalmente, a taxa de erro médio geral foi de 9,06% e desvio padrão de $\pm$ 4,67%.

Tabela 18. Taxas de erro médio e desvios padrão.

MÉTODO	C1	C2	C3
JRIP	7,89 $\pm$3,76	11,72 $\pm$3,67	9,80 $\pm$4,18
BAGGING	8,45 $\pm$3,63	9,40 $\pm$5,06	8,92 $\pm$4,90
BOOSTING	7,30 $\pm$4,14	9,65 $\pm$5,15	8,47 $\pm$4,82

Com os resultados apresentados na Tabela 17 pode-se observar que os métodos BAGGING e BOOSTING melhoraram o percentual de exemplos classificados corretamente, sendo BOOSTING o melhor dentre eles para os experimentos realizados. Da mesma forma, pode-se observar na Tabela 18 que as menores taxas de erro médio e desvio padrão são do método BOOSTING. Assim, pode-se concluir que as técnicas de combinação de classificadores podem ser eficazes em situações onde os atributos disponíveis/selecionados correspondem apenas a um pequeno número, visto que as diferenças entre os valores da coluna C1 e C2 não são grandes. Estes

resultados também foram verificados em outros trabalhos realizados em nosso grupo de pesquisa (LOPES, 2007).

3.4 Avaliação Subjetiva de Padrões Interessantes

A avaliação subjetiva dos padrões obtidos foi realizada por meio da avaliação da qualidade das regras obtidas, conforme a visão de um especialista/prático atuante em clínica de estética. Dentre as regras obtidas, um número reduzido para cada conjunto de treinamento é apresentado na seqüência, visando facilitar a exposição. De imediato pode-se observar que as regras apontam para um paciente enquadrado como sendo: àquele que sempre apresenta alguma alteração pós-cirúrgica; procura pelo procedimento preferencialmente no 3º ou no 4º trimestre de cada ano; quando ele é usuário de algum medicamento, a recorrência é sobre os antidepressivos, anti-hipertensivos e anticoncepcionais; a principal conduta pós-cirúrgica é drenagem linfática manual; os pacientes que se encaixam no menor número de procedimentos e realizam preferencialmente mais próximo do final de ano; há uma boa parcela dos pacientes cuja ocupação é estudante e possui hábitos de vida saudável com idade menor de 22 anos.

As próximas tabelas mostram algumas regras agrupadas por atributo meta. Pode-se observar, por exemplo, nas regras da Tabela 21, que os medicamentos mais usados são: antidepressivo, anti-hipertensivo e anticoncepcional.

Tabela 19. Exemplo de regras obtidas tendo como atributo meta ALERGIA.

R1	**SE** (ATIVIDADEFISICA = PILATES) **ENTÃO** ALERGIA=PO (284.0/0.0)
R2	**SE** (OCUPACAO = ESTUDANTE) E (ANTECEDENTES = RETIROU-AMIGDALA-E-ADENOIDE) **ENTÃO** ALERGIA=DIPIRONA (40.0/0.0)
R3	**SE** (OCUPACAO = ADVOGADA) **E** (MEDICAMENTO = ANTICONCEPCIONAL) **E** (TRIMESTRE = 3) **ENTÃO** ALERGIA= NAO (40.0/0.0)
R4	**DEFAULT** ALERGIA=NAO (3275.0/183.0)

Tabela 20. Exemplo de regras obtidas tendo como atributo meta ALTERAÇÃO.

R1	**SE** (LOCALIZACAO = COXA-INTERNA-DIREITA) **ENTÃO** ALTERACAO=HEMATOMA-1 (58.0/0.0)
R2	**SE** (LOCALIZACAO = COXA-INTERNA-ESQUERDA) **ENTÃO** ALTERACAO=HEMATOMA-1 (45.0/0.0)
R3	**SE** (TAMANHO = PEQUENO) **E** (LOCALIZACAO = MAMA-ESQUERDA) **ENTÃO** ALTERACAO=FIBROSE (53/10)
R4	**SE** (COR = ROXO) **ENTÃO** ALTERACAO=HEMATOMA (65.0/0.0)
R5	**SE** (TAMANHO = PEQUENO) **ENTÃO** ALTERACAO=HEMATOMA (58.0/0.0)

Tabela 21. Exemplo de regras obtidas tendo como atributo meta MEDICAMENTO.

R1	**SE** (OCUPACAO = ASSISTENTE-CONTABIL) **ENTÃO** MEDICAMENTO=ANTIDEPRESSIVO (324.0/0.0)
R2	**SE** (ANTECEDENTES = NAO) **E** (ALERGIA = NAO) **E** (OCUPACAO = MOTO-GIRL) **ENTÃO** MEDICAMENTO=ANTICONCEPCIONAL (338.0/0.0)
R3	**SE** (ANTECEDENTES = HISTERECTOMIA) E (ATIVIDADEFISICA = PILATES) **ENTÃO** MEDICAMENTO=ANTIHIPERTENSIVO (284.0/0.0)
R4	**SE** (OCUPACAO = SERVENTE) **ENTÃO** MEDICAMENTO=ANTIDEPRESSIVO (191.0/17.0)

Tabela 22. Exemplo de regras obtidas tendo como atributo meta CONDUTA.

R1	**SE** (ASSIDUIDADE = '(3.8-6.6]') **E** (LOCALIZACAO = COSTAS-DORSO) **ENTÃO** CONDUTA=DLM-DE-COSTAS (149.0/0.0)
R2	**SE** (LOCALIZACAO = PALPEBRA) E (ASSIDUIDADE = '(-inf-3.8]') **ENTÃO** CONDUTA=DLM-DE-FACE (181.0/0.0)
R3	**SE** (ASSIDUIDADE = '(9.4-12.2]') **E** (VOLUME = '(-inf-500]') **ENTÃO** CONDUTA=DLM-DE-MEMBROS-INFERIORES (129.0/0.0)

Tabela 23. Exemplo de regras obtidas tendo como atributo meta ASSIDUIDADE.

R1	**SE** (CONDUTA = DLM-DE-COSTAS) **ENTÃO** ASSIDUIDADE ='(3.8-6.6]' (234.0/0.0)
R2	**SE** (CONDUTA = DLM-DE-MAMA) **ENTÃO** ASSIDUIDADE ='(3.8-6.6]' (136.0/0.0)
R3	**SE** (CONDUTA = DLM-DE-MEMBROS-INFERIORES) **ENTÃO** ASSIDUIDADE ='(3.8-6.6]' (83.0/0.0)
R4	**DEFAULT** ASSIDUIDADE ='(6.6-9.4]' (1135/100)

Tabela 24. Exemplo de regras obtidas tendo como atributo meta OCUPAÇÃO.

R1	**SE** (MEDICAMENTO = ANTIDEPRESSIVO) **E** (ANTECEDENTES = NAO) **ENTÃO** OCUPACAO=ASSISTENTE-CONTABIL (324.0/0.0)
R2	**SE** (TABAGISMO = SIM) **E** (ANTECEDENTES = NAO) **ENTÃO** OCUPACAO=MOTO-GIRL (338.0/0.0)
R3	**SE** (MEDICAMENTO = NAO) **E** (ATIVIDADEFISICA = NAO) **E** (ANTECEDENTES = CESARIA) **E** (TABAGISMO = NAO) **ENTÃO** OCUPACAO=CABELEIREIRA (100.0/2.0)
R4	**DEFAULT** OCUPACAO=PROMOTORA (393.0/64.0)

Em resumo, as informações relevantes observadas após a execução dos diferentes objetos em termos de descoberta de conhecimentos são:

- os pacientes quando usurários de algum medicamento, que não anticoncepcional, a recorrência é sobre os antidepressivos ou anti-hipertensivos;
- as alterações, quando ocorrem, são em geral: hematoma e fibrose, proporcionalmente nesta seqüência;
- a conduta pós-cirúrgica mais empregada é a drenagem linfática, seja ela em decorrência de um procedimento cirúrgico realizado nas costa, nas coxas, no abdômen, nas mamas de uma paciente; e
- o número de eventos cirúrgicos e/ou fisioterapêutico que uma paciente realiza está entre quatro e sete ou entre sete e dez.

3.4.1 Transformação de Atributos

Relembrando que o nosso principal objetivo é descobrir regras que minimizem o tempo e o esforço de um tratamento fisioterapêutico pós-cirúrgico. Entretanto, há alguns atributos da base de treinamento que eram inicialmente contínuos e foram usados em nossos experimentos como atributos metas, por exemplo, Total (quantidade de procedimentos/eventos que um paciente submeteu-se). A discretização dos valores foi alvo de experimentos, à medida que os algoritmos usados não geram classificadores quando o atributo meta é um valor contínuo. Sendo assim, foram feitas as discretizações das variáveis contínuas, onde Total foi representada por Assiduidade. Duas formas diferentes de realizar a discretização foram utilizadas, a saber: a discretização dinamicamente feita pelo próprio algoritmo de classificação e a discretização estática feita *off-line*. A discretização estática permite, por exemplo, que sejam escolhidos a quantidade de intervalos (*bins*), o tamanho dos intervalos e a freqüência de itens em cada intervalo (HUSSIAN, et al. 1999).

3.5 Modelo de Interpretação das Regras Descobertas/Explicitadas

O sistema PATY é composto por cinco módulos principais. O módulo Operacional (*OP)* é responsável por registrar as operações do dia-a-dia da clínica. O módulo *d/w* é responsável em importar os dados gerados pelo *OP* e disponibilizar tais dados em um esquema multidimensional. O módulo OLAP é responsável pela geração de relatórios dinâmicos. O módulo *MP* é responsável pela geração dos modelos de previsão. E finalmente, o módulo *MIR* é responsável por interpretar as regras geradas pelos classificadores.

O módulo *MIR* recebe como entrada um arquivo texto contendo tais regras (cf. Figura 20). O núcleo deste módulo permite aplicar diferentes conjuntos ordenados de regras. Por exemplo, um evento *e* é submetido respectivamente aos modelos de previsão: MP1, MP2 e MP3. Sendo que MP1 é um modelo simples gerado pelo JRIP, MP2 é um modelo composto gerado pelo BAGGING+JRIP e MP3 é um modelo composto gerado pelo BOOSTING+JRIP. A Figura 21 mostra o modelo conceitual do *MIR*.

```
JRIP rules:
===========

(ATIVIDADEFISICA = BICICLETA) and (HABITOSALIMENTARE = NORMAL) => MEDICAMENTO=PREDINIZONA (36.0/4.0)
(OCUPACAO = COMERCIANTE) and (TRIMESTRE = 1) and (ANTECEDENTES = HISTERECTOMIA) => MEDICAMENTO=FLORAIS (21.0/0.0)
(OCUPACAO = DO-LAR) and (TRIMESTRE = 2) and (ATIVIDADEFISICA = CAMINHADA) => MEDICAMENTO=DIURETICO (30.0/0.0)
(OCUPACAO = EMPRESARIA) and (ALERGIA = INSETO) and (CONDUTA = ?) => MEDICAMENTO=RAUCUTAM (42.0/0.0)
(OCUPACAO = DO-LAR) and (ANTECEDENTES = MENTOPLASTIA) => MEDICAMENTO=SINVASTATINA (31.0/4.0)
(OCUPACAO = DO-LAR) and (ANTECEDENTES = GASTROPLASTIA) and (ALERGIA = NAO) => MEDICAMENTO=SINVASTATINA (15.0/1.0)
(OCUPACAO = BANCARIA) and (TRIMESTRE = 3) => MEDICAMENTO=FENOPREPOREX-SUBTRAMINA (42.0/14.0)
(ANTECEDENTES = LIPOASPIRACAO) and (ALERGIA = PENICILINA) => MEDICAMENTO=FENOPREPOREX-SUBTRAMINA (17.0/0.0)
(FAIXAETARIA = (39-45]) and (ANTECEDENTES = CESARIA) and (CONDUTA = ?) => MEDICAMENTO=ANTICONCEPCIONAL (19.0/0.0)
(OCUPACAO = PROMOTORA) and (TRIMESTRE = 1) and (ANTECEDENTES = CESARIA) and (ATIVIDADEFISICA = MUSCULACAO) and
(CONDUTA = ?) => MEDICAMENTO=ANTICONCEPCIONAL (8.0/0.0)
[ .... ]
  => MEDICAMENTO=NAO (1954.0/268.0)
```

Figura 20. Exemplo de regras no formato JRIP/WEKA.

O modelo exibido na Figura 21 define um conjunto de conceitos. De forma prática, podemos lê-lo da seguinte maneira: um *MIR* pode ter diferentes camadas de decisão. Cada camada pode ter diferentes classificadores. Cada classificador pode ter diferentes regras. Cada regra tem um conseqüente e um antecedente. O antecedente de uma regra pode ser vazio para o caso da regra *default*. Cada clausula do antecedente de uma regra é formada por um operador relacional, envolvendo sempre uma variável e uma constante. O conseqüente de uma regra é uma constante. Cada camada pode receber um evento *e*. Este último é persistido na forma de um registro de dados no Sistema PATY, realizado por meio do módulo *OP*. Este registro é transformado em um exemplo/vetor *p* a ser classificado. Finalmente, tal exemplo *p* é submetido aos diferentes classificadores das diferentes camadas do módulo *MIR*. Na aplicação em questão, cada camada reúne três classificadores, gerados respectivamente pelos métodos JRIP, BAGGING+JRIP, BOOSTING+JRIP.

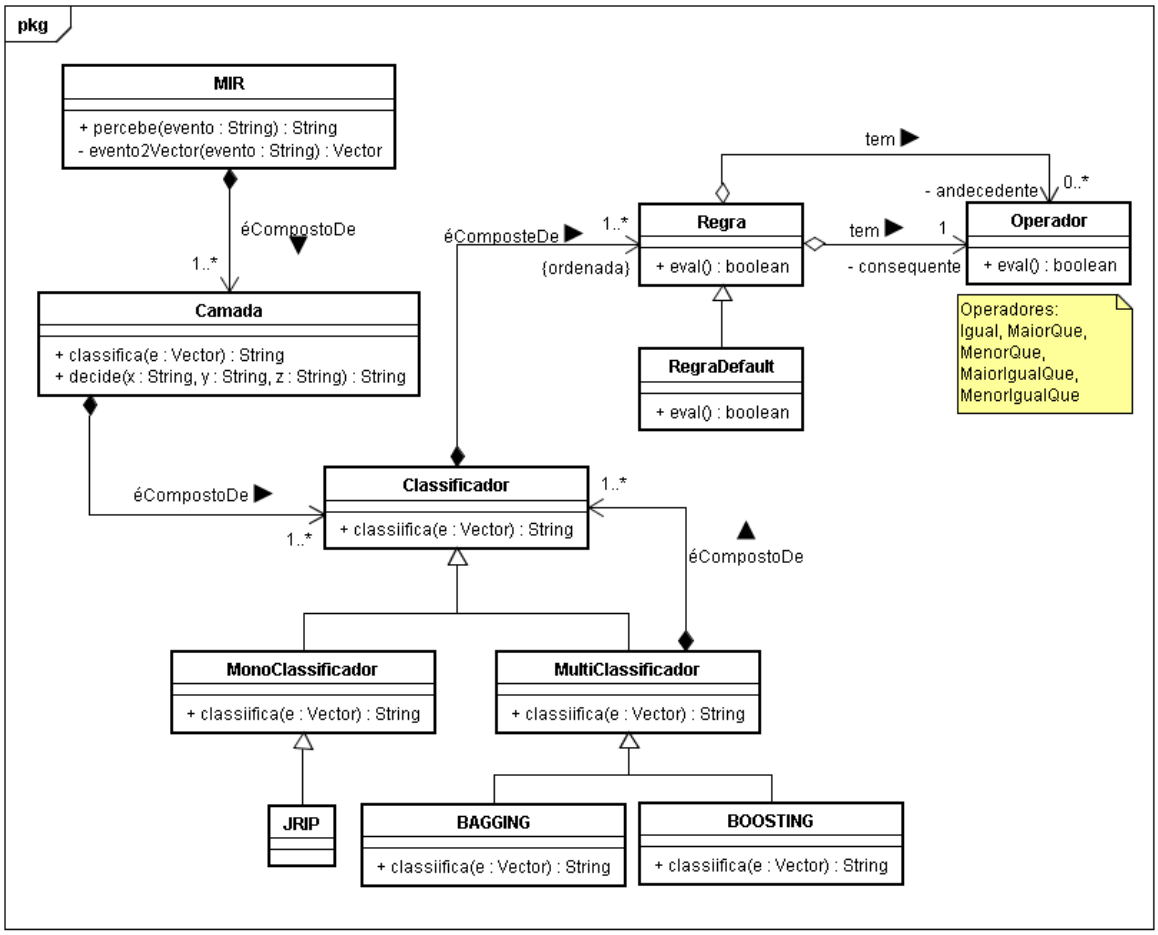

Figura 21. Modelo conceitual do interpretador de regras (MIR).

Relembrando, têm-se da Tabela 11 a Tabela 16 os dicionários de dados para cada objetivo/atributo meta definido em nossos experimentos. Para cada atributo meta foram gerados seis classificadores, dos quais foram selecionados os três melhores, em termos de taxa de acerto, para compor uma das camadas de decisão. A Figura 22 ilustra um cenário envolvendo uma camada chamada de *assiduidade* e três classificadores: c1, c2 e c3, onde c1 é um classificador gerado a partir do JRIP, c2 é um classificador gerador a partir do método BAGGING e c3 é um classificador gerado a parir do método BOOSTING. Como haverá três respostas, que podem ser distintas, será adotada a resposta do classificador que possuir maior taxa de acerto. Este cenário pode ser visto como o cenário para a estruturação em torno do atributo meta ASSIDUIDADE (Tabela 15).

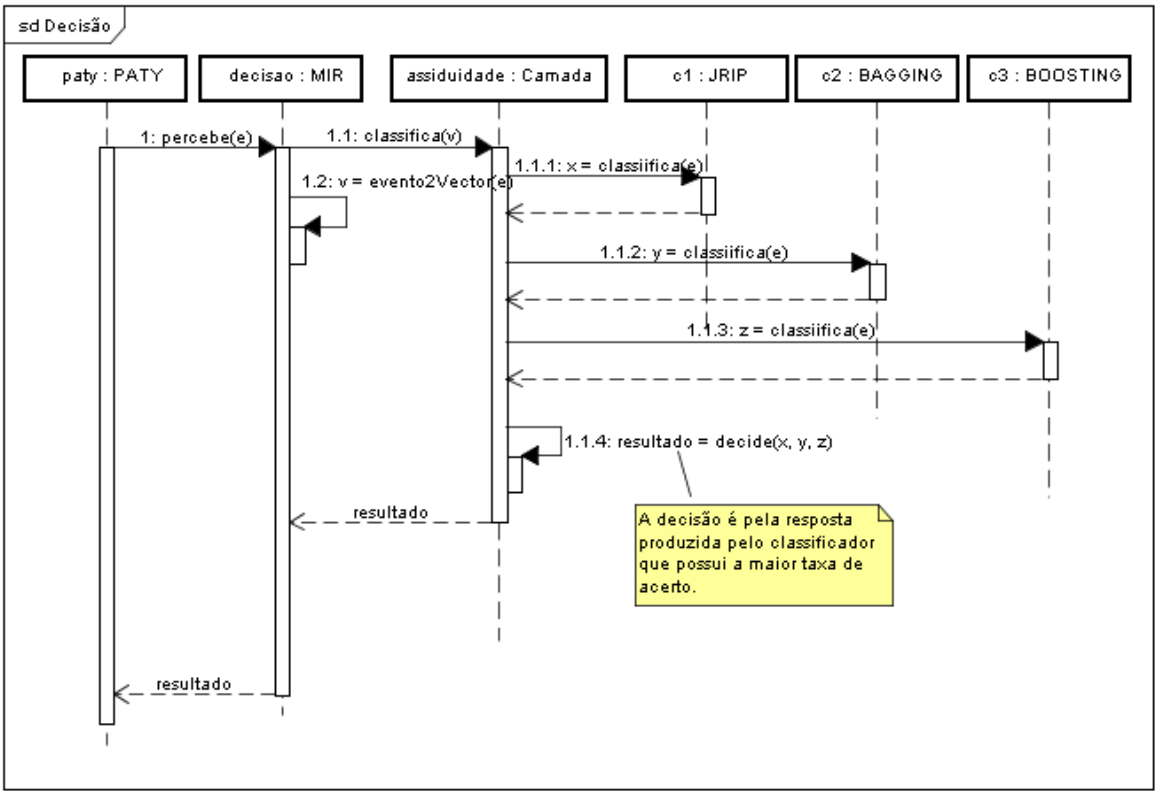

Figura 22. Cenário padrão de classificação de um evento, envolvendo sempre um classificador JRIP, um classificador BAGGING e um classificador BOOSTING.

A representação completa do cenário encerrar-se-ia com a inclusão, na Figura 22, de mais cinco objetos do tipo Camada, ou seja, um objeto/camada adicional para um dos seguintes atributos meta:

- deALTERAÇÃO, visa a descobrir/explicitar os conhecimentos que caracterizam as alterações pós-cirúrgicas (Tabela11);
- ALERGIA, visa a descobrir/explicitar as alergias relatadas pelos pacientes que submetem-se a procedimentos estéticos (Tabela 12);
- MEDICAMENTO, visa a descobrir/explicitar os conhecimentos que caracterizam os pacientes que tomam (ou tomaram recentemente) certos medicamentos (Tabela 13);
- CONDUTA, visa a descobrir/explicitar os conhecimentos que caracterizam as condutas seguidas/indicadas aos pacientes que se submeteram a certo número de cirurgias e procedimentos estéticos (Tabela 14); e
- OCUPAÇÃO, visa descobrir explicitar a ocupação dos pacientes e realizam cirurgias em clínicas de estética (Tabela 15).

A Figura 23 mostra, na forma de um diagrama UML, as principais atividades e fluxos de objetos entre os módulos do Sistema PATY. De forma resumida, um usuário preenche um formulário descrevendo um evento e solicita o registro do mesmo ao subsistema Modelo OP. Este último registra-o na base de dados operacional. O fluxo continua e os dados são repassados ao d/w. O Módulo Md/w registra os dados passados de acordo com o esquema multidimensional do d/w. O fluxo continua e os dados do evento no novo esquema são repassados para serem classificados, segundo o conjunto de objetivos gerenciado pelo módulo MIR. Este último realiza as transformações necessárias para que o evento seja corretamente classificado para cada objetivo gerenciado pelo MIR; finalmente, cada classificação é reunida em um único relatório e mostrada ao usuário.

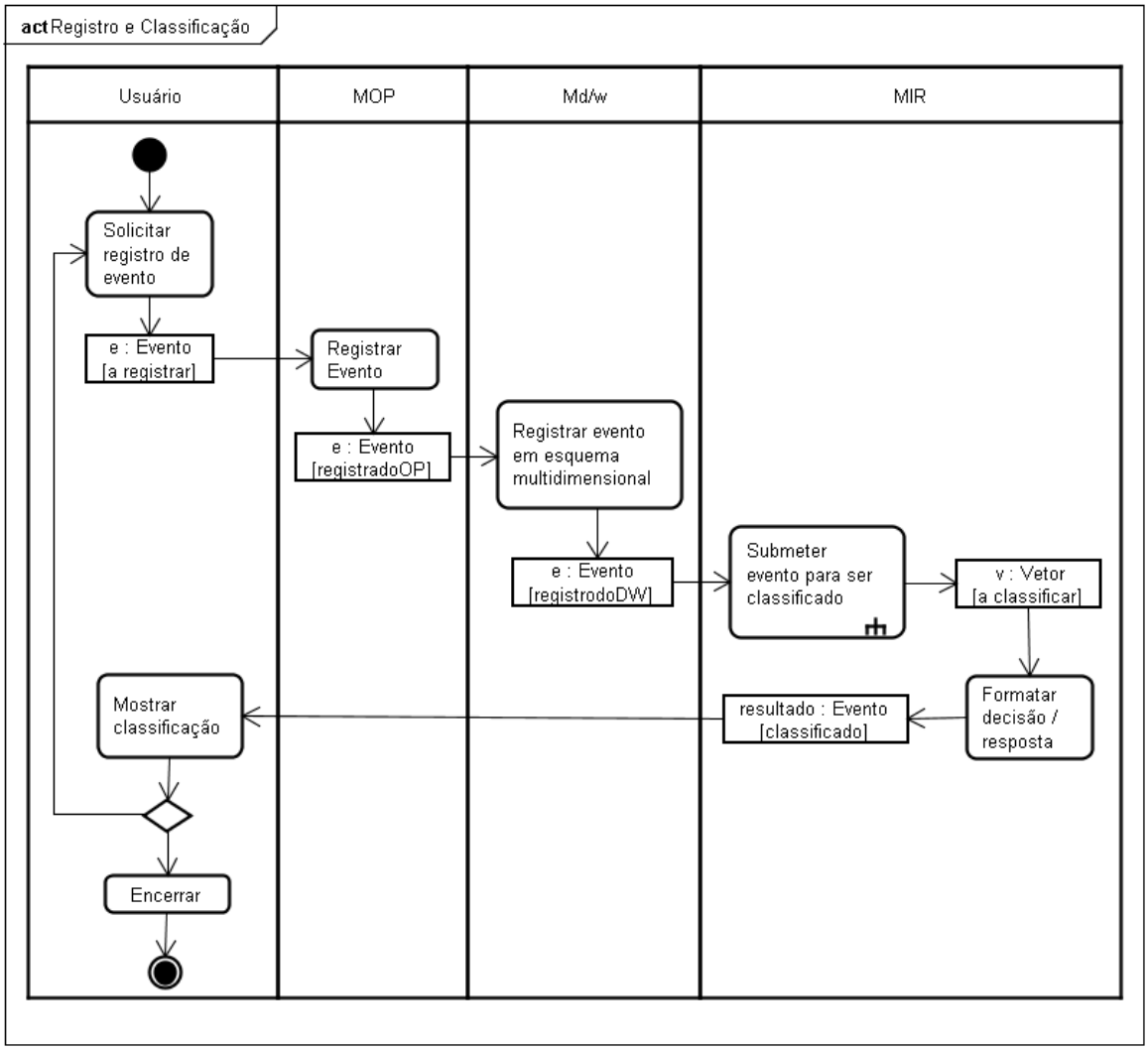

Figura 23. Cenário padrão de registro e classificação de um evento no Sistema PATY.

A Figura 24 mostra, na forma de um diagrama UML, as principais atividades e fluxos de objetos entre os classificadores do módulo MIR; tal esquema representa o detalhamento da atividade "*Submeter evento para ser classificado*" presente no diagrama da Figura 23.

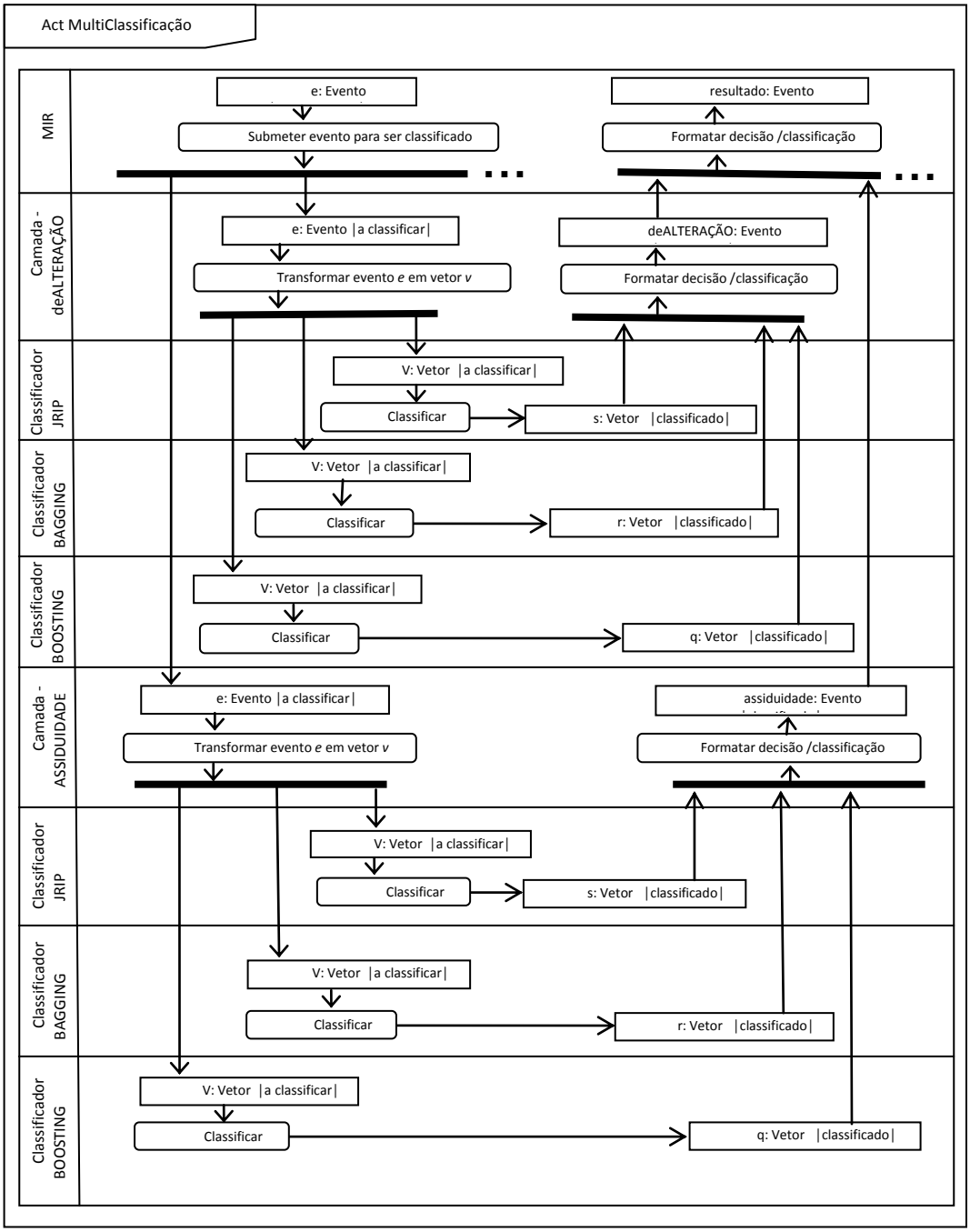

Figura 24. Cenário padrão de classificação de um evento no Sistema PATY multi-objetivo.

A Figura 24 mostra parcialmente o esquema de classificação multi-objetivo do Sistema PATY. Para completar seria necessário incluir as seguintes camadas com os seus classificadores específicos: ALERGIA, MEDICAMENTO, CONDUTA E OCUPAÇÃO.

3.6 Considerações Finais

A metodologia apresentada segue, em linhas gerais, as etapas clássicas do processo de descobrimento de conhecimento e modelagem de dados multidimensional. Esta última mostra-se interessante à medida que se podem explorar os dados por meio de ferramentas de OLAP. Tal navegação nos dados ajuda também a formulação de objetivos a serem explorados por meio de ferramentas de aprendizagem de máquina.

Em termos práticos, definiu-se também uma forma interativa e flexível de usar os conhecimentos descobertos/explicitados. Em outras palavras, à medida que o histórico do paciente é formado, o usuário, obtém informações parciais a partir de classificações feitas pelo conjunto de classificadores do módulo MIR. No início da montagem do histórico do paciente, as respostas dos classificadores são baseadas, em particular, nas regras default de cada classificador. Já com o incremento do histórico do paciente as respostas dos classificadores tendem a uma maior taxa de acerto, à medida que as regras mais específicas são aplicadas.

4 RESULTADO E DISCUSSÃO

Relembrando que o trabalho proposto visa descobrir padrões, a partir de dados coletados por meio de diferentes fontes em uma clínica de cirurgia plástica, para ajudar no planejamento e na execução de um bom tratamento fisioterapêutico para cada paciente. A consecução, deste objetivo, inclui diferentes tarefas não triviais, a saber: (i) um estudo detalhado sobre os dados coletados provenientes de diferentes bases especialistas atuantes na clínica; (ii) a execução de inúmeros experimentos visando extrair conhecimentos úteis para o processo de planejamento e execução de um bom tratamento fisioterapêutico; e (iii) uma análise criteriosa para validação das regras, realizado por meio da comparação das taxas de acerto dos classificadores.

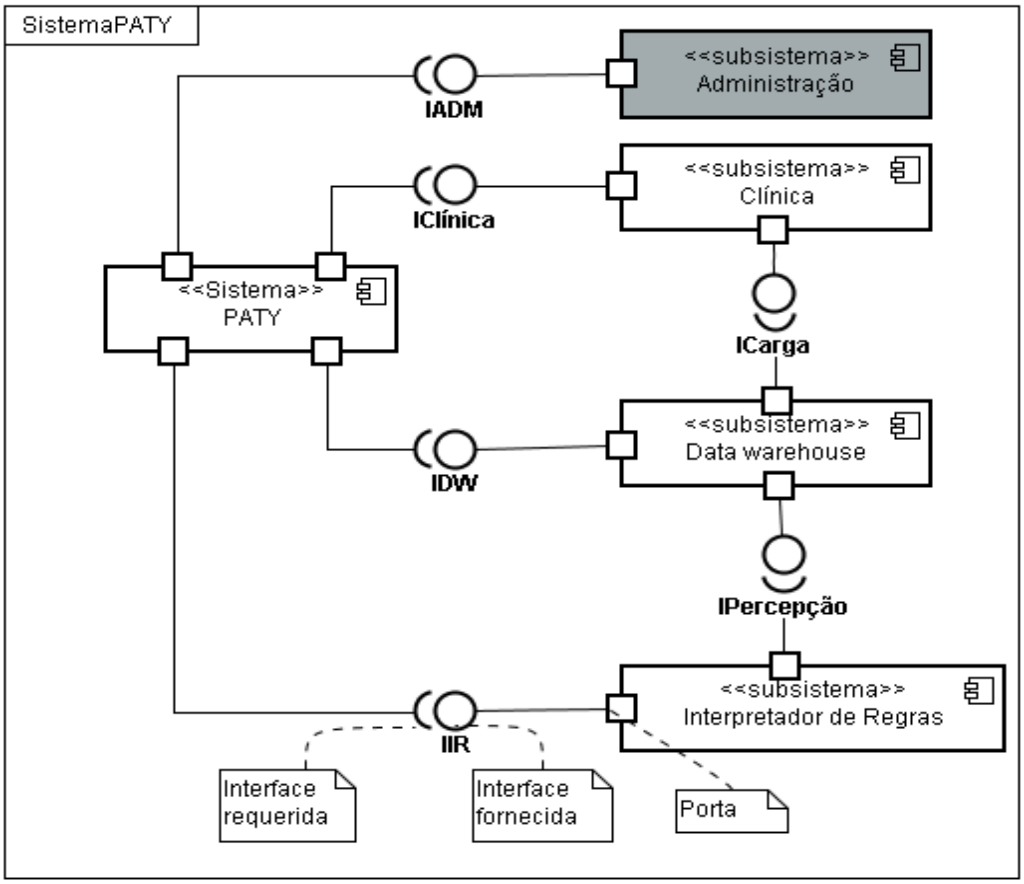

Figura 25. Sistema PATY e seus componentes.

A Figura 25 ilustra, na forma de um diagrama de componentes UML, o conjunto de módulo do Sistema PATY; o módulo na cor cinza não foi desenvolvido. Importante destacar que módulo *MOP* fornece uma interface, denominada *ICarga*, usada pelo módulo d/w para fazer a carga da base de dados multidimensional. Ainda destaca-se o módulo d/w que fornece uma interface, denominada *IPercepção*, usada pelo módulo *MIR* para recuperar/perceber os novos eventos registrados na base de dados multidimensional a serem classificados.

Como já visto anteriormente, o funcionamento geral do sistema, ilustrado por meio do esquema da Figura 23 e o funcionamento detalhado do módulo de interpretação, ilustrado por meio do esquema da Figura 24, mostra um comportamento interativo; este último é um dos elementos importantes dos resultados do projeto em questão. O cenário a seguir ilustra tal dinâmica.

O fluxo ilustrado na Tabela 25 é o seguinte: à medida que o histórico do paciente com a clínica é construído, o módulo de MIR, submete cada novo evento gerado aos diferentes classificadores; os resultados/classificações são apresentados ao usuário a cada novo evento. Desta forma, o usuário constrói pouco a pouco uma idéia, mesmo que parcial, do paciente. Isto deve ajudar o usuário a planejar o acompanhamento do paciente. O cenário a seguir ilustra tecnicamente como os dados são formatados e repassados ao módulo MIR, o qual classifica um dado evento (e.g., coluna1 da Tabela 25) e devolve ao usuário um relatório (coluna1 da Tabela 26).

Em resumo, a Tabela 25 apresenta oito eventos. O evento com o rótulo 1 refere-se ao primeiro registro “anamnese” no sistema de um paciente. Na seqüência os eventos com os rótulos de 2 a 6 referem-se a diferentes cirurgias. Os dois últimos eventos referem-se às alterações pós-cirúrgicas.

Tabela 25. Exemplos de eventos/vetores classificados de um dado paciente.

ATRIBUTOS	EVENTOS							
	ANAMNESE	CIRURGIA					ALTERACAO	
	1	2	3	4	5	6	7	8
CODPAC	1	1	1	1	1	1	1	1
SEQ	(-inf-3.8]	(-inf-3.8]	(-inf-3.8]	(3.8-6.6]	(3.8-6.6]	(3.8-6.6]	(-inf-3.8]	(-inf-3.8]
ALERGIA	NÃO	NÃO	NÃO	NAO	NAO	NAO	NAO	NÃO
ALTERACAO	?	?	?	?	?	?	HEMATOMA	FIBROSE
ANTECEDENTE	AMIG.	AMIG.	AMIG.	AMIG.	AMIG.	AMIG.	AMIG.	AMIG.
A.FISICA	NÃO	NÃO	NÃO	NAO	NAO	NAO	NAO	NÃO
CIRURGIA	?	ABDOMEN	BLEFARO	C.CICATRIZ	DORSOP	P.LABIOS	?	?
CONDUTA	?	DLM-R-ABD	DLM-F	DLM-C	DLM-C	DLM-M	?	?
VOLUME	0	2200	0	0	650	0	0	0
COR	?	?	?	?	?	?	AMARELO	?
DOR	?	?	?	?	?	?	?	?
H.ALIMENTARE	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL	NORMAL
INTENSIDADE	?	?	?	?	?	?	?	?
LOCALIZACAO	?	ABDOMEN	PALPEBRA	CICATRIZ-U	C.DORSO	P.-LABIOS	COCIX	ABDOMEN
MEDICAMENTO	ANTIC	ANTIC	ANTIC	ANTIH	ANTIH	ANTIH	ANTIH	ANTIH
OCUPACAO	ADV	ADV	ADV	ADV	ADV	ADV	ADV	ADV
FAIXAETARIA	(21-27]	(27-33]	(27-33]	(27-33]	(27-33]	(21-27]	(21-27]	(21-27]
TRIMESTRE	1	1	1	1	1	1	1	1
PARECER	?	?	?	?	?	?	?	?
PERIMETRIA	?	?	?	?	?	?	?	?
QUEIXA	?	?	?	?	?	?	?	?
TABAGISMO	NÃO	NÃO	NÃO	NAO	NAO	NAO	NAO	NÃO
TAMANHO	?	?	?	?	?	?	MEDIO	PEQUENO
T.CIRURGICA	?	CLASSICA	CLASSICA	CLASSICA	CLASSICA	CLASSICA	?	?
T.FISIO	?	?	?	?	?	?	?	?
QUANTIDADE	1	1	1	1	1	1	1	1
ANO	2008	2008	2008	2008	2008	2008	2008	2008
MÊS	3	3	3	3	3	3	3	3
DIA	5	5	5	5	5	5	5	5

LEGENDA: ANTIC=ANTICONCEPCIONAL, DLMM=DRENAGEM LINFÁTICA MANUAL DE MAMA, DLMRA=DRENAGEM LINFÁTICA MANUAL REVERSA DE ABDOMEN, DLMC=DRENAGEM LINFÁTICA MANUAL DE COSTAS.

A Tabela 26 mostra as classificações dos eventos da Tabela 25, os quais foram submetidos a seis classificadores identificados por meio da primeira coluna da Tabela 26. Esta última tabela lê-se da seguinte forma: a informação “ANTIC” do cruzamento da linha MEDICAMENTO e da coluna 1

corresponde ao resultado da classificação do evento 1 vis-à-vis ao classificador que possui como atributo preditor MEDICAMENTO, ANTIC é anticoncepcional classificado como correto.

Tabela 26. Resultado da classificação dos eventos da Tabela 25 – classificadores BOOSTING.

CLASSIFI-CADORES	CLASSIFICAÇÕES							
	CONVERSA	CIRURGIA					ALTERACAO	
	1	2	3	4	5	6	7	8
ALERGIA	Não	Não	Não	Não	Não	Não	Não	Não
ALTERAÇÃO	?	?	?	?	?	?	HEMATOMA	FIBROSE
ASSIDUIDADE	(-inf-3.8	(-inf-3.8	(-inf-3.8	(3.8-6.6]	(3.8-6.6]	(3.8-6.6]	(-inf-3.8	(-inf-3.8
CONDUTA	?	DLM-REVERSA-DE-ABDOMEN	DLM-DE-FACE	DLM-DE-COSTAS	DLM-DE-COSTAS	DLM-DE-MAMA	?	?
MEDICAMENTO	ANTIC	ANTIC	ANTIC	ANTIH	ANTIH	ANTIH	ANTIH	ANTIH
OCUPAÇÃO	ADV	ADV	ADV	ADV	ADV	ADV	ADV	ADV

LEGENDA: DLMM=DRENAGEM LINFÁTICA MANUAL DE MAMA, DLMA=DRENAGEM LINFÁTICA MANUAL DE ABDOMEN, DLMC=DRENAGEM LINFÁTICA MANUAL DE COSTAS, SA=SEM ALTERAÇÃO, ADV=ADVOGADA

Em resumo, a Tabela 26 mostra que à medida que os eventos ocorrem à quantidade de informações/classificações mais confiáveis também acontecem. Devem-se observar as classificações do primeiro evento, onde há certamente ausência de informação para possíveis classificações corretas. Em outras palavras, o primeiro evento fornece apenas informação da anamnese com conseqüente default e, no entanto, o classificador CONDUTA não conseguiu classificá-la, porém no segundo evento já sugeriu DLMReversa de Abdomen, e assim sucessivamente conforme o número de eventos foi aumentando. O mesmo ocorreu com o classificador deALTERAÇÃO, que sugeriu nas primeiras classificações SA (SEM ALTERAÇÃO) e, na seqüência, os hematomas ocorreram. São estas informações sugeridas que permitem ao profissional montar de forma progressiva as suas hipóteses para planejar e acompanhar a evolução de um paciente.

Em particular, o classificador ASSIDUIDADE informou por meio da anamnese que se tratava de uma paciente que se submeteria a *quatro* ou mais procedimentos. E, além disso, que a conduta seria *drenagem linfática manual reversa de abdomen*.

A Tabela 27 mostra a contabilidade das classificações corretas (T) e não corretas (F) de dois pacientes para seis classificadores, representados pelos nomes das colunas. Aqui deve-se observar que as classificações não corretas acontecem nos primeiros eventos, onde

potencialmente são as regras *defaults* as avaliadas e, posteriormente, conforme o número de eventos aumenta, as classificações tendem a apresentar valores corretos (T).

Tabela 27. Cruzamento da contabilidade de classificações corretas e não corretas no tempo.

Contar de CLASSIFICAÇÃO	Rótulos de Coluna						
Rótulos de Linha	ALERGIA	ALTERAÇÃO	ASSIDUIDADE	CONDUTA	MEDICAMENTO	OCUPAÇÃO	Total geral
BOOSTING – MÉTODO	16	16	16	16	16	16	96
1 – PACIENTE	8	8	8	8	8	8	48
2008 – ANO	8	8	8	8	8	8	48
7 º MÊS	4	4	4	4	4	4	24
5 – DIA	1	1	1	1	1	1	6
F				1	1		2
T	1	1	1			1	4
16 – DIA	3	3	3	3	3	3	18
F				2	1		3
T	3	3	3	1	2	3	15
8º MÊS	4	4	4	4	4	4	24
23 – DIA	4	4	4	4	4	4	24
F				1	1		2
T	4	4	4	3	3	4	22
2 – PACIENTE	8	8	8	8	8	8	48
2008 – ANO	8	8	8	8	8	8	48
7º MÊS	4	4	4	4	4	4	24
15 – DIA	4	4	4	4	4	4	24
F				3	2		5
T	4	4	4	1	2	4	19
8º MÊS	4	4	4	4	4	4	24
8 – DIA	3	3	3	3	3	3	18
T	3	3	3	3	3	3	18
15 – DIA	1	1	1	1	1	1	6
F				1	1		2
T	1	1	1			1	4
Total geral	**16**	**16**	**16**	**16**	**16**	**16**	**96**

Nota: T – Correta; F – Não Correta

Lendo a Tabela 27 constata-se que o paciente 1 teve quarenta e oito eventos classificados no ano de 2008 pelo método BOOSTING, onde vinte e quatro eventos foram realizados no mês 07 e vinte e quatro no mês 08, no mês 07, seis eventos foram classificados no dia 5 e dezoito no dia 16, dos eventos classificados no dia 5 dois foram classificados incorretamente e quatro corretamente, já no dia 16 três foram classificados incorretos e quinze corretos.

A Figura 26 mostra, em outro formato a mesma informação da Tabela 27, a contabilidade das classificações corretas (T) e não corretas (F) de três pacientes para seis classificadores, representados pelos nomes das colunas da Tabela 27. Aqui, deve-se observar que, de um lado, as classificações não corretas acontecem nos primeiros eventos e, de outro lado, quando o número de eventos é apenas um ou dois as classificações não corretas permanecem. Por outro lado, quando o número de evento é maior que dois ou três, as classificações não corretas passam, na sua grande maioria, para corretas.

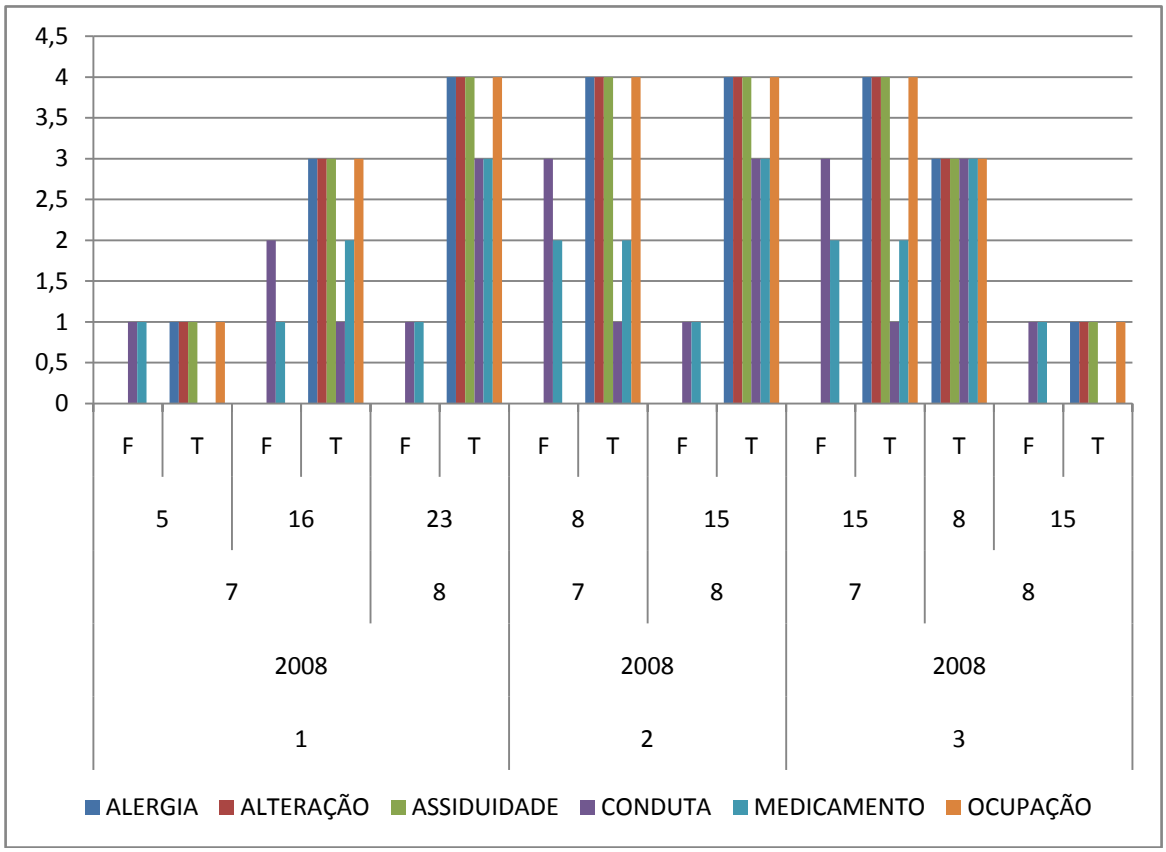

Figura 26. Contabilidade de classificações corretas e não corretas no tempo.

Finalmente, a presente pesquisa propôs a extração de modelos/classificadores confiáveis e compreensíveis que evidenciem a evolução antecipada de um paciente que se submete a uma ou mais cirurgias estética. A aplicação de técnicas de descoberta de conhecimento, sobretudo métodos de aprendizagem de máquina simbólica, gera informações úteis à medida que o histórico de um paciente é construído. As regras descobertas permitiram a construção de um módulo de software que fornece informações/classificações ao profissional de fisioterapia, de forma interativa, já a partir do primeiro evento registrado sobre um determinado paciente; a taxa de acerto dos classificadores melhora a partir que o histórico é incrementado.

Em termos do domínio de aplicação, os conhecimentos descobertos ajudam realizar a definição do plano de tratamento fisioterapêutico de um paciente que se submeteu a uma ou várias cirurgias plásticas. O ganho esperado concerne às sugestões que o sistema fornece por meio dos modelos e, desta forma, podem-se antecipar procedimentos e cuidados, bem como a alocação de recursos. Por exemplo, já no registro da informação sobre a anamnese, os modelos podem indicar que a conduta para aquele paciente será *drenagem linfática manual de mama* e que o número de procedimentos será superior a sete.

4.1 DISCUSSÃO

Os registros em saúde têm muito a oferecer no que diz respeito à informação, conhecimento e gerenciamento de dados. Porém, na área da fisioterapia há muito que fazer; são poucos os registros de estudos relacionados à fisioterapia. Trabalhos relacionados à aquisição de conhecimento em saúde já foram produzidos junto ao Programa de Mestrado em Tecnologia em Saúde,como identificação de características relacionadas à mortalidade infantil utilizando a descoberta de conhecimento em base de dados da saúde pública (VIANNA, 2007), e alguns trabalhos na área de fisioterapia, porém nenhum que utilizasse os métodos de classificação C4.5, BAGGING e BOOSTING, muitos destes trabalhos já realizados são da área de sistemas de apoio a decisão com Sistema Especialista, como exemplo, as pesquisas: sistema de apoio a decisão para o desmame da ventilação mecânica (ZANON, 2008); sistema de apoio à decisão na monitoração do paciente em assistência ventilatória invasiva (RIELLA, 2007); sistema de apoio à decisão em exames ortopédicos da coluna vertebral para auxílio nos diagnósticos fisioterapêuticos da região cervical e lombar (PEIXOTO, 2006); aprendizagem de máquina baseada na combinação de classificadores em base de dados da área da saúde (LOPES, 2007) entre outros. Estes trabalhos são

unânimes no que diz respeito à aquisição, registro e manutenção dos dados, da importância da construção de uma base de dados que assegure à credibilidade dos dados, produtividade e possibilidade de transformar dados em informações.

A construção de um sistema de informação para registro e armazenamento dos dados, foi uma das etapas importante deste trabalho, e que possibilitou a transformação dos dados em informação. Tal informação pode ser vista na forma de relatórios e gráficos que facilitaram a visualização dos resultados encontrados.

A estruturação multidimensional dos dados/assuntos permitiu que ferramentas de OLAP sejam utilizadas na recuperação, organização, investigação e resumo dos dados. As informações obtidas por meio destas ferramentas foram capazes de responder de forma dinâmica perguntas triviais dentro de uma clínica de cirurgia plástica, como: quais são as idades dos pacientes que mais operaram nos últimos três anos? A escolha do d/w como sistema de armazenamento dos dados ocorreu visto que ele facilita a aplicação de técnicas de descoberta de conhecimento.

A transformação da informação em conhecimento foi realizada por meio da classificação simbólica por meio dos classificadores simples (e.g. C4.5) e classificadores compostos (e.g. BAGGING e BOOSTING). Estes modelos de classificação foram escolhidos por representarem seus resultados por meios simbólicos de fácil compreensão, como as regras de produção.

Trabalhos semelhantes a esta dissertação, que também utilizaram a comparação de métodos como C4.5, BAGGING e BOOSTING como modelos de classificação, já foram desenvolvidos em diversas áreas. Pode-se citar Costa, et al.(2005) que utilizou a comparação de técnicas individuais de aprendizagem de máquina (árvores de decisão, naives bayes, redes neurais...) com modelos multiclassificadores (BAGGING, BOOSTING, STACKINGC...). Estes últimos apresentaram desempenho superior, com destaque para a combinação de árvore de decisão com BOOSTING; Quinlan (1996), também, comprovou a eficiência do método BOOSTING. Outros trabalhos de destaque são trabalhos desenvolvidos em nossa instituição como Lopes (2007) e Borges (2009), que da mesma maneira deste estudo, comprovou uma melhor classificação com menor taxa de erro pelo método BOOSTING.

BOOSTING foi o método que melhor classificou os eventos referentes a este trabalho. É importante salientar também que para todos os três métodos aplicados foi observado que as classificações incorretas foram diminuindo conforme o número de eventos foi aumentando. Isto

mostra que a interatividade do usuário com o sistema, enriquece o prontuário do paciente e proporciona uma melhor chance de acerto das recomendações sobre o mesmo paciente.

Finalmente, a existência de um modelo de dados e de um sistema operacional, bem como de um modelo de dados multidimensional organizado por assuntos relevantes e de um sistema informacional (d/w), permite-nos vislumbrar os seguintes benefícios para uma clínica de cirurgia estética:

- um espaço compartilhado entre profissionais para o registro e acesso de dados clínicos;
- um espaço compartilhado de termos para descrever minimante os eventos envolvendo os pacientes;
- a geração facilitada de informações a partir de uma base de dados integrada e consistida;
- a extração de padrões de comportamento dos dados;
- a aplicação de forma facilitada destes padrões no dia-a-dia da operação da clínica com vista a um atendimento de maior qualidade para o paciente.

5 CONCLUSÕES E TRABALHOS FUTUROS

O trabalho proposto visa descobrir padrões a partir de dados coletados por meio de diferentes fontes que possam auxiliar as necessidades dos diversos profissionais que atuam em uma clínica de cirurgia plástica. Tais padrões devem, em geral, ajudar no planejamento e na execução de um bom tratamento fisioterapêutico para cada paciente, como também no gerenciamento e na rotina profissional de uma clínica de estética, lembrando que para tal foram analisados e processados os dados utilizados por diferentes profissionais como: enfermeiras, médicos, fisioterapeutas, instrumentadoras, secretarias e administradora. A consecução deste trabalho incluiu diferentes tarefas não triviais, a saber: (i) um estudo detalhado sobre os dados coletados provenientes de diferentes bases especialistas atuantes na clínica; (ii) a execução de inúmeros experimentos visando extrair conhecimentos úteis para o processo de planejamento e execução de um bom tratamento fisioterapêutico; e (iii) uma análise criteriosa para validar as regras, a qual foi feita por meio da comparação das taxas de acerto dos classificadores.

Para obter bons modelos foi realizado um estudo aprofundado sobre os algoritmos e métodos mais comuns de aprendizagem de máquina, mas que produzissem bons resultados. Nestes termos, optou-se pelo clássico algoritmo C4.5. Uma das particularidades interessantes do C4.5 é que ele gera modelos simbólicos de fácil leitura. Além dos modelos simbólicos, buscou-se também obter modelos com boa taxa de acerto. Para tal, se utilizou dos métodos de combinação de modelos/classificadores: BAGGING e BOOSTING. Como era esperado, em termos teóricos, o método BOOSTING produziu resultados melhores que os demais. Destacando-se também os métodos que foram aplicados: (i) sobre conjuntos de treinamentos que continham todos os atributos da base de dados simulada total; e (ii) sobre conjuntos de treinamentos reduzidos por processo de filtragem de atributos, que continham em média 25% dos atributos da base de dados simulada total. Considerando tais dimensionalidades dos conjuntos de treinamentos, notou-se que as taxas de acerto dos classificadores não se alteram significativamente. Isto sugere, quando necessário, que se proceda à redução da dimensão da base de dados.

Era de fundamental importância que houvesse o fechamento de ciclo, o mais completo possível, envolvendo os sistemas computacionais de uma clínica, a saber: operacional, informacional e apoio a decisão. Ou seja, um sistema que realizasse o registro dos eventos do dia a dia de uma clínica de estética, um sistema que estruturasse os dados em modelo multidimensional e finalmente, um sistema que usasse em particular os dados no esquema multidimensional para apoiar tarefas de descobertas de conhecimentos e operacionalizasse a aplicação dos achados. Deste conjunto de sistemas, é importante destacar:

- a forma de interação do usuário com o módulo MIR, onde o mesmo fornece ao usuário informações, obtidas a partir da aplicação das regras dos diferentes modelos, de forma dinâmica à medida que se constrói o histórico do paciente com a clínica;
- a forma dinâmica de montar-se relatório, por meio de ferramentas de OLAP, a partir da base de dados construída sobre um esquema multidimensional, classicamente denominada de *data warehouse* ou simplesmente *d/w*.

O uso dos modelos, por meio do módulo MIR, pode ser feito por pessoas com objetivos diferentes como, por exemplo:

- um profissional, pode usá-lo como uma memória auxiliar no planejamento e acompanhamento de um paciente à medida que o histórico do mesmo se constrói;
- um novato, pode usá-lo como um simulador ou uma extensão de sua memória para avaliar casos e começar a formar suas primeiras experiências;
- um profissional, como uma forma de geração de diferentes alertas.

Finalmente, pode-se resumir dizendo que o trabalho realizado nesta dissertação, mesmo com dados simulados, apresenta uma base, em termos de conceitos, tecnologia e uso dos mesmos, para fazer evoluir os sistemas de informação e de apoio a decisão na área em questão.

5.1 TRABALHOS FUTUROS

O sistema proposto nesta dissertação apresentou resultados satisfatórios no que diz respeito ao auxílio ao profissional fisioterapeuta à tomada de decisão. Porém, estes resultados são baseados em uma base de dados simulada, ficando como desafio para trabalhos futuros a aplicação do

sistema em bases de dados reais e a validação dos conhecimentos descobertos por profissionais capacitados da área.

Ainda podem ser realizados estudos referentes à necessidade de substituição do modelo que possa estar obsoleto por um novo que deve ser gerado e incorporado em seu lugar. Para tal, há diferentes técnicas que podem ser empregadas com diferentes níveis de complexidade. A mais simples consiste em estabelecer que ao final de um determinado período de tempo, por exemplo, um mês todos os modelos são re-gerados, incluindo os dados deste período no conjunto de treinamento e teste. Uma forma mais sofisticada consistiria em incorporar um módulo de detecção de mudanças. Tal módulo decidiria se um ou mais modelos devem ser substituídos afim do sistema recuperar ou melhorar a sua taxa de acerto.

Outra evolução ao sistema seria incorporar novos dados de pesquisa que englobem de forma mais generalizada uma clínica de estética, como: os custos de uma cirurgia, logística e encaminhamento para hospitais com melhor custo/benefício, retorno ou não de um paciente para uma nova cirurgia, identificação dos pacientes que mais indicam a clínica, marketing e outros.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BAUER, ERIC e KOHAVI, RON. 1999. *An Empirical Comparison of Voting Classification Algorithms: BAGGING, BOOSTING, and Variants*. Machine Learning, 36(1/2). 1999, pp. 105-139.

BARLETT, P.; FREUND, Y.; LEE, W.S.; SCHAPIRE, R.E.:, **1997.** *Boosting the margin: A new explanation for the effectiveness of voting methods*. Proc 14th International Conference on Machine Learning**.** Morgan Kaufmann, San Francisco (1997) 320-330.

BOLFARINE, H. BUSSAB, W. O. 2005. *Elementos de Amostragem*. Editora Edgar Blücher, 290p, 2002.

BORGES, ANDRÉ PINZ. 2009. *Descoberta de Regras de Condução de Trens de Carga.* Curitiba : s.n., 2009. Dissertação – Mestrado – Informática Aplicada – Pontifícia Universidade Católica do Paraná – Centro de Ciências Exatas e de Tecnologia.

BREIMAN, L. 1996. *BAGGING predictors*. Machine learning, 24(2). 1996, pp. 123-140.

CANESTRARO J., DIAS J., MALUCELLI A., SHMEIL M. *Sistema de Apoio à Decisão para Infarto Agudo do Miocárdio.* Anais do X CBIS, Florianópolis-SC, 2006.

CHEN, L., WRIGHT, P. e NEJDL, W. 2009. *Improving music genre classification using collaborative tagging data*. Proceedings of the Second ACM International Conference on Web Search and Data Mining. 2009, pp. 84-93.

CIOS, K. J., et al. 2007. *Data Mining: A Knowledge Discovery Approach.* s.l. : Springer, 2007. p. 606. ISBN: 978-0-387-33333-5.

COSTA, J. A. ; BITTENCOURT,V. G. e SOUTO, M. C. P. 2005. *Aplicação de Multi-classificadores no Reconhecimento de Classes Estruturais de Proteínas*. Press XXVIII CNMAC. 2005 snato Amaro – S.P.

DIAMANTIDIS, N.A., KARLIS, D. e GIANKOUMAKIS, E.A. 2000. *Unsupervised stratification of cross-validation for accuracy estimation.* Artificial Intelligence, 116. 2000, pp. 1-16.

FAYYAD, U., PIATETSKI-SHAPIRO, G. e PADHRAIC, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM.* 1996, pp. 27-34.

FAYYAD, U. M. 1996. Data mining and knowledge discovery: making sense out of data. *IEEE Expert.* 5, 1996, Vol. 11, pp. 20-25.

FREUND, Y. e SCHAPIRE, R.E. 1996. Experiments with a new BOOSTING algorithm. *Proceedings of the 13th International Conference on Machine Learning.* 1996, pp. 148-156.

FRONZA C.F., OSÓRIO F.C.A. 2006. *O Prontuário Eletrônico como Instrumento de Avaliação e Apoio à Decisão Clínica Fisioterapêutica.* Anais do X-CBIS, Florianópolis-SC, 2006.

GRANDVALET, Y. Bagging Equalizes Influence**.** *Machine Learning*, 55(3): 251-270, 2004.

GRAY,R.M. Entropy and information theory. *Springer Verlag*, New York, 1990.

GROSSMAN M, ARONSON JE, McCARTHY RV. 2005. *Does UML make the grade? Insights from the software development community.* Information and Software Technology, 2005; 383-397.

GUIRRO E,GUIRRO R. Fisioterapia Dermato-funcional, 3a. ed, São Paulo: Manole; 2004.

HALL, M. A. 2000. Correlation-based feature selection for discrete and numeric class machine learning. *Proc. of the 17th Int. Conf. on Machine Learning.* 2000, pp. 359-366.

—. 1998. *Correlation-based Feature Selection for Machine Learning.* Hamilton : Department of Computer Science, University of Waikato, 1998. Ph.D. thesis.

HALL, M. A. e HOLMES, G. 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering.* 2003, Vol. 15, 3.

HAN, JIAWEI e KAMBER, MICHELINE. 2006. *Data Mining: Concepts and Techniques.* Second Edition. San Franciso, CA : Morgan Kaufmann, 2006. p. 772.

HUAN, LIU e MOTODA, HIROSHI. 1998. *Feature Selection for Knowledge Discovery and Data Mining.* s.l. : Kluwer Academic Publishers, 1998.

HUSSAIN, FARHAD, et al. 1999. Discretization: An Enabling Technique. *The National University of Singapure. Technical Report.* Junho de 1999, pp. 1-29.

INMON, W. H. 2002. *Building the Data Warehouse.* 3 ed. Toronto : Willey Computer Publishing, 2002.

KALOS, A. e REY, T. 2005. *Data mining in the chemical industry.* Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005, pp. 763-769.

KIMBALL, R.e MERS, R. 2000. *Data Webhouse: Construindo um Data Warehouse para a Web*. Rio de Janeiro: Campus.

KOBUS, L. ENEMBRECK, F. SCALABRIN, E. DIAS, J. HONORATO, S. *Automatic Knowledge Discovery and Case Management: an effective way to use databases to enhance health care management.* In: 5th IFIP Conference on Artificial Intelligence Applications & Innovations, 2009, Thessaloniki. 5th IFIP Conference on Artificial Intelligence Applications & Innovations, 2009. v. 1.

KOHAVI, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence (IJCAI) p. 1137-1145.

KOHAVI, R. e JOHN, G. 1996. Wrappers for Feature Subset Selection. *AIK special issue on relevance.* 1996.

KOLLER, D. e SAHAMI, M. 1996. Toward optimal feature selection. *Proc. of the 13th Int. Conf. on Machine Learning.* 1996, pp. 284-292.

LEE, H. D., MONARD, M. C. e BARANAUSKAS, J. A. 1999. *Empirical comparison of wrapper and filter approaches for feature subset selection.* São Carlos : ICMC-USP, 1999. Disponível em: ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_094.pdf.

LIU, H.; MOTODA, H. 1998. Feature selection for knowledge discovery and data mining. *Kluwer*, 1998.

LIU, HUAN e YU, LEI. 2005. Toward Integration Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering.* 2005, pp. 491-502.

LOPES, LUCELENE. 2007. *Aprendizagem de máquina baseada na combinação de classificadores em bases de dados da área de saúde.* Curitiba : s.n., 2007. Dissertação – Mestrado – Tecnologia em Saúde – Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde.

MALOOF, MARCUS A. e MICHALSKI, RYSZARD S. 2000. *Selecting Examples for Partial Memory Learning.* s.l. : Machine Learning Journal, 2000. pp. 27-52. Vol. 41, citeseer.ist.psu.edu/maloof00selecting.html.

MITCHELL, T. 1997. *Machine Learning.* New York : McGraw-Hill, 1997.

NILSSON, N.J. 1996. *Introduction to machine learning.* Stanford : Stanford university, 1996.

OCKHAM, W. 1999. *Prólogo da Exposição dos Oitos Livros da Física.* São Paulo : Nova Cultural, 1999.

PEIXOTO, L. 2006. *Sistema de apoio à decisão em exames ortopédicos da coluna vertebral para auxílio nos diagnósticos fisioterapêuticos da região cervical e lombar* , 2006. Dissertação – Mestrado – Tecnologia em Saúde – Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde.

PRATI, R. C. 2006. Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos. São Carlos : s.n., 2006. p. 191. Tese de doutorado.

QUINLAN, J. R. 1993. *C4.5: Programs for machine learning.* San Francisco : Morgan Kaufman, 1993.

QUINLAN, J. R. 1986. *Induction of decision trees.* Machine Learning, 1(1): 81-106, 1986.

QUINLAN, J. R. 1987. *Generation Production Rules from Decision Trees.* s.l. : In Proc. of IJCAI 87, 1987. pp. 304-307.

QUINLAN, J. R. 1996. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research.* 1996, Vol. IV, pp. 77-90.

QUINLAN, L. R. 1996. Bagging, Boosting and C4.5. *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference*, AAAI 96, IAAI 96, August 4-8, 1996, Portland, Oregon - Volume 1. AAAI Press / The MIT Press, 1996.

RODDICK, J., FULE, P. e GRACO, W. 2003. Exploratory medical knowledge discovery: experiences and issues. *ACM SIGKDD Explorations Newsletter.* 1, 2003, Vol. 5, pp. 94-99.

ROMÃO, W. 2002. Descoberta de Conhecimento Relevante em Banco de Dados sobre Ciência e Tecnologia; Tese – Doutorado – Engenharia da Produção – Universidade Federal de Santa Catarina, 2002.

RIELLA, C. 2007. *Sistema de apoio à decisão na monitoração do paciente em assistência Ventilatória Invasiva,* 2007; Dissertação – Mestrado – Tecnologia em Saúde – Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde.

SALVADOR, V. F. M. BRITTO, M.; MOURA JR, L. A. e ALMEIDA JR, J. R. 2006. Qualidade de Dados para Gestão do Conhecimento na Área de Saúde. X Congresso Brasileiro de Informática em Saúde. Florianópolis- SC., 2006. http://www.sbis.org.br/cbis/arquivos/758.pdf. Acessado: 20/02/2009.

SCHAPIRE, R.E.F.; LEE, W.S. 1997. *Boosting the margin: A new explanation for the effectiveness of voting methods.* In: Fischer DH [editor]. Proc Fourteenth International Conference on Machine Learning, San Francisco: Morgan Kaufmann, 332-340, 1997.

SEBBAN, M.; NOCK,R. e LALLICH, S., *Boosting Neighborhood- Based Classififers.* Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco: Morgan Kaufmann, 505-512.

SIGULEM D., ANÇÃO, M. S., RAMOS, M. P., LEÃO, B. F. "Sistemas de apio à Medicina". Extraído da versão eletrônica do livro "Atualização terapêutica – Manual Prático de Diagnóstico e Tratamento", 1998.

SILVA, DB. A Fisioterapia Dermato-Funcional como Potencializadora no Pré e Pós-Operatório de Cirurgia Plástica. Disponível em: http:infonet.com.br/fisioterapia/materia20.htm. Acessado em 07/03/2006.

TAN, M., STEINBACH, V. e KUMAR, A.W. 2006. *Introduction to Data Mining.* Minnesota : Addison Wesley, 2006.

THOMSEN, E. 2002. OLAP: Construindo Sistemas de Informações Multidimensionais. 2ª edição, Rio de Janeiro: Campus, 2002.

VIANNA, R. C. X. F. 2007. *Identificação de Características Relacionadas à Mortalidade Infantil Utilizando a Descoberta de Conhecimento em Base de Dados de Saúde Pública*. 2007. Dissertação – mestrado – Tecnologia em Saúde **-** Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde.

VREEMAN,D.;TAGGARD, S.; RHINE, M. D. e WORREL,T. W. 2006. Evidence for Electronic Health Record Systems in Physical Therapy. Physical Therapy, vol.86, num.3, march 2006.

WEKA. 2008. Data Mining with Open Source Machine Learning Software in JAVA. *WEKA.* [Online] 2008. http://www.cs.waikato.ac.nz/ml/weka/ Acesso em: 15/01/2008.

WITTEN, I.H e FRANK, E. 2005. *Data Mining: Practical machine learning tools and techniques.* 2 ed. San Francisco : Morgan Kaufmann, 2005.

WITTEN, I.H; FRANK, E. 2000. Data mining: practical machine learning tools and techniques with Java implementations, *Morgan Kaufmann*, 2000.

ZANON, B. 2008. *Sistema de Apoio a Decisão para o Desmame da Ventilação Mecânica*. Dissertação – Mestrado – Tecnologia em Saúde – Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde.

7 ANEXO I – GERAÇÃO DE DADOS SIMULADOS

Nas próximas páginas constam os detalhes dos dados para a geração dos exemplos de treinamento que foram usados nesta dissertação. De forma pragmática, tais foram descritos por meio das seguintes tabelas:

TABELA A1. Descrição de dados para a geração de exemplos de “paciente”.
TABELA A2. Descrição de dados para a geração de exemplos de “cirurgia”.
TABELA A3. Descrição de dados para a geração de exemplos de “alteração”.
TABELA A4. Descrição dos dados para a geração de exemplos “queixa principal”
TABELA R. Estrutura da tabela que conterá os exemplos gerados a partir das tabelas A1, A2, A3 e A4.

A geração propriamente dita é feita por meio da invocação da função geraExemplo descrita na próxima página.

```
Inteiro k = geraExemplo(A1, ref R, 16, 1);   {geração de exemplos para anamnese}
k = geraExemplo(A2, ref R, 16, k);           {geração de exemplos para cirurgia}
k = geraExemplo(A3, ref R, 16, k);           {geração de exemplos para alteração}
k = geraExemplo(A4, ref R, 16, k);           {geração de exemplos para queixa principal}
```

ou simplesmente

```
k = geraExemplo(A4, ref R, 16,
                    geraExemplo(A3, ref R, 16,
                                geraExemplo(A2, ref R, 16,
                                            geraExemplo(A1, ref R, 16, 1))));
```

O número total de exemplos gerados foi 4826. Tal informação é fornecida pela variável k.

Funções para a a geração aleatoria de exemplos

```
função  geraExemplo(A, var R : TABELA, α, k : inteiro ) : inteiro
{A é tabela que descreve os dados para a geração de exemplos.}
{R é referência da tabela que conterá os eventos/exemplos gerados a partir de A}
{α é número máximo de exemplos que serão  gerados aleatoriamente para cada valor da coluna 1 da Tabela A.}
{k é o apontador para a próxima linha vazia da Tabela R, onde será inserido o próximo exemplo gerado}
inicio
    inteiro: i;       {linha corrente da Tabela A}
    inteiro: n;       {número de linhas da Tabela A}
    inteiro: j ;      {coluna/atributo corrente da Tabela A}
    inteiro: m ;      {número de colunas da Tabela A}
    String: X ;       {valores possíveis para o atributo j correspondente ao atributo 1 da linha i}
    inteiro: h;       {variável auxiliar}
    para cada i de 1 até n passo 1 faça
    inicio
            h = randon(α );          {retorna um número inteiro aleatório ≤ u}
            para cada q de 1 até h passo 1 faça
            inicio
                    R[ k ][ 1 ] = A[i][1];
                    para cada j de 2 até m passo 1 faça
                    inicio
                            R[k][j] = pegaValor(A[i][j]);
                            k = k + 1;
                    fimpara
            fimpara
    fimpara
    retorne k;      {o valor de k indica o número de exemplos +1 constante na Tabela R}
fim.
```

```
Função pegaValor(Vetor<String> V) : String
{V é um vetor de cadeias de caracteres (String)}
{retorna um dos elementos de V selecionados aleatoriamente}
inicio
      inteiro: u = V.lenth();      {retorna o número de elementos de V}
      inteiro: h = randon( u );    {retorna um número aleatório ≤ u}
      retorne V[ h ];
fim.
```

TABELA A1. Descrição de dados para a geração de exemplos de “paciente”.

OCUPAÇÃO	TRIMESTRE	HAB_ALIMENTARES	ATIVIDADEFISICA	TABAGISMO	ANTE_PATOLOGICOS	ALERGIAS	MEDICAMENTOS	IDADE
ADMINISTRADORA	1;2;1	NORMAL	NAO	NAO	AMIGDALA	NAO	ANTICONCEPCIONAL	26;30;1
ADVOGADA	1;2;1	NORMAL	ESTEIRA; MUSCULACAO; NAO	SIM; NAO	GASTROPLASTIA; CESARIA; REDUCAO-DE-MAMA;NAO	PENICILINA; NAO	ANTIHIPERTENSIVO;ANTICONCEPCIONAL;NAO	27;39;1
ARQUITETA	2;3;1	NORMAL	NAO	NAO	NAO	NAO	ANTICONCEPCIONAL	24;30;1
ARTISTA	2;3;1	NORMAL	NAO	SIM	GESTACOES	NAO	NAO	27;37;1
ASSISTENTE-CONTABIL	2;3;1	NORMAL	MUSCULACAO	NAO	NAO	NAO	ANTIDEPRESSIVO	27;37;1
ATLETA	3;4;1	REEDUCACAO-ALIMENTAR	CORRIDA	NAO	CESARIA	NAO	NAO	27;37;1
BANCARIA	3;4;1	NORMAL;REEDUCACAO-ALIMENTAR	DANCA;NAO	SIM;NAO	LIPOASPIRACAO;NAO	METIOLATE;NAO	FENOPREPOREX-SUBTRAMINA;ANTIDEPRESSIVO	26;51;2
CABELEIREIRA	3;4;1	NORMAL	NAO	NAO	CESARIA;NAO	DIPIRONA; NAO	NAO	27;37;1
COMERCIANTE	2;3;1	NORMAL;REEDUCACAO-ALIMENTAR	CAMINHADA;MUSCULACAO; NAO	NAO	CESARIA;HIPERTENSAO;VARIZES;APENDICE;CESARIA;GASTROPLASTIA;HISTERECTOMIA;NAO	PO;NAO;INSETO	ANTIHIPERTENSIVO;ANTIDEPRESSIVO;HOMEOPATIA;ANTICONCEPCIONAL;FLORAIS;NAO	33;55;2
DESIGNER	2;3;1	NORMAL	NAO	SIM;NAO	CESARIA;NAO	PO;NAO	NAO	22;35;2
DO-LAR	3;4;1	NORMAL;REEDUCACAO-ALIMENTAR	CAMINHADA;HIDROGINASTICA;MUSCULACAO;NAO	SIM;NAO	ABDOMENOPLASTIA;CESARIA;GASTROPLASTIA;GASTROPLASTIA;HERNIA-HIATO;HIDROCEFALIA;HISTERECTOMIA;PARTO-NORMAL;PROTESE-MAMARIA;RETENCAO-DE-LIQUIDO;VARIZES;NAOLAPAROTOMIA;LIPOASPIRACAO;MAMOPLASTIA;MENTOPLASTIA;	PENICILINA;LACTOSE; PERFUME;PO;NAO	NAO;FENOPREPOREX-SUBTRAMINA;ANTIHIPERTENSIVO;SINVASTATINA;ANTIDEPRESSIVO;ANTICONCEPCIONAL;EUTIROX-GLIFAGE;DIURETICO;REPOSICAO-HORMONAL	26;63;3

OCUPAÇÃO	TRIMESTRE	HAB_ALIMENTARES	ATIVIDADEFISICA	TABAGISMO	ANTE_PATOLOGICOS	ALERGIAS	MEDICAMENTOS	IDADE
EMPRESARIA	3;4;1	NORMAL	CORRIDA;ESTEIRA;MUSCULACAO;NAO	SIM;NAO	NAO;CESARIA;COLECISTECTOMIA;LIPOASPIRACAO;RETIROU-AMIGDALA-E-ADENOIDE;GLAUCOMA;PROTESE-MAMARIA	NAO;MICROPORE;CORTICOIDE;INSETO	ANTICONCEPCIONAL;ANTIDEPRESSIVO;RAUCUTAM	25;48;2
ENFERMEIRA	1;2;1	NORMAL	NAO	NAO	CESARIA	NAO	NAO	27;37;1
ESTUDANTE	3;4;1	REEDUCACAO-ALIMENTAR;NORMAL	NAO;BICICLETA;CAMINHADA;MUSCULACAO	SIM;NAO	RETIROU-AMIGDALA-E-ADENOIDE;OTOPLASTIA;RINOPLASTIA;GASTROPLASTIA;NAO;ADENOIDE;ADERENCIA-INTESTINAL;ASMA	PO;NAO;AAS;DIPIRONA	FENOPREPOREX-SUBTRAMINA;ANTICONCEPCIONAL;PREDINIZONA;ANTIDEPRESSIVO;NAO	15;21;2
FISIOTERAPEUTA	1;2;1	NORMAL	NAO	NAO	NAO	NAO	ANTICONCEPCIONAL	27;37;1
FUNCIONARIA-PUBLICA	3;4;1	NORMAL	NAO	NAO	HERNIA-INGNAL;RINOPLASTIA;NAO	DIPIRONA;NAO	ANTICONCEPCIONAL;NAO	22;33;2
LOJISTA	1;2;1	NORMAL	NAO	NAO	RINOPLASTIA	NAO	NAO	27;37;1
MEDICA	3;4;1	REEDUCACAO-ALIMENTAR	CORRIDA;NAO	NAO	PROTESE-MAMARIA;HISTERECTOMIA	NAO	NAO;ANTIDEPRESSIVO	40;58;2
MOTO-GIRL	3;4;1	NORMAL	NAO	SIM	NAO	NAO	ANTICONCEPCIONAL	27;37;1
POLICIAL-MILITAR	3;4;1	REEDUCACAO-ALIMENTAR	NAO	NAO	RETIROU-TROMPAS	LATEX	NAO	27;37;1
PROFESSORA	3;4;1	REEDUCACAO-ALIMENTAR;NORMAL	HIDROGINASTICA;NAO	SIM;NAO	CESARIA;GESTACOES;CIRURGIA-DE-COLUNA;PARTO-NORMAL;LIPOASPIRACAO;NAO	LACTOSE;MICROPORE;DIPIRONA;NAO	ANTIHIPERTENSIVO;ANTICONCEPCIONAL;NAO	21;50;3
PROMOTORA	1;2;1	NORMAL	MUSCULACAO;NAO	NAO	PROTESE-MAMARIA;CESARIA	NAO	ANTICONCEPCIONAL;NAO	26;41;2
PSICOLOGA	3;4;1	NORMAL	PILATES	EX	HISTERECTOMIA	PO	ANTIHIPERTENSIVO	37;47;2
PUBLICITARIA	3;4;1	NORMAL	CAMINHADA;NAO	NAO	NAO	NAO	ANTIHIPERTENSIVO;ANTICONCEPCIONAL	26;52;3
SECRETARIA	3;4;1	NORMAL;REEDUCACAO-ALIMENTAR	NAO	SIM;NAO	GASTROPLASTIA;REDUCAO-DE-MAMA	NAO	ANTIDEPRESSIVO;ANTICONCEPCIONAL;NAO	28;53;3
SERVENTE	1;2;1	NORMAL;REEDUCACAO-ALIMENTAR	CAMINHADA	NAO	RETIRORU-OVARIO;BLEFAROPLASTIA	METAL;NAO	ANTIDEPRESSIVO;ORMIGREN	36;50;3
VENDEDORA	3;4;1	NORMAL	MUSCULACAO;NAO	SIM;NAO	GASTROPLASTIA;CESARIA;VARIZES;NAO	NAO	VITAMINAS;ORMIGREN;ANTICONCEPCIONAL;NAO	20;42;3

TABELA A2. Descrição de dados para a geração de exemplos de “cirurgia”.

CIRURGIA	TECNICA	LOCALIZAÇÃO	VOLUME	CONDUTA
ABDOMENOPLASTIA	ANCORA; CLASSICA	ABDOMEN	1600;2620;150	DLM-DE-ABDOMEN;DLM-REVERSA-DE-ABDOMEN
BLEFAROPLASTIA	CLASSICA	PALPEBRA	0;0;0	DLM-DE-FACE
CORRECAO-DE-CICATRIZ	CLASSICA	CICATRIZ-UMBILICAL	0;0;0	DLM-DE-COSTAS;DLM-DE-MEMBROS-INFERIORES
DORSOPLASTIA	CLASSICA	COSTAS-DORSO	500;700;50	DLM-DE-COSTAS
INTIMA-PEQ-LABIOS	CLASSICA	PEQUENOS-LABIOS	0;0;0	DLM-DE-MAMA
LIFITING-DE-BRACO	CLASSICA	BRACO-D-E	100;237;50	DLM-DE-MAMA
LIPOASPIRACAO	ILLOUZ	ABDOMEN	300;2500;300	DLM-DE-COSTAS;DLM-DE-ABDOMEN;ULTRA-SOM;DLM-REVERSA-DE-ABDOMEN
LIPOASPIRACAO-1	ILLOUZ-1	BRACO DIREITO;BRACO ESQUERDO	100;200;30	DLM-DE-MEMBROS-SUPERIORES
LIPOASPIRACAO-2	ILLOUZ-2	COSTAS-DORSO	1900;2600;200	DLM-DE-COSTAS;DLM-DE-MAMA;ULTRA-SOM
LIPOASPIRACAO-3	ILLOUZ-3	MEMBROS-INFERIORES	300;800;100	DLM-DE-MEMBROS-INFERIORES
LIPOASPIRACAO-4	ILLOUZ-4	CULOTE	0;0;0	DLM-DE-MEMBROS-INFERIORES
LIPOASPIRACAO-5	ILLOUZ-5	FLANCO DIREITO;FLANCO ESQUERDO	250;2600;300	DLM-DE-MEMBROS-INFERIORES;DLM-DE-COSTAS
LIPOASPIRACAO-6	ILLOUZ-6	JOELHO DIREITO;JOELHO ESQUERDO	300;700;100	DLM-DE-MEMBROS-INFERIORES
MAMOPEXIA	T-INVERTIDO	MAMA DIREITA;MAMA ESQUERDA	0;0;0	DLM-DE-MAMA
MAMOPLASTIA-REDUTORA	T-INVERTIDO	MAMA DIREITA;MAMA ESQUERDA	100;500;70	ULTRA-SOM
OTOPLASTIA	PEREIAROLAR	ORELHA DIREITA;ORELHA ESQUERDA	0;0;0	DLM-DE-MEMBROS-INFERIORES
PROTESE-MAMARIA	T-INVERTIDO	MAMA DIREITA;MAMA ESQUERDA	100;500;100	DLM-DE-MAMA
PROTESE-MAMARIA-1	PERIAREOLAR	MAMA DIREITA;MAMA ESQUERDA	100;500;100	DLM-DE-MAMA
PROTESE-MAMARIA-2	SULCO-MAMARIO	MAMA DIREITA;MAMA ESQUERDA	100;500;100	DLM-DE-MAMA
RINOPLASTIA	CLASSICA	NARIZ	0;0;0	DLM-DE-FACE
RINOPLASTIA-1	L	NARIZ	0;0;0	DLM-DE-FACE
VARIZES	CROCHETAGEM	MEMBROS-INFERIORES	0;0;0	DLM-DE-MEMBROS-INFERIORES

TABELA A3. Descrição de dados para a geração de exemplos de “alteração”.

ALTERAÇÃO	LOCALIZAÇÃO	TAMANHO	COR
DEISCENCIA	ABDOMENOPLASTIA;MAMA-DIREITA;MAMA-ESQUERDA	PEQUENO;MEDIO;GRANDE	?
DESLOCAMENTO-PROTESE	MAMA-DIREITA;MAMA-ESQUERDA	?	?
FIBROSE	ABDOMENOPLASTIA;ABDOMEN;MAMA-DIREITA;MAMA-ESQUERDA	PEQUENO;MEDIO;GRANDE	?
FIBROSE-1	FLANCO-DIREITO;FLANCO-ESQUERDO;PALPEBRA-DIREITA;PALPEBRA-ESQUERDA	?	?
FIBROSE-2	PALPEBRA-ESQUERDA;PALPEBRA-DIREITA;PUBIS	?	?
HEMATOMA	ABDOMEN;BRACO-DIREITO;BRACO-ESQUERDO;COCIX;	PEQUENO;MEDIO;GRANDE	ROXO;VERDE;AMARELO
HEMATOMA-1	COSTAS-DORSO;COXA-INTERNA-DIREITA;COXA-INTERNA-ESQUERDA;	?	?
HEMATOMA-2	CUDE;CULOTE;FLANCO-DIREITO;FLANCO-ESQUERDO;JOELHO;	?	?
HEMATOMA-3	MANA-DIREITA;MAMA-ESQUERDA;NARIZ;PALPEBRA-DIREITA;PALPEBRA-ESQUERDA;PUBIS	?	?
SEROMA	ABDOMEN;COCIX;COSTAS-DORSO;MAMA-DIREITA;MAMA-EQUERDA	PEQUENO;MEDIO;GRANDE	?
SEM-ALTERACAO	?	?	?

TABELA A4. Descrição dos dados para a geração de exemplos “queixa principal”

LOCALIZACAO	TIPO	INTENSIDADE
ABDOMEN		
BRACO		
CICATRIZ-ABDOMENOPLASTIA		
COSTASDORSO		
COXA		
CULOTE		
FLANCO-D-E		
FLANCO-E	DOR-DE-ESTOMAGO	
MAMA-D	DIFICULDADE-RESPIRATORIA	
MAMA-D-E	INSONIA	LEVE
MAMA-E	SEM-QUEIXA	INTENSA
PALPEBRA	CONSTIPACAO-INTESTINAL	MODERADA
PUBIS	DOR	INTENSA

TABELA R. Estrutura da tabela que conterá os exemplos gerados a partir das tabelas A1, A2, A3 e A4.

CODIGO	CODPAC	SEQ	ALERGIA	ALTERACAO	ANTECEDENTES	ATIVIDADEFISICA	CIRURGIA	CONDUTA	VOLUME	COR

DOR	HABITOSALIMENTARE	INTENSIDADE	LOCALIZACAO	MEDICAMENTO	OCUPACAO	FAIXAETARIA	TRIMESTRE	PARECER

PERIMETRIA	QUEIXA	TABAGISMO	TAMANHO	TECNICACIRURGICA	QUANTIDADE	ANO	MES	DIA	EVENTO

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo