UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

UM ESTUDO DE CASO UTILIZANDO
O MÓDULO DE ANÁLISE DE REGRAS DO *Rule System*

ALAN KELLER GOMES
MARIA CAROLINA MONARD

Nº 158

RELATÓRIOS TÉCNICOS

São Carlos – SP
Fev/2002

Um Estudo de Caso Utilizando o Módulo de Análise de Regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$.*

Alan Keller Gomes
Maria Carolina Monard

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
e-mail: {akgomes, mcmonard}@icmc.sc.usp.br

Resumo

Encontra-se em desenvolvimento no Laboratório de Inteligência Computacional — LABIC — um projeto de grande porte denominado DISCOVER, que tem objetivo fornecer um ambiente integrado para apoiar as etapas do processo de descoberta de conhecimento.

No intuito de testar algumas idéias que poderão futuramente ser incorporadas no ambiente DISCOVER, foi proposto um sistema computacional protótipo, implementado na linguagem de programação lógica Prolog, denominado $\mathcal{R}_{ule}\mathcal{S}_{ystem}$.

Um dos principais módulos do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$, o Módulo de Análise de Regras (MAR), tem como objetivo efetuar a análise quantitativa e qualitativa de regras de conhecimento induzidas por algoritmos de Aprendizado de Máquina simbólico.

Neste trabalho é apresentado um estudo de caso utilizando o MAR para efetuar a análise automática de diferentes conjuntos de regras induzidas utilizando dois algoritmos de AM e um conjunto de dados do mundo real, relacionados ao processamento de sêmen diagnóstico.

Palavras-Chave: Aprendizado de Máquina, Análise de Regras, Prolog, Processamento de Sêmen Diagnóstico.

Março 2002

*Trabalho realizado com auxílio do CNPq, CAPES e FAPESP.

Sumário

Lista de Figuras

Lista de Tabelas

[1]Este documento foi elaborado com o LaTeX, sua bibliografia é mantida com o auxílio da ferramenta BibView (Prati et al., 1999) e gerada automaticamente pelo BibTeX

1 Introdução

Encontra-se em desenvolvimento no Laboratório de Inteligência Computacional —LABIC[1]— um projeto de grande porte denominado DISCOVER, inicialmente proposto por Baranauskas e Batista (Baranauskas and Batista, 2001). O projeto DISCOVER tem como objetivo fornecer um ambiente integrado para apoiar as etapas do processo de descoberta de conhecimento, oferecendo funcionalidades voltadas para Aprendizado de Máquina (AM) (Batista, 1997; Caulkins, 2000; Milaré, 2000; Martins, 2001; Pila, 2001), Mineração de Dados[2](MD) (Batista, 2000; Félix, 1998; Horst, 1999; Lee, 2000; Nagai, 2000; Pugliesi, 2001; Baranauskas, 2001) e Mineração de Textos[3](MT) (Imamura, 2001).

As funcionalidades voltadas para AM consideram, entre outros, um formato padrão para as regras induzidas por algoritmos de AM simbólico, denominado $\mathcal{PBM}$ (Prati et al., 2001b; Prati et al., 2001a) bem como um formato padrão para os exemplos utilizados (Batista, 2001).

No intuito de testar algumas idéias que poderão futuramente ser implementadas no ambiente DISCOVER, foi proposto em (Gomes, 2001) e (Bernardini, 2001) um sistema computacional, denominado $\mathcal{R}_{ule}\mathcal{S}_{ystem}$, que é composto de três módulos:

1. Módulo de Análise de Regras (MAR);
2. Módulo de Combinação e Explicação (MCE) (Bernardini and Monard, 2002a); e
3. Módulo Auxiliar (MA).

O Módulo de Análise de Regras tem como objetivo efetuar a análise quantitativa e qualitativa de regras induzidas por algoritmos de AM simbólico, implementando diversas medidas de regras propostas na literatura.

Os procedimentos implementados no Módulo de Análise de Regras são:

1. Cálculo de Informações de Avaliação de Regras - `evaluateAllSetOfRuleFrequency/1`.
2. Cálculo de Medidas de Avaliação de Regras - `calculateMeasuresOverAllSetOFRules/1`.
3. Cálculo do Grau de Surpresa de Pequenos Disjuntos - `smallDisjSurp/3`.
4. Cálculo do Grau de Surpresa dos Atributos Individuais Discretos da Regra - `indAttSurp/2`.

O estudo de caso a ser apresentado nesse trabalho tem por objetivo efetuar a análise automática de regras através da execução dos procedimentos do Módulo de Análise de Regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Essas regras compõem diferentes classificadores simbólicos que são induzidos a partir de um conjunto de dados denominado *Proc-a-gmg-d*. Esse conjunto de dados refere-se ao processamento de sêmen diagnóstico.

[1] http://labic.icmc.sc.usp.br
[2] Data Mining
[3] Text Mining

O trabalho está organizado da seguinte forma: na Seção 2 é sucintamente apresentado o Módulo de Análise de Regras. Na Seção 3 são feitos alguns comentários a respeito do processamento de sêmen diagnóstico. Na Seção 4 são apresentadas as etapas desse estudo de caso. Na Seção 5 são discutidos os resultados obtidos e, finalmente, na Seção 6, são apresentadas algumas considerações finais sobre o trabalho realizado.

2 Módulo de Análise de Regras

Algoritmos de Aprendizado de Máquina (AM) simbólicos são capazes de induzir um classificador **h** de tal forma que o conceito descrito por **h** é facilmente interpretável por seres humanos. Geralmente, esse classificador pode ser transformado em conjuntos de regras *if-then*, ou seja, regras do tipo *Corpo* $\rightarrow$ *Cabeça* ou *Body* $\rightarrow$ *Head*. Cada regra de **h** pode então ser resumidamente denotada como $B \rightarrow H$.

Pode-se medir um conjunto de regras **h** dado um conjunto de exemplos S. Neste caso, **h** consiste de um conjunto de regras com NR regras $R_u, u = 1, ..., NR$, ou seja, $\mathbf{h} = \{R_1, ..., R_{NR}\}$, o qual denominamos classificador simbólico. Além de medir a precisão de **h** como um todo, isto é, como um classificador tipo caixa preta, é possível avaliar separadamente cada uma das regras que constituem **h**.

Regras podem ser avaliadas com o objetivo de saber quais são aquelas melhor sustentadas pelos dados — aspecto qualitativo — ou ainda podem ser avaliadas com o intuito de selecionar aquelas que possam trazer algum conhecimento surpreendente ou inesperado para o usuário, ou seja, o objetivo é selecionar as regras mais interessantes — aspecto quantitativo.

O Módulo de Análise de Regras (MAR), descrito em detalhes em (Gomes and Monard, 2002), tem como objetivo efetuar, de maneira automática, a análise qualitativa e quantitativa de regras induzidas por diferentes algoritmos de AM simbólico. A análise qualitativa é efetuada por meio de procedimentos que implementam medidas de avaliação de regras (Lavrac et al., 1999). A análise quantitativa é implementada por meio do cálculo de medidas objetivas de interessabilidade de regras (Freitas, 1998b; Freitas, 1998a)

A seguir são sucintamente apresentadas as funcionalidades implementadas no MAR.

2.1 Análise Qualitativa de Regras

Dado um exemplo $\mathbf{x}_i$ e a classe y_i a qual ele pertence, a avaliação da regra no formato $B \rightarrow H$ é feita através de um procedimento que determina se B é verdadeiro ou falso e se H é verdadeiro ou falso para esse exemplo. Por meio desse procedimento é possível determinar, para cada regra R_u de um classificador simbólico $\mathbf{h} = \{R_1, ..., R_{NR}\}$, as seguintes informações:

hb número de exemplos para os quais H é verdade e B é verdade e
$\overline{h}b$ número de exemplos para os quais H é falso e B é verdade e
$\overline{hb}$ número de exemplos para os quais H é falso e B é falso e
$h\overline{b}$ número de exemplos para os quais H é verdade e B é falso.

No cálculo dessas informações deve ser considerado a forma que as regras do classificador simbólico

h foram induzidas pelo algoritmo de AM. O Módulo de Análise de Regras calcula essas informações para regras *unordered*, *ordered* e *interclass*. Maiores detalhes a respeito do cálculo dessas informações são apresentados em (Gomes and Monard, 2002).

Após determinar as informações de avaliação para cada uma das regras R_u de um classificador simbólico **h**, os valores encontrados podem ser expressos sob a forma de freqüências relativas as quais são utilizadas como uma estimativa de probabilidade. Por exemplo, considerando a informação hb, a probabilidade $P(HB)$ pode ser determinada da seguinte forma:

$$P(HB) = f_{bh} = \frac{hb}{N}$$

onde N é o número total de exemplos utilizados para calcular a informação hb.

De forma semelhante podem ser determinados os valores das probabilidades $P(\overline{H}B), P(\overline{HB})$ e $P(H\overline{B})$. Conhecidas essas probabilidades, os valores de $P(B)$, $P(\overline{B})$, $P(H)$ e $P(\overline{H})$ podem ser determinados. Por exemplo:

$$P(B) = P(\overline{H}B) + P(HB) \quad ou \quad P(\overline{B}) = P(H\overline{B}) + P(\overline{HB})$$

No MAR essas probabilidades são utilizadas para calcular as medidas de avaliação de regras propostas no *framework* de Lavrac em (Lavrac et al., 1999). Essas medidas podem ser divididas nas seguintes categorias:

1. Medidas genéricas de avaliação;

Precisão	$Acc(B \rightarrow H) = P(H\|B) = \frac{hb}{b}$
Erro	$Err(B \rightarrow H) = P(\overline{H}\|B) = \frac{\overline{h}b}{b}$
Confiança negativa	$NegRel(B \rightarrow H) = P(\overline{H}\|\overline{B}) = \frac{\overline{hb}}{\overline{b}}$
Sensitividade	$Sens(B \rightarrow H) = P(B\|H) = \frac{hb}{h}$
Especificidade	$Spec(B \rightarrow H) = P(\overline{B}\|\overline{H}) = \frac{\overline{hb}}{\overline{h}}$
Cobertura	$Cov(B \rightarrow H) = P(B) = \frac{b}{N}$
Suporte	$Sup(B \rightarrow H) = P(HB) = \frac{hb}{N}$
Novidade	$Nov(B \rightarrow H) = P(HB) - P(H)P(B) = \frac{1}{N}(hb - \frac{h \cdot b}{N})$
Satisfação	$Sat(B \rightarrow H) = \frac{(P(\overline{H}) - P(\overline{H}\|B))}{P(\overline{H})} = \frac{\frac{\overline{h}}{N} - \frac{\overline{h}b}{b}}{\frac{\overline{h}}{N}} = 1 - \left(\frac{\overline{h}b}{b} \cdot \frac{N}{\overline{h}}\right)$

2. Medidas relativas de avaliação;

Precisão relativa	$RAcc(B \rightarrow H) = P(H\|B) - P(H) = \frac{hb}{b} - \frac{h}{N}$
Confiança negativa relativa	$RNegRel(B \rightarrow H) = P(\overline{H}\|\overline{B}) - P(\overline{H}) = \frac{\overline{hb}}{\overline{b}} - \frac{\overline{h}}{N}$
Sensitividade relativa	$RSens(B \rightarrow H) = P(B\|H) - P(B) = \frac{hb}{h} - \frac{b}{N}$
Especificidade relativa	$RSpec(B \rightarrow H) = P(\overline{B}\|\overline{H}) - P(\overline{B})\ \frac{\overline{hb}}{\overline{h}} - \frac{\overline{b}}{N}$

3. Medidas relativas de avaliação com peso;

Precisão relativa com peso	$WRAcc(B \rightarrow H) = P(B)(P(H\|B) - P(H)) = \frac{b}{N}\left(\frac{hb}{b} - \frac{h}{N}\right)$
Confiança negativa relativa com peso	$WRNegRel(B \rightarrow H) = P(\overline{B})(P(\overline{H}\|\overline{B}) - P(\overline{H})) = \frac{\overline{b}}{N}\left(\frac{\overline{hb}}{\overline{b}} - \frac{\overline{h}}{N}\right)$
Sensitividade relativa com peso	$WRSens(B \rightarrow H) = P(H)(P(B\|H) - P(B)) = \frac{h}{N}\left(\frac{hb}{h} - \frac{b}{N}\right)$
Especificidade relativa com peso	$WRSpec(B \rightarrow H) = P(\overline{H})(P(\overline{B}\|\overline{H}) - P(\overline{B})) = \frac{\overline{h}}{N}\left(\frac{\overline{hb}}{\overline{h}} - \frac{\overline{b}}{N}\right)$

Baseada na medida de novidade, Lavrac mostra que as medidas relativas de avaliação com peso são iguais entre si e iguais a medida de novidade da regra, i.e.

$$WRAcc(B \rightarrow H) = WRSens(B \rightarrow H) = WRSpec(B \rightarrow H) = WRNegRel(B \rightarrow H) = Nov(B \rightarrow H)$$

Maiores detalhes a respeito da abordagem utilizada no trabalho de Lavrac para unificação dessas medidas de avaliação de regras podem ser encontrados em (Gomes, 2001).

Os procedimentos do MAR que possibilitam a análise qualitativa de regras consideram um conjunto de classificadores simbólicos $\mathcal{H} = \{\mathbf{h}_1,...,\mathbf{h}_L\}$ e um conjunto de exemplos $S = \{(\mathbf{x}_1,y_1),...,(\mathbf{x}_N,y_N)\}$. Esses procedimentos são:

1. Cálculo de Informações de Regras - `evaluateAllSetOfRuleFrequency/1`.
2. Cálculo de Medidas de Avaliação de Regras - `calculateMeasuresOverAllSetOFRules/1`.

O procedimento `evaluateAllSetOfRuleFrequency/1` implementa o cálculo das informações para cada classificador simbólico $\mathbf{h} \in \mathcal{H}$, aplicando todos os exemplos de S em cada regra R_u de $\mathbf{h}$. Essas informações são expressas sob a forma de freqüências relativas, i.e. f_{hb}, $f_{\overline{h}b}$, $f_{h\overline{b}}$, $f_{\overline{hb}}$ e são calculadas levando em conta, separadamente, exemplos com valores conhecidos e exemplos com valores desconhecidos. O procedimento `evaluateAllSetOfRuleFrequency/1` adiciona as informações expressas como freqüências relativas ao final de cada regra R_u de cada hipótese $\mathbf{h} \in \mathcal{H}$.

O procedimento `calculateMeasuresOverAllSetOFRules/1` implementa o cálculo das medidas de avaliação de regra, utilizando as informações expressas como freqüências relativas calculadas pelo procedimento `evaluateAllSetOfRuleFrequency/1`. As medidas de avaliação de cada regra podem ser referenciadas, após terem sido calculadas pelo procedimento `calculateMeasuresOverAllSetOFRules/1`, segundo os nomes na Tabela 1.

`accR`:	Precisão da regra
`errR`:	Erro da regra
`negrelR`:	Confiança negativa da regra
`sensR`:	Sensitividade da regra
`specR`:	Especificidade da regra
`covR`:	Cobertura da regra
`supR`:	Suporte da regra
`novR`:	Medida que mostra o quanto uma regra é nova, interessante ou fora do comum
`satR`:	Satisfação da regra
`raccR`:	Precisão relativa da regra
`rnegrelR`:	Confiança negativa relativa da regra
`rsensR`:	Sensitividade relativa da regra
`rspecR`:	Especificidade relativa da regra
`wraccR`:	Precisão relativa com peso da regra
`wrnegrelR`:	Confiança negativa relativa com peso da regra
`wrsensR`:	Sensitividade relativa com peso da regra
`wrspecR`:	Especificidade relativa com peso da regra

Tabela 1: Identificação das Medidas de Avaliação da Regra

Geralmente, medidas de avaliação são úteis para determinar quais são as "melhores" regras de um classificador simbólico **h**. As "melhores" regras podem ser aquelas que apresentam o maior (ou menor) valor para uma determinada medida de avaliação. Por exemplo, considerando um classificador simbólico **h** pode ser considerado que as "melhores" regras R_u são aquelas que possuem maior valor para a medida de Precisão (ou menor valor para a medida de Erro).

2.2 Análise Quantitativa de Regras

No Módulo de Análise de Regras a análise quantitativa é implementada por meio do cálculo de medidas objetivas de interessabilidade de regras propostas nos trabalhos de Freitas (Freitas, 1998b; Freitas, 1998a) que foram inicialmente estudas e implementadas por Horst em (Horst, 1999).

Essas medidas são:

1. Surpresa de Pequenos Disjuntos e
2. Surpresa dos Atributos Individuais da Regra.

Maiores detalhes sobre cada uma dessas medidas objetivas de interessabilidade veja (Gomes and Monard, 2002).

A idéia geral da medida de Surpresa de Pequenos Disjuntos considera um *pequeno disjunto* como conhecimento surpresa quando esse disjunto prediz uma classe diferente das classes preditas pelas suas generalizações mínimas. Saber se uma regra é um pequeno disjunto depende de algum critério específico relacionado com os exemplos, tal como um *threshold* ou outro critério mais flexível. Geralmente, o *threshold* adotado é o número de exemplos cobertos pela regra.

Um pequeno disjunto de m condições terá g generalizações, onde $g = 1..m$. Uma maneira de obter GM generalizações mínimas de um disjunto é através da remoção de uma condição do disjunto original. A classe y_i mais freqüente nos exemplos cobertos por uma generalização mínima GM é a classe predita por essa generalização. O número de vezes que as classes y_i das generalizações mínimas diferem da classe C_v do pequeno disjunto é o valor do grau de surpresa do disjunto. Essa medida de interessabilidade da regra pode ser calculada através da fórmula:

$$SurpDisj(R) = \sum_{g=1}^{m} diferente(y_i, C_v) \tag{1}$$

onde

$$diferente = \begin{cases} 1 & \text{se } y_i \neq C_v \\ 0 & \text{caso contrário} \end{cases}$$

O valor do grau de surpresa pode ser normalizado pelo número de condições m presentes no disjunto. Essa normalização procura evitar a confusão entre medidas de complexidade sintática e a medida de surpresa de disjuntos.

$$SurpDisj_{norm}(R) = \frac{SurpDisj(R)}{m} \tag{2}$$

A segunda medida proposta, Surpresa dos Atributos Individuais da Regra, no caso de atributos discretos, é determinada através do poder de predição dos atributos individuais discretos presentes no corpo de uma regra R_u, ou seja, por meio dos atributos que tenham alto ganho de informação. Quanto mais atributos com alto ganho de informação estiverem presentes no corpo de uma regra, mais interessante a regra será. Essa medida de surpresa da regra pode ser calculada através da fórmula:

$$SurpAtr(R) = \frac{1}{\sum_{p=1}^{m} \frac{GInfo(At_p)}{m}} \tag{3}$$

na qual

$$GInfo(At_p) = Info(C) - Info(C|At_p) \tag{4}$$

com

$$Info(C) = -\sum_{v=1}^{NCl} \Pr(C_v) \log \Pr(C_v) \tag{5}$$

$$Info(C|At_p) = \sum_{q=1}^{NClp} \Pr(At_{pq})(-\sum_{v=1}^{NCl} \Pr(C_v|At_{pq}) \log \Pr(C_v|At_{pq})) \quad (6)$$

onde:

$Info(C)$ é a informação da classe;
$Info(C|At_p)$ é a informação do atributo classe C dado o atributo At_p;
At_{pq} denota o q-ésimo valor do atributo At_p;
C_v denota o v-ésimo valor da classe C;
NCl é o número de classes;
$NClp$ representa o número de valores possíveis do atributo At_p;
todos os log estão na base 2.

As probabilidades $\Pr(C_v)$, $\Pr(At_{pq})$ e $\Pr(C_v|At_{pq})$ podem ser obtidas aproximando-as das respectivas freqüências relativas.

Os procedimentos do MAR que possibilitam a análise quantitativa de regras, implementando medidas objetivas de interessabilidade, são:

1. Cálculo do Grau de Surpresa de Pequenos Disjuntos - `smallDisjSurp/3`
2. Cálculo do Grau de Surpresa dos Atributos Individuais da Regra - `indAttSurp/2`

Considerando uma hipótese **h**, o procedimento `smallDisjSurp/3` encontra todas as regras $R_u \in \mathbf{h}$ cujo número de exemplos cobertos é menor ou igual ao valor Vt (pequenos disjuntos). Encontrados os pequenos disjuntos, o procedimento efetua o cálculo (inclusive normalizado) do grau de Surpresa dos Pequenos Disjuntos, para cada regra $R_u \in \mathbf{h}$ segundo a Equação 1.

Considerando uma hipótese **h**, o procedimento`indAttSurp/2` efetua o cálculo do grau de Surpresa dos Atributos Individuais Discretos da Regra para cada regra $R_u \in \mathbf{h}$, segundo a Equação 3.

As medidas de interessabilidade aqui apresentadas são calculadas de forma independente do domínio de aplicação e de impressões que o usuário tem a respeito da regra. Num processo de análise quantitativa, quanto mais alto o valor de uma dessas medidas para uma regra R_u, maior é a chance de que o conhecimento descrito por R_u seja um conhecimento interessante ou inesperado.

3 Processamento de Sêmen Diagnóstico

O estudo de caso sobre processamento de sêmem diagnóstico teve seu início em (Lee, 2000), onde é descrita a importância desse processamento diagnóstico no tratamento para a reprodução assistida. A explicação a seguir está baseada nos trabalhos de Huei Diana Lee (Lee and Monard, 2000b; Lee and Monard, 2000a; Esteves et al., 2001) e foi desenvolvida de forma conjunta com Flávia Cristina Bernardini (Bernardini and Monard, 2002b). Esse exame permite:

1. Quantificar com precisão a qualidade do sêmen (processamento de sêmen diagnóstico);

2. Recuperar a maior quantidade possível de espermatozóides para a utilização na reprodução assistida (processamento de sêmen terapêutico).

A quantidade de espermatozóides recuperados pelo processamento de sêmen influencia na escolha da técnica que será utilizada no tratamento. São utilizadas três técnicas no tratamento para a reprodução assistida:

1. Inseminação Intra Uterina — IUI;
2. Fertilização In Vitro — FIV e
3. Injeção Intracitoplasmática do Espermatozóide no Oócito — ICSI.

Deve ser observado que o processamento de sêmen é bastante custoso. A realização do processamento de sêmen pode elevar o custo do exame em aproximadamente 80% do valor de um espermograma. Essa elevação de custo se deve principalmente a três fatores: necessidade de equipamentos especiais, mão de obra qualificada e tempo gasto para a realização do exame (Esteves and Bento, 1998; Esteves et al., 2000; Ferro et al., 2002). Assim, um dos interesses do estudo desse tema é tentar predizer qual será a quantidade de espermatozóides recuperados pelo processamento de sêmen antes mesmo da realização desse exame, a partir de exames menos custosos, como o espermograma. Dessa forma, dependendo da qualidade da predição, o especialista poderia decidir por uma técnica sem a necessidade da realização do processamento de sêmen. Um outro interesse para o estudo desse tema é a extração e avaliação do conhecimento adquirido.

O processamento de sêmen, através do fornecimento de melhores condições, permite que o maior número de espermatozóides móveis (motilidade) seja recuperado. Os espermatozóides são classificados em graus A, B, C e D dependendo de sua motilidade:

- grau A: espermatozóides que apresentam o maior grau de motilidade;
- grau B: espermatozóides que apresentam um menor grau de motilidade;
- grau C: espermatozóides que se movem em círculos;
- grau D: espermatozóides que são imóveis.

Os graus de motilidade são atributos medidos tanto no espermograma quanto no processamento de sêmen. Os atributos medidos durante o processamento de sêmen são utilizados para a determinação das classes e os atributos medidos através do espermograma são fornecidos como atributos ao algoritmo de indução. Em outras palavras, os valores dos atributos de cada exemplo (paciente) são medidos utilizando um exame pouco custoso (espermograma, neste caso) enquanto que as classes desses exemplos são determinadas através de um outro exame mais custoso (processamento de sêmen, neste trabalho). A idéia é tentar verificar se é possível descobrir um relacionamento entre eles tal que, pelo menos para novos pacientes a necessidade de realizar o exame mais custoso seja minimizada. Este é um procedimento freqüentemente utilizado na área de medicina a fim de tentar diminuir, mas com segurança, o custo de um tratamento.

O especialista sugeriu duas possíveis opções para os valores considerados na classificação dos exemplos:

- A: considerar apenas a percentagem de espermatozóides classificados como de motilidade grau A no processamento de sêmen ou
- AB: considerar a percentagem de espermatozóides classificados como de motilidade grau A e grau B no processamento de sêmen.

Assim, foram definidas três classes, baseadas nas técnicas que são utilizadas no tratamento para a reprodução assistida:

- 1: $x < 1$ — IUI;
- 2: $1 \geq x < 5$ — FIV e
- 3: $x \geq 5$ — ICSI.

onde x representa milhões de espermatozóides por mililitro (ml).

Os casos para a composição desses conjunto de dados foram extraídos de casos reais do Centro de Fertilização, em Campinas, SP fornecidos pelo especialista Dr. Sandro Esteves. Em (Lee, 2000), foram calculados para cada caso os valores A e AB, gerando dois conjuntos de dados, *Proc-a* e *Proc-ab*, respectivamente, cada um com 231 exemplos. Diversos experimentos foram conduzidos considerando-se esses dois conjuntos separadamente. Com o decorrer do trabalho, necessitou-se criar novos atributos como combinação de atributos originais, criando-se diferentes conjuntos de dados (Lee and Monard, 2000b). Um desses conjunto de dados que mostrou um bom resultado com relação à predição dos classificadores induzidos foi o denominado *Proc-a-gmg-d*, o qual considera somente os exemplos em que o processamento de sêmen é diagnóstico.

Posteriormente, o conjunto de exemplos *Proc-a* do trabalho inicial foram acrescentados novos casos. Tomando-se desse conjunto de exemplos somente os exemplos em que o processamento de sêmen é diagnóstico, foi obtido um segundo conjunto de exemplos *Proc-a-gmg-d*,semelhante ao construído com os exemplos *Proc-a* do trabalho inicial, mas incluindo os novos casos, o qual é utilizado neste trabalho. Deve ser observado que todo o trabalho de coleta, preparação e limpeza dos dados foi realizada por Huei Diana Lee.

4 Etapas do Estudo de Caso

Conforme mencionado anteriormente, o objetivo do estudo apresentado neste trabalho é efetuar a análise qualitativa e quantitativa das regras induzidas utilizando os dados descritos na Seção 4.1, através da execução dos procedimentos do Módulo de Análise de Regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Assim, vários conjuntos de regras, utilizando diferentes algoritmos de AM, foram induzidos a partir do conjunto de dados *Proc-a-gmg-d*.

As etapas envolvidas nesse estudo são:

- Descrição e preparação do conjunto de dados;
- Indução e preparação dos conjuntos de regras induzidas por diferentes algoritmos de AM;

- Análise quantitativa dos conjuntos de regras; e
- Análise qualitativa dos conjuntos de regras.

4.1 Descrição e Preparação do Conjunto de Dados

A Tabela 2 sumariza as características do conjunto de dados utilizado neste estudo. Ela mostra o número de exemplos (#Exemplos), número e porcentagem de exemplos com valores duplicados (que aparecem mais de uma vez) ou conflitantes (que possuem o mesmo atributo-valor mas têm diferentes classificações), número de atributos (#Atributos) contínuos e nominais incluindo atributo classificatório, distribuição de classes, o erro majoritário e se o conjunto de dados tem ao menos um valor desconhecido[4]. Na Tabela 3 são descritos os atributos que compõem o conjunto de dados *Proc-a-gmg-d*.

Conjunto de Dados	# Exemplos	# Duplicados ou Conflitantes (%)	# Atributos (cont.,nom.)	Classes	Classe %	Erro CM	Valores Desconhecidos
Proc-a-gmg-d	240	1 (0.416%)	22 (17, 5)	um	42.92%	57.08%	Sim
				dois	22.08%	sobre um	
				tres	35.00%		

Tabela 2: Sumário das Características do Conjunto de Dados *Proc-a-gmg-d*

O conjunto de dados *Proc-a-gmg-d* é utilizado tanto na indução dos classificadores simbólicos quanto na análise dos conjuntos de regras realizada pelo MAR.

Para executar-se os procedimentos implementados no MAR, o formato do conjunto de dados *Proc-a-gmg-d* deve primeiramente ser convertido para a sintaxe padrão Prolog do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ (Gomes et al., 2002). Para ilustrar, na Figura 1 é apresentado apenas um dos 240 exemplos e os atributos do conjunto de dados *Proc-a-gmg-d* na sintaxe padrão Prolog do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$.

4.2 Indução e Preparação dos Conjuntos de Regras

Foram induzidos dois classificadores simbólicos a partir do conjunto de dados *Proc-a-gmg-d*. Utilizando o algoritmo de AM $\mathcal{CN}2$ (Clark and Niblett, 1989) foi induzido um conjunto de regras *unordered* — $\mathbf{h}_1$ — apresentado no Apêndice A. Utilizando o algoritmo $\mathcal{S}$`See5` [5] foi induzido um conjunto de regras *interclass* — $\mathbf{h}_2$ — apresentado no Apêndice B.

Logo após, utilizando o conversor descrito em (Gomes et al., 2002), ambos conjuntos de regras foram convertidos para a sintaxe padrão Prolog do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Nessa conversão automática, cada regra do classificador $\mathbf{h}_1$ passou a ser identificada por `id_`105 e as do classificador $\mathbf{h}_2$ por `id_`104. Os Apêndices C e D mostram, respectivamente, as regras que constituem $\mathbf{h}_1$ e $\mathbf{h}_2$ na sintaxe padrão Prolog, incluindo os valores das freqüências adicionadas ao final de cada regra — vide Seção 4.3.

[4] Essa informação foi obtida utilizando a ferramenta *info* da $\mathcal{MLC}$++.
[5] *http://www.rulequest.com*

Número da Feature	Nome da Feature	Descrição	#Valores Distintos		
			possíveis	atuais	tipo
#0	idade	Idade do paciente	—	26	contínuo
#1	hora	Período da coleta da amostra de sêmen			
		1: manhã; 2: tarde	2	3	discreto
#2	processamento	Processamento do sêmen realizado após			
		esta quantidade de minutos após a coleta	—	12	contínuo
#3	tempo-abs	Tempo de abstinência	—	12	contínuo
#4	volume	Volume de sêmen coletado	—	55	contínuo
#5	cor	Cor do sêmen coletado			
		1: branco-opalescente (normal);			
		2: amarelo-opalescente; 3: translúcido	3	4	discreto
#6	odor	Odor do sêmen coletado			
		1: característico; 2: forte; 3: urina	3	3	discreto
#7	pH	pH do sêmen coletado	—	13	contínuo
#8	viscosidade	Viscosidade do sêmen coletado			
		1: normal; 2: aumentada	2	2	discreto
#9	liquefacao	Liquefação do sêmen coletado			
		1: completa; 2: imconpleta	2	2	discreto
#10	concentracao	Concentração de espermatozóides			
		por ml coletado	—	198	contínuo
#11	concentracao-total	Concentração total de espermatozóides	—	219	contínuo
const #12	motilidade	% de espermatozóides móveis	—	74	contínuo
#13	class-A	Classificação da motilidade - Grau A	—	66	contínuo
#14	class-B	Grau B	—	53	contínuo
#15	class-C	Grau C	—	44	contínuo
#16	class-D	Grau D	—	76	contínuo
#17	vitalidade	% de espermatozóides vivos	—	59	contínuo
#18	num-leu	Número de leucócitos	—	66	contínuo
#19	Kruger	Morfologia Estrita de Kruger			
		- % de formas normais	—	28	contínuo
#20	HP	Teste Hipo-osmótico - % de inchados	—	54	contínuo

Tabela 3: Descrição dos Atributos do Conjunto de Dados *Proc-a-gmg-d*

```
ex(0,[31, manha, 20, 2, 4.9, translucido, caracteristico, 8,
aumentada, incompleta, 9.6, 96.04, 0, 73, 5, 22, 73, 79, 0, 3, 63,
dois]).

feature(0,idade,integer).
feature(1,hora,enumerate(manha,tarde)).
feature(2,processamento,integer).
feature(3,tempo_abs,integer).
feature(4,volume, float).
feature(5,cor,enumerated(branco,amarelo,translucido)).
feature(6,odor, enumerated(caracteristico, forte, urina)).
feature(7,ph, float).
feature(8,viscosidade, enumerated(normal, aumentada)).
feature(9,liquefacao,enumerated(completa, incompleta)).
feature(10,concentracao, float).
feature(11,concentracao_total,float).
feature(12,motilidade, integer).
feature(13,class_a, float).
feature(14,class_b, float).
feature(15,class_c, float).
feature(16,class_d, float).
feature(17,vitalidade, integer).
feature(18,num_leu, float).
feature(19,kruger, float).
feature(20,hp,integer).
feature(21,class,enumerated(um,dois,tres)).

classFeature(class).
```

Figura 1: Um Exemplo e os Atributos do Conjunto de Dados *Proc-a-gmg-d* na Sintaxe Padrão Prolog do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$

Além de induzir as hipóteses $\mathbf{h}_1$ e $\mathbf{h}_2$ utilizando todos os exemplos disponíveis (240 exemplos) foi também estimado o erro de $\mathbf{h}_1$ e $\mathbf{h}_2$ utilizando a técnica *10-fold-cross-validation*. A Tabela 4 mostra os resultados obtidos.

Indutor Utilizado	Classificador	Número Regras Induzidas[a]	Forma de Indução das Regras	Erro Aparente	Erro ± Desvio Padrão
$\mathcal{CN}2$	$\mathbf{h}_1$	32	*unordered*	$12,9\%$	$41.66\% \pm 5.13\%$
See5	$\mathbf{h}_2$	19	*interclass*	$15,4\%$	$44.20\% \pm 2.90\%$

[a]Na contagem do número de regras não foi considerada a regra *default*.

Tabela 4: Informações sobre os Conjuntos de Regras Induzidas

Como pode ser observado, a estimativa de erro de cada classificador em novos exemplos ainda não vistos, i.e., $41.66\% \pm 5.13\%$ para $\mathbf{h}_1$ e $44.20\% \pm 2.90\%$ para $\mathbf{h}_2$, ainda que inferior ao erro da classe majoritária (57.08% na Tabela 2) é muito alto para a área médica. Isso significa que $\mathbf{h}_1$ e $\mathbf{h}_2$, considerados como caixa preta, não são apropriadas para classificar (rotular) qualquer outro novo caso desse domínio.

Entretanto, analisando individualmente as regras que constituem cada classificador, existe a possibilidade de encontrar, entre outros, algumas regras apropriadas para classificar com bastante confiança alguns desses novos casos.

4.3 Análise Qualitativa dos Conjuntos de Regras

Como mencionado anteriormente, os procedimentos do MAR que possibilitam a análise qualitativa de regras são:

1. Cálculo de Informações de Regras - `evaluateAllSetOfRuleFrequency/1`.
2. Cálculo de Medidas de Avaliação de Regras - `calculateMeasuresOverAllSetOFRules/1`.

Ao ser executado o procedimento `evaluateAllSetOfRuleFrequency/1` as informações f_{hb}, $f_{\overline{h}b}$, $f_{\overline{hb}}$, $f_{h\overline{b}}$, referentes aos exemplos com valores conhecidos e desconhecidos, são calculadas e adicionadas ao final de cada regra de $\mathbf{h}_1$ e $\mathbf{h}_2$.

Deve ser ressaltado que essas informações, ao serem calculadas para as regras de $\mathbf{h}_1$, diferem um pouco dos valores apresentados no classificador equivalente induzido pelo algoritmo $\mathcal{CN}2$. Isso deve-se ao fato de que, se o exemplo apresentar atributos com valores desconhecidos, ele sempre será contado como verdade no cálculo das informações realizado pelo MAR. Em outras palavras o procedimento `evaluateAllSetOfRuleFrequency/1` sempre conta 1 para B e/ou H verdade para esse exemplo. Entretanto, o $\mathcal{CN}2$ trata esse caso contando 0.5 para cada exemplo com valores desconhecidos, coberto por uma regra.

Como já mencionado, os Apêndices C e D mostram, respectivamente, as regras de $\mathbf{h}_1$ e $\mathbf{h}_2$ incluindo o cálculo das freqüências realizado pelo procedimento `evaluateAllSetOfRuleFrequency/1`. Prosseguindo com a análise qualitativa desses dois conjuntos de regras, foram calculadas medidas de avaliação por meio da execução do procedimento `calculateMeasuresOverAllSetOFRules/1`.

Os valores das medidas genéricas encontradas para as regras de $\mathbf{h}_1$ e $\mathbf{h}_2$ são apresentados, respectivamente, nas Tabela 5 e 6. As medidas relativas, com e sem peso, calculadas para as regras de $\mathbf{h}_1$ e $\mathbf{h}_2$, são apresentadas, respectivamente, nas Tabelas 7 e 8. Para melhor compreensão, em cada uma dessa tabelas são apresentados os valores hb, $\overline{h}b$, $\overline{hb}$, $h\overline{b}$ utilizados no cálculo das freqüências f_{hb}, $f_{\overline{h}b}$, $f_{\overline{hb}}$, $f_{h\overline{b}}$.

# da	Medidas Genéricas						Valores Absolutos
Regra	`accR`	`negrelR`	`sensR`	`covR`	`supR`	`novR`	$[hb, \overline{h}b, h\overline{b}, \overline{hb}, N]$
1	1.0000	0.6587	0.3107	0.1333	0.1333	0.0761	[32, 0, 71, 137, 240]
2	1.0000	0.6343	0.2330	0.1000	0.1000	0.0571	[24, 0, 79, 137, 240]
3	1.0000	0.6143	0.1650	0.0708	0.0708	0.0404	[17, 0, 86, 137, 240]
4	1.0000	0.6313	0.2233	0.0958	0.0958	0.0547	[23, 0, 80, 137, 240]
5	1.0000	0.6116	0.1553	0.0667	0.0667	0.0381	[16, 0, 87, 137, 240]
6	1.0000	0.6524	0.2913	0.1250	0.1250	0.0714	[30, 0, 73, 137, 240]
7	1.0000	0.5905	0.0777	0.0333	0.0333	0.0190	[8, 0, 95, 137, 240]
8	1.0000	0.5880	0.0680	0.0292	0.0292	0.0166	[7, 0, 96, 137, 240]
9	1.0000	0.5805	0.0388	0.0167	0.0167	0.0095	[4, 0, 99, 137, 240]
10	1.0000	0.6171	0.1748	0.0750	0.0750	0.0428	[18, 0, 85, 137, 240]
11	1.0000	0.5732	0.0097	0.0042	0.0042	0.0024	[1, 0, 102, 137, 240]
12	1.0000	0.8026	0.1321	0.0292	0.0292	0.0227	[7, 0, 46, 187, 240]
13	1.0000	0.8095	0.1698	0.0375	0.0375	0.0292	[9, 0, 44, 187, 240]
14	1.0000	0.7890	0.0566	0.0125	0.0125	0.0097	[3, 0, 50, 187, 240]
15	1.0000	0.7857	0.0377	0.0083	0.0083	0.0065	[2, 0, 51, 187, 240]
16	1.0000	0.7857	0.0377	0.0083	0.0083	0.0065	[2, 0, 51, 187, 240]
17	1.0000	0.8026	0.1321	0.0292	0.0292	0.0227	[7, 0, 46, 187, 240]
18	1.0000	0.7824	0.0189	0.0042	0.0042	0.0032	[1, 0, 52, 187, 240]
19	1.0000	0.7824	0.0189	0.0042	0.0042	0.0032	[1, 0, 52, 187, 240]
20	1.0000	0.7824	0.0189	0.0042	0.0042	0.0032	[1, 0, 52, 187, 240]
21	1.0000	0.6996	0.2024	0.0708	0.0708	0.0460	[17, 0, 67, 156, 240]
22	1.0000	0.6964	0.1905	0.0667	0.0667	0.0433	[16, 0, 68, 156, 240]
23	1.0000	0.7027	0.2143	0.0750	0.0750	0.0488	[18, 0, 66, 156, 240]
24	1.0000	0.7091	0.2381	0.0833	0.0833	0.0542	[20, 0, 64, 156, 240]
25	1.0000	0.6933	0.1786	0.0625	0.0625	0.0406	[15, 0, 69, 156, 240]
26	1.0000	0.6933	0.1786	0.0625	0.0625	0.0406	[15, 0, 69, 156, 240]
27	1.0000	0.6610	0.0476	0.0167	0.0167	0.0108	[4, 0, 80, 156, 240]
28	1.0000	0.6582	0.0357	0.0125	0.0125	0.0081	[3, 0, 81, 156, 240]
29	1.0000	0.6638	0.0595	0.0208	0.0208	0.0135	[5, 0, 79, 156, 240]
30	1.0000	0.6527	0.0119	0.0042	0.0042	0.0027	[1, 0, 83, 156, 240]
31	1.0000	0.6667	0.0714	0.0250	0.0250	0.0163	[6, 0, 78, 156, 240]
32	1.0000	0.6783	0.1190	0.0417	0.0417	0.0271	[10, 0, 74, 156, 240]

Tabela 5: Valores das Medidas Genéricas para as Regras de $\mathbf{h}_1$

4.4 Análise Quantitativa dos Conjuntos de Regras

Como mencionado anteriormente, no Módulo de Análise de Regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ os procedimentos que possibilitam a análise quantitativa de regras, implementando medidas objetivas de interessabilidade, são:

1. Cálculo do Grau de Surpresa de Pequenos Disjuntos - `smallDisjSurp/3`
2. Cálculo do Grau de Surpresa dos Atributos Individuais da Regra - `indAttSurp/2`

# da Regra	Medidas Genéricas									Valores Absolutos
	accR	errR	negrelR	sensR	specR	covR	supR	novR	satR	$[hb, \overline{h}b, h\overline{b}, \overline{h}\overline{b}, N]$
1	0.9111	0.0889	0.6821	0.3981	0.9708	0.1875	0.1708	0.0904	0.8443	[41, 4, 62, 133, 240]
2	1.0000	0.0000	0.5957	0.0971	1.0000	0.0417	0.0417	0.0238	1.0000	[10, 0, 93, 137, 240]
3	1.0000	0.0000	0.5931	0.0874	1.0000	0.0375	0.0375	0.0214	1.0000	[9, 0, 94, 137, 240]
4	0.8800	0.1200	0.6233	0.2136	0.9781	0.1042	0.0917	0.0470	0.7898	[22, 3, 81, 134, 240]
5	0.8493	0.1507	0.7545	0.6019	0.9197	0.3042	0.2583	0.1278	0.7360	[62, 11, 41, 126, 240]
6	1.0000	0.0000	0.5805	0.0388	1.0000	0.0167	0.0167	0.0095	1.0000	[4, 0, 99, 137, 240]
7	0.8750	0.1250	0.5862	0.0680	0.9927	0.0333	0.0292	0.0149	0.7810	[7, 1, 96, 136, 240]
8	0.8125	0.1875	0.5982	0.1262	0.9781	0.0667	0.0542	0.0256	0.6715	[13, 3, 90, 134, 240]
9	1.0000	0.0000	0.7991	0.1132	1.0000	0.0250	0.0250	0.0195	1.0000	[6, 0, 47, 187, 240]
10	1.0000	0.0000	0.7957	0.0943	1.0000	0.0208	0.0208	0.0162	1.0000	[5, 0, 48, 187, 240]
11	0.8571	0.1429	0.7983	0.1132	0.9947	0.0292	0.0250	0.0186	0.8167	[6, 1, 47, 186, 240]
12	1.0000	0.0000	0.7824	0.0189	1.0000	0.0042	0.0042	0.0032	1.0000	[1, 0, 52, 187, 240]
13	0.8000	0.2000	0.7915	0.0755	0.9947	0.0208	0.0167	0.0121	0.7433	[4, 1, 49, 186, 240]
14	0.7500	0.2500	0.7881	0.0566	0.9947	0.0167	0.0125	0.0088	0.6791	[3, 1, 50, 186, 240]
15	0.7500	0.2500	0.7881	0.0566	0.9947	0.0167	0.0125	0.0088	0.6791	[3, 1, 50, 186, 240]
16	1.0000	0.0000	0.7857	0.0377	1.0000	0.0083	0.0083	0.0065	1.0000	[2, 0, 51, 187, 240]
17	1.0000	0.0000	0.7027	0.2143	1.0000	0.0750	0.0750	0.0488	1.0000	[18, 0, 66, 156, 240]
18	0.8571	0.1429	0.7576	0.4286	0.9615	0.1750	0.1500	0.0888	0.7802	[36, 6, 48, 150, 240]
19	0.8161	0.1839	0.9150	0.8452	0.8974	0.3625	0.2958	0.1690	0.7171	[71, 16, 13, 140, 240]

Tabela 6: Valores das Medidas Genéricas para as Regras de $\mathbf{h}_2$

De acordo com a medida de Surpresa de Pequenos Disjuntos, nas Figura 2 e 3 são apresentados, respectivamente, em ordem crescente dos valores calculados pelo MAR[6], os pequenos disjuntos que fazem parte dos conjuntos de regras $\mathbf{h}_1$ e $\mathbf{h}_2$. Esses valores foram obtidos admitindo-se como critério para determinar quais regras são pequenos disjuntos, o número de exemplos cobertos pela regra. No caso aqui apresentado, pequenos disjuntos são regras que cobrem no máximo 10% dos exemplos (ou seja, regras com $Cov(R) \leq 0.10$).

De acordo com a medida de Surpresa dos Atributos Individuais Discretos, as Figura 4 e 5 mostram, respectivamente, as regras selecionadas do conjunto de regras $\mathbf{h}_1$ e $\mathbf{h}_2$.

[6]Foram desconsiderados pequenos disjuntos com grau de surpresa igual a 0 (zero).

# da Regra	Medidas Relativas					Valores Absolutos
	`raccR`	`rnegrelR`	`rsensR`	`rspecR`	`wraccR`[a]	$[hb, \overline{h}b, h\overline{b}, \overline{hb}, N]$
1	0.5708	0.0878	0.1773	0.1333	0.0761	[32, 0, 71, 137, 240]
2	0.5708	0.0634	0.1330	0.1000	0.0571	[24, 0, 79, 137, 240]
3	0.5708	0.0435	0.0942	0.0708	0.0404	[17, 0, 86, 137, 240]
4	0.5708	0.0605	0.1275	0.0958	0.0547	[23, 0, 80, 137, 240]
5	0.5708	0.0408	0.0887	0.0667	0.0381	[16, 0, 87, 137, 240]
6	0.5708	0.0815	0.1663	0.1250	0.0714	[30, 0, 73, 137, 240]
7	0.5708	0.0197	0.0443	0.0333	0.0190	[8, 0, 95, 137, 240]
8	0.5708	0.0171	0.0388	0.0292	0.0166	[7, 0, 96, 137, 240]
9	0.5708	0.0097	0.0222	0.0167	0.0095	[4, 0, 99, 137, 240]
10	0.5708	0.0463	0.0998	0.0750	0.0428	[18, 0, 85, 137, 240]
11	0.5708	0.0024	0.0055	0.0042	0.0024	[1, 0, 102, 137, 240]
12	0.7792	0.0234	0.1029	0.0292	0.0227	[7, 0, 46, 187, 240]
13	0.7792	0.0304	0.1323	0.0375	0.0292	[9, 0, 44, 187, 240]
14	0.7792	0.0099	0.0441	0.0125	0.0097	[3, 0, 50, 187, 240]
15	0.7792	0.0065	0.0294	0.0083	0.0065	[2, 0, 51, 187, 240]
16	0.7792	0.0065	0.0294	0.0083	0.0065	[2, 0, 51, 187, 240]
17	0.7792	0.0234	0.1029	0.0292	0.0227	[7, 0, 46, 187, 240]
18	0.7792	0.0033	0.0147	0.0042	0.0032	[1, 0, 52, 187, 240]
19	0.7792	0.0033	0.0147	0.0042	0.0032	[1, 0, 52, 187, 240]
20	0.7792	0.0033	0.0147	0.0042	0.0032	[1, 0, 52, 187, 240]
21	0.6500	0.0496	0.1315	0.0708	0.0460	[17, 0, 67, 156, 240]
22	0.6500	0.0464	0.1238	0.0667	0.0433	[16, 0, 68, 156, 240]
23	0.6500	0.0527	0.1393	0.0750	0.0488	[18, 0, 66, 156, 240]
24	0.6500	0.0591	0.1548	0.0833	0.0542	[20, 0, 64, 156, 240]
25	0.6500	0.0433	0.1161	0.0625	0.0406	[15, 0, 69, 156, 240]
26	0.6500	0.0433	0.1161	0.0625	0.0406	[15, 0, 69, 156, 240]
27	0.6500	0.0110	0.0310	0.0167	0.0108	[4, 0, 80, 156, 240]
28	0.6500	0.0082	0.0232	0.0125	0.0081	[3, 0, 81, 156, 240]
29	0.6500	0.0138	0.0387	0.0208	0.0135	[5, 0, 79, 156, 240]
30	0.6500	0.0027	0.0077	0.0042	0.0027	[1, 0, 83, 156, 240]
31	0.6500	0.0167	0.0464	0.0250	0.0163	[6, 0, 78, 156, 240]
32	0.6500	0.0283	0.0774	0.0417	0.0271	[10, 0, 74, 156, 240]

[a]Medidas Relativas com Pesos `wraccR` = `wrnegrelR` = `wrsensR` = `wrspecR` = `novR` (Lavrac et al., 1999)

Tabela 7: Valores das Medidas Relativas, com e sem Peso, para as Regras de $\mathbf{h}_1$

# da Regra	Medidas Relativas					Valores Absolutos
	`raccR`	`rnegrelR`	`rsensR`	`rspecR`	`wraccR`[a]	$[hb, \overline{h}b, h\overline{b}, \overline{h}\overline{b}, N]$
1	0.4819	0.1112	0.2106	0.1583	0.0904	[41, 4, 62, 133, 240]
2	0.5708	0.0248	0.0554	0.0417	0.0238	[10, 0, 93, 137, 240]
3	0.5708	0.0222	0.0499	0.0375	0.0214	[9, 0, 94, 137, 240]
4	0.4508	0.0524	0.1094	0.0823	0.0470	[22, 3, 81, 134, 240]
5	0.4201	0.1837	0.2978	0.2239	0.1278	[62, 11, 41, 126, 240]
6	0.5708	0.0097	0.0222	0.0167	0.0095	[4, 0, 99, 137, 240]
7	0.4458	0.0154	0.0346	0.0260	0.0149	[7, 1, 96, 136, 240]
8	0.3833	0.0274	0.0595	0.0448	0.0256	[13, 3, 90, 134, 240]
9	0.7792	0.0200	0.0882	0.0250	0.0195	[6, 0, 47, 187, 240]
10	0.7792	0.0166	0.0735	0.0208	0.0162	[5, 0, 48, 187, 240]
11	0.6363	0.0191	0.0840	0.0238	0.0186	[6, 1, 47, 186, 240]
12	0.7792	0.0033	0.0147	0.0042	0.0032	[1, 0, 52, 187, 240]
13	0.5792	0.0123	0.0546	0.0155	0.0121	[4, 1, 49, 186, 240]
14	0.5292	0.0090	0.0399	0.0113	0.0088	[3, 1, 50, 186, 240]
15	0.5292	0.0090	0.0399	0.0113	0.0088	[3, 1, 50, 186, 240]
16	0.7792	0.0065	0.0294	0.0083	0.0065	[2, 0, 51, 187, 240]
17	0.6500	0.0527	0.1393	0.0750	0.0488	[18, 0, 66, 156, 240]
18	0.5071	0.1076	0.2536	0.1365	0.0888	[36, 6, 48, 150, 240]
19	0.4661	0.2650	0.4827	0.2599	0.1690	[71, 16, 13, 140, 240]

[a]Medidas Relativas com Pesos `wraccR` = `wrnegrelR` = `wrsensR` = `wrspecR` = `novR` (Lavrac et al., 1999)

Tabela 8: Valores das Medidas Relativas, com e sem Peso, para as Regras de $\mathbf{h}_2$

```
rule(id_105,30,[less(tempo_abs,1.5),less(volume,1)],class(tres)).
SurpDisj : 1.0000 SurpDisjNorm : 0.5000

rule(id_105,16,[(less(idade,34),greater(class_d,47.5),
greater(vitalidade,91.5))],class(dois)).
 SurpDisj : 2.0000 SurpDisjNorm : 0.6667

rule(id_105,15,[less(idade,25.5),less(volume,2.05)],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm :1.0000

rule(id_105,18,[greater(processamento,55),greater(tempo_abs,3.5)],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm : 1.0000

rule(id_105,19,[greater(idade,34),less(class_c,9)],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm : 1.0000

rule(id_105,20,[less(idade,33.5),greater(concentracao,593.4)],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm : 1.0000
```

Figura 2: Surpresa dos Pequenos Disjuntos de $\mathbf{h}_1$

```
rule(id_104,12,[(equal(hora,manha),equal(cor,branco),
equal(liquefacao,completa),lessORequal(concentracao_total,46.2),
greater(hp,62))],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm : 0.4000

rule(id_104,16,[(equal(hora,manha),equal(cor,branco),
lessORequal(concentracao_total,46.2),greater(hp,62))],class(dois)).
SurpDisj : 2.0000 SurpDisjNorm : 0.5000
```

Figura 3: Surpresa dos Pequenos Disjuntos de $\mathbf{h}_2$

```
rule(id_105,32,[(equal(hora,manha),greater(class_d,61),
greater(hp,83.5))],class(tres)).
SurpAttR : 1399.3011
```

Figura 4: Surpresa dos Atributos Individuais Discretos das Regras Selecionadas de $\mathbf{h}_1$

```
rule(id_104,3,[(greater(idade,30),lessORequal(idade,33),
equal(cor,branco),greater(concentracao_total,95.2),
lessORequal(concentracao_total,338.8),lessORequal(motilidade,15),
greater(class_c,18),lessORequal(vitalidade,79),
lessORequal(hp,83))],class(um)).
SurpAttR : 24.4009

rule(id_104,4,[(greater(volume,2.4),equal(cor,branco),
lessORequal(concentracao_total,95.2),lessORequal(class_a,33),
lessORequal(hp,88))],class(um)).
SurpAttR : 24.4009

rule(id_104,7,[(greater(idade,35),lessORequal(idade,36),
equal(cor,branco),lessORequal(motilidade,15))],class(um)).
SurpAttR : 24.4009

rule(id_104,9,[(lessORequal(volume,2.4),equal(cor,branco),
lessORequal(concentracao_total,95.2),greater(motilidade,17),
lessORequal(class_a,33),lessORequal(hp,88))],class(dois)).
SurpAttR : 24.4009

rule(id_104,10,[(greater(idade,33),lessORequal(idade,35),
greater(volume,2.1),equal(cor,branco),greater(concentracao_total,95.2),
lessORequal(concentracao_total,338.8),lessORequal(motilidade,15),
greater(class_c,18),greater(hp,56))],class(dois)).
SurpAttR : 24.4009

rule(id_104,16,[(equal(hora,manha),equal(cor,branco),
lessORequal(concentracao_total,46.2),greater(hp,62))],class(dois)).
SurpAttR : 47.9654

rule(id_104,12,[(equal(hora,manha),equal(cor,branco),
equal(liquefacao,completa),lessORequal(concentracao_total,46.2),
greater(hp,62))],class(dois)).
SurpAttR : 63.8402
```

Figura 5: Surpresa dos Atributos Individuais Discretos das Regras Selecionadas de $\mathbf{h}_2$

5 Discussão dos Resultados Obtidos

Nesta seção é apresentada uma análise dos valores das medidas de avaliação de regras (análise qualitativa) bem como dos valores das medidas objetivas de interessabilidade de regras (análise quantitativa) obtidos nesse estudo de caso.

5.1 Da Análise Qualitativa

Conforme mencionado anteriormente, as "melhores" regras de um classificador simbólico **h** são aquelas que apresentam o maior (ou menor) valor para uma determinada medida de avaliação. Entretanto, como essas medidas representam probabilidades, é importante considerar o número de exemplos utilizados para calcular essas probabilidades. Isto pode ser visto claramente no caso das 32 regras que constituem $\mathbf{h}_1$. Todas elas possuem o mesmo valor de `accR` = 1 — Tabela 5.

Entretanto, a regra 1 cobre corretamente 32 exemplos, enquanto outras regras cobrem menos exemplos. Nesse caso, se a medida de `accR` é importante para o especialista, um outro critério válido seria considerar primeiro as regras com `accR` = 1 que cobrem corretamente mais exemplos. A medida `supR` fornece essa informação. Como pode ser observado, são vários os critérios que podem ser utilizados pelo especialista para analisar as regras segundo essas medidas.

Para melhor visualizar a classificação das regras que constituem $\mathbf{h}_1$ e $\mathbf{h}_2$ segundo as diversas medidas apresentadas, respectivamente, nas Tabelas 5 e 7 para $\mathbf{h}_1$ e nas Tabelas 6 e 8 para $\mathbf{h}_2$, as regras de cada hipótese foram ordenadas de forma decrescente, segundo o "melhor"valor das medidas consideradas. Assim, na Tabela 9 são mostradas as regras de $\mathbf{h}_1$ ordenadas pelo "melhor"valor das medidas de avaliação (genéricas e relativas com e sem peso). Da mesma forma, nas Tabelas 10 e 11 são mostrados as regras de $\mathbf{h}_2$ ordenadas pelas medidas genéricas e relativas, respectivamente.

5.1.1 Considerando o Conjunto de Regras $\mathbf{h}_1$

De um total de 14 diferentes medidas calculadas pelo MAR, sob o conjunto de regras $\mathbf{h}_1$, 3 dessas medidas (`accR`, `specR` e `satR`) apresentaram valor 1 para todas as regras. Assim, as 9 medidas de avaliação restantes foram sintetizadas nas Tabelas 5 e 7. Por sua vez, essas 9 medidas foram utilizadas para elencar as regras de $\mathbf{h}_1$ pela ordem decrescente dos valores calculados — Tabela 9. Deve ser observado que a medida `errR` é complementar à medida `accR` e por isso não foi considerada nessa tabela. Ainda deve ser observado que `novR`=`wraccR` e que, para esse conjunto de regras `covR`=`supR`.

Observando a Tabela 9 é possível verificar que as regras 1 e 6 apresentaram "melhor"valor para 6 (seis) das 9 (nove) medidas consideradas. Dessa forma, se o especialista não considerar que a precisão na predição de exemplos não cobertos pela regra (ou seja, `negrelR`) não é um critério importante, pode-se dizer que as regras regras 1 e 6 são as que apresentam as melhores medidas de avaliação. Analogamente, as regras $2, 4, 24$ e 23 são as próximas regras a apresentar boas medidas de avaliação. Isso dá uma indicação ao especialista para analisar mais cuidadosamente o subconjunto composto pelas regras $\{1, 6, 2, 4, 24$ e $23\}$. Esse subconjunto pode ser referenciado como as "melhores"regras de $\mathbf{h}_1$.

Ordem	Medidas Genéricas						
	negrelR	sensR	covR[a]	novR[b]	rnegrelR	rsensR	rspecR
1	13	1	1	1	1	1	1
2	17,12,	6	6	6	6	6	6
3	14	24	2	2	2	24	2
4	16,15	2	4	4	4	23	4
5	20,19,18	4	24	24	24	2	24
6	24	23	23,10,	23	23	13	23, 10
6	23	21	21,3	21	21	21	21,3
7	21	22	22,5	22	22	4	22,5,
8	22	26,25	26,25	10	10	22	26
9	26,25	10	32	26,25	3	26,25	32
10	32	13	13	3	26	17,12	13
11	31	3	7	5	25	10	7
12	29	5	17,12,8	13	5	3	17,12,8
13	27	17	31	32	13	5	31
14	1	12	29	17,12	32	32	29
15	28	32	27,9	7	17,12	31	9,27
16	30	7	28,14	8	7	7	28,14
17	6	31	16,15	31	8	14	16,15
18	2	8	30,20,19,18,11	29	31	8	30,20,19,18,11
19	4	29		27	29	29	
20	10	14		14	27	27	
21	3	27		28	14	16,15	
22	5	9		16,15	9	28	
23	7	16,15		20,19,18	28	9	
24	8	28		30,11	16,15	20,19,18	
25	9	20,19,18			20,19,18	30	
26	11	30			30	11	
27		11			11		

[a] covR = supR.
[b] novR = wraccR.

Tabela 9: Regras de $\mathbf{h}_1$ Ordenadas pelos Valores da Medida de Avaliação

Inicialmente, um aspecto que chama a atenção ao ser observado, refere-se às classes preditas por cada regra do subconjunto das "melhores"regras de $\mathbf{h}_1$. Observando o Apêndice C verifica-se que as regras $1, 2, 4$ e 6 predizem a classe `um`, enquanto que as regras 23 e 24 são da classe `tres`. Conseqüentemente, não se destacam regras que predizem a classe `dois`. É interessante observar que a classe dois é a classe minoritária do conjunto de exemplos utilizados, 22.08% — Tabela 2 — a classe `um` é a classe majoritária com 42.92% dos exemplos, enquanto que a classe `tres` possui 35.00% dos exemplos.

Outro aspecto que pode ser considerado refere-se aos atributos que compõem cada uma das regras desse subconjunto de $\mathbf{h}_1$. Observando novamente o Apêndice C é possível verificar que dos 21 atributos (exceto o atributo classe) do conjunto de dados utilizado apenas 10 estão presentes nos corpos das "melhores"regras de $\mathbf{h}_1$. Esses atributos são `concentracao`, `vitalidade`, `hp`, `class_b`, `idade`, `motilidade`, `class_a`, `class_c`, `class_d` e `concentracao_total`.

Finalmente, um outro aspecto que pode ser levado em conta refere-se a análise dos valores hb, $\overline{hb}$ de cada regra, ou seja, a análise dos valores referentes à cobertura da regra — Tabela 5, por exemplo. Como a precisão de todas as regras de $\mathbf{h}_1$ é igual a 1, o valor $\overline{hb}$ é igual a 0, ou seja, todas as regras de $\mathbf{h}_1$ não cobrem exemplos de uma classe diferente da predita pela regra. Considerando as regras $\{1, 6, 2, 4, 24$ e $23\}$ é possível observar que o valor hb da regra 1 é maior que o hb da regra 6, e assim por diante. Dessa forma, a regra 1 cobre corretamente mais exemplos que a regra 6, que por sua vez, cobre mais exemplos que a regra 2 e assim segue.

5.1.2 Considerando o Conjunto de Regras $\mathbf{h}_2$

A partir das 14 diferentes medidas calculadas pelo MAR, apresentadas nas Tabelas 6 e 8, as regras $\mathbf{h}_2$ foram dispostas em ordem decrescente dos valores calculados para as medidas genéricas e relativas — Tabelas 10 e 11, respectivamente. Deve ser lembrado que a medida `errR` é complementar a `accR` e que `novR = wraccR`.

Dentre as 12 (doze) diferentes medidas consideradas, 7 (sete) apontam as regras 19 e 5 como sendo "melhores". As próximas melhores regras são $9, 10, 16$ e 12, apontadas por 4 (quatro) medidas não consideradas entre as sete anteriores. Assim, o subconjunto das "melhores"regras de $\mathbf{h}_2$ é composto pelas regras $\{19, 5, 9, 10, 16$ e $12\}$

Do subconjunto das "melhores"de regras de $\mathbf{h}_2$, con relação às classes preditas por cada regra, a regra 5 prediz a classe `um`, as regras $9, 10, 12$ e 16 predizem a classe `dois`, enquanto que a regra 19 prediz a classe `tres` — Apêndice D.

Com relação aos atributos que compõem cada regra, pode ser observado no Apêndice D que em todas regras o atributo `concentracao_total` está presente. Dos 21 atributos (exceto o atributo classe) do conjunto de dados utilizado, novamente apenas 10 estão presentes nos corpos das "melhores"regras de $\mathbf{h}_2$. Esses atributos são `concentracao_total`, `motilidade`, `hp`, `hora`, `cor`, `liquefação`, `idade`, `volume`, `class_a` e `class_c` .

Considerando agora os valores hb, $\overline{hb}$, mostrados na Tabela 6 página 15, de cada regra do subconjunto composto pelas "melhores"regras de $\mathbf{h}_2$ é possível observar que o valor hb e $\overline{hb}$ da regra 19 é maior que os valores hb e $\overline{hb}$ da regra 5. É interessante observar por esses valores que a regra 19, mesmo cobrindo mais exemplos de classes diferentes da sua ($\overline{hb}$ maior) foi apontada

Medidas Genéricas							
accR	negrelR	sensR	specR	covR	supR	novR[a]	satR
17,16,12,10,9,6,3,2	19	19	17,16,12,10,9,6,3,2	19	19	19	17,16,12,10,9,6,3,2
1	9	5	15,14,13,11	5	5	5	1
4	11	18	7	1	1	1	11
7	10	1	8,4	18	18	18	4
18,11	13	17	1	4	4	17	7
5	15,14	4	18	17	17	4	18
19	16	8	5	8	8	8	13
8	12	11,9	19	2	2	2	5
13	18	2		3	3	3	19
15,14	5	10		7	7	9	15
	17	3		11	11,9	11	14
	1	13		9	10	10	8
	4	7		13,1	13,6	7	
	8	15, 14		15,14,6	15,14	13	
	2	6		16	16	6	
	3	16		12	12	15,14	
	7	12				16	
	6					12	

[a] novR = wraccR.

Tabela 10: Regras de $\mathbf{h}_2$ Ordenadas pelos Valores das Medidas Genéricas

Medidas Relativas			
raccR	rnegrelR	rsensR	rspecR
16,12,10,9	19	19	19
17	5	5	5
11	1	18	1
13	18	1	18
6,3,2	17	17	4
15,14	4	4	17
18	8	9	8
1	2	11	2
19	3	10	3
4	9	8	7
7	11	2	9
5	10	13	11
8	7	3	10
	13	15,14	6
	6	7	13
	15,14	16	15,14
	16	6	16
	12	12	12

Tabela 11: Regras de $\mathbf{h}_2$ Ordenadas pelos Valores das Medidas Relativas.

como "melhor"que a regra 5, i.e. apresentou "melhores"valores para as medidas de avaliação calculadas.

Com relação às regras restantes $\{9, 10, 16$ e $12\}$, é possível observar que o valor hb da regra 9 é maior que o hb da regra 10 e assim por diante. É interessante observar que o valor de hb dessas regras é muito menor se comparados com os valores de hb das regras 19 e 5 mas, sob o ponto de vista da medida de precisão `accR`, as $9, 10, 16$ e 12 são "melhores"que as regras 19 e 5.

5.2 Da Análise Quantitativa

Num processo de análise quantitativa, quanto mais alto o valor de uma medida para uma determinada regra, maior é a chance de que o conhecimento descrito por ela seja um conhecimento interessante ou inesperado. Considerando os valores calculados pelo MAR, as regras mais interessantes de $\mathbf{h}_1$ e $\mathbf{h}_2$ são apresentadas na Tabela 12 na ordem decrescente dos valores calculados para cada medida objetiva de interessabilidade.

$\mathbf{h}_1$		$\mathbf{h}_2$	
`SurpDisj`	`SurpAttR`	`SurpDisj`	`SurpAttR`
20,19,18,15	32	12	3,4,7,9,10
16		16	12
30			16

Tabela 12: Regras de $\mathbf{h}_1$ e $\mathbf{h}_2$ Ordenadas pelos Valores da Medidas de Interessabilidade

5.2.1 Considerando o Conjunto de Regras $\mathbf{h}_1$

De um total de trinta e duas (32) regras de $\mathbf{h}_1$, foram selecionadas pelo MAR, através da medida de Surpresa de Pequenos Disjuntos, as regras $20, 19, 18, 15, 16$ e 30 (Tabela 12). Pela medida de Surpresa dos Atributos Individuais Discretos foram selecionada apenas a regra 32 (Tabela 12).

Pode ser observado — Apêndice C — que desse subconjunto de regras apontadas como "mais interessantes", as regras $20, 19, 18, 15$ e 16 predizem a classe `dois` (classe minoritária), que as regras 30 e 32 predizem a classe `três` e, portanto, nenhuma dessas regras predizem a classe `um` (classe majoritária).

Dos 21 atributos (exceto o atributo classe) do conjunto de dados utilizado, apenas 11 estão presentes nos corpos das "mais interessantes"regras de $\mathbf{h}_1$. Esses atributos são `hora`, `idade`, `concentracao`, `class_c`, `vitalidade`, `hp`, `volume`, `class_d`, `processamento` e `tempo_abs`. Outro aspecto a ser observado é que os corpos das regras são compostos por menos de 4 diferentes atributos.

Considerando agora os valores hb, $\overline{hb}$, mostrados na Tabela 5 página 14, é interessante observar que todas as regra tem $\overline{hb} = 0$ e ainda, que dentre essas regras, a regra 32 chama atenção pois trata-se da única regra de $\mathbf{h}_1$ que possui em seu corpo atributos discretos e ainda, o número de exemplos corretamente cobertos (hb) por essa regra é maior que as demais regras selecionadas como "mais interessantes".

5.2.2 Considerando o Conjunto de Regras $\mathbf{h}_2$

Do conjunto de regras $\mathbf{h}_2$ com dezenove (19) regras, foram selecionadas através do cálculo da medida de Surpresa de Pequenos Disjuntos efetuado pelo MAR, as regras 12 e 16 (Tabela 12). Através da medida de Surpresa dos Atributos Individuais Discretos, foram selecionadas as regras $3, 4, 7, 9, 10, 12$ e 16 (Tabela 12).

Desse subconjunto de regras "mais interessantes"composto pelas regras $\{3, 4, 7, 9, 10, 12$ e $16\}$, pode ser observado — Apêndice D — que as regras $3, 4$ e 7 predizem a classe `um`(classe majoritária) e que as regras $10, 12$ e 16 predizem a classe `dois` e, portanto, nenhuma regra selecionada prediz a classe `tres`.

Pode ser facilmente constatado que as regras 12 e 16 foram selecionadas pelas duas medidas. Espera-se que essas regras tenham, potencialmente, mais interesse para o especialista do que as outras.

Dos 21 atributos (exceto o atributo classe) do conjunto de dados utilizado, novamente apenas 11 estão presentes nos corpos das "mais interessantes"regras de $\mathbf{h}_2$. Esses atributos são `hora`, `idade`, `cor`, `concentracao_total`, `motilidade`, `class_c`, `vitalidade`, `hp`, `volume`, `class_a` e `liquefação`. Deve ser observado que as regras desse subconjunto de $\mathbf{h}_2$ tem corpos compostos por mais de 4 atributos e ainda, que o atributo `cor` está presente em todas essas regras —- Apêndice D.

Finalmente, se forem comparados os subconjuntos de regras selecionadas pelas medidas qualitativas com o subconjunto de regras selecionadas pelas medidas quantitativas,considerando valores hb, $\overline{hb}$ de cada regra — por exemplo na Tabela 6 página 15 — pode ser observado que as regras $9, 10, 12$ e 16 estão presentes em ambos subconjuntos de regras de $\mathbf{h}_2$. Dessa forma, é curioso observar que essas quatro (4) regras cobrem poucos exemplos mas, ao mesmo tempo, cobrem corretamente esses exemplos. Ainda deve ser observado que as regras 12 e 16 apresentam o valor hb menor que as outras regras selecionadas de $\mathbf{h}_2$.

6 Considerações Finais

No estudo apresentado neste trabalho, foi sucintamente descrita a importância do processamento de sêmen diagnóstico no tratamento para a reprodução assistida e também foi rapidamente descrito o Módulo de Análise de Regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Foram apresentadas as etapas envolvidas nesse estudo e ainda foram discutidos os resultados obtidos.

Vale ressaltar que inexiste uma metodologia estabelecida para a discussão de resultados obtidos através da utilização de medidas de avaliação e interessabilidade de regras. Neste trabalho optou-se por elencar as regras de acordo com o maior valor das medidas aqui consideradas e então observar quais regras que se destacam.

Entretanto, um dos aspectos positivos do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$ é que por ter sido implementado na linguagem de programação Prolog, após realizado os cálculos das melhores medidas implementadas no MAR, o usuário do sistema pode realizar quaisquer tipo de interrogações *on-line* tais como:

Quais as regras que possuem máximo valor de `accR` *e cobrem pelo menos* $X\%$ *do número de exemplos?*
Quais as regras com máximo valor para as medidas `covR` *e* `novR`*?*
Quais as regras comuns selecionadas pelas medidas `SurpDisjR` *e* `SurpAtrR`*?*

Neste trabalho os resultados foram discutidos levando em conta aspectos considerados relevantes sob o ponto de vista da análise objetiva dos resultados. Na análise subjetiva, ou seja, a análise dos resultados feita por um especialista, podem ser utilizados ou não os resultados da análise objetiva. A idéia envolvida neste estudo de caso é que, mediante o uso das medidas de avaliação e de interessabilidade, o especialista possa realmente focalizar sua atenção sob aquelas regras que mais se destacam, evitando assim que esse especialista examine uma a uma cada regra do conjunto de regras que constituem a(s) hipótese(s). Nesse contexto, a participação do especialista do domínio é fundamental que pode, inclusive discordar dos resultados apresentados para algumas regras.

A Conjunto de Regras Induzidas pelo Algoritmo $\mathcal{CN}2$ — $\mathbf{h}_1$.

```
**RULE FILE** @ Time: [ Sun Jan 13 14:48:50 2002 ]
Examples: proc.exs
Algorithm: UNORDERED
Error_Estimate: LAPLACIAN
Threshold: 0.00 Star: 5 @

*UNORDERED-RULE-LIST*

IF    concentracao < 66.70  AND vitalidade < 49.50  AND HP < 70.00
THEN  class = um  [29.50 0 0]

IF    idade < 46.00  AND pH > 7.70  AND motilidade < 5.50
  AND class_A > 10.00  AND HP < 58.50
THEN  class = um  [23.50 0 0]

IF    processamento < 35.00  AND tempo_abs < 6.00
  AND concentracao_total < 18.55 THEN  class = um  [17 0 0]

IF    pH < 8.15  AND concentracao < 72.45  AND class_B > 6.00
  AND class_C > 57.50
THEN  class = um  [21.50 0 0]

IF    tempo_abs < 4.50  AND 49.05 < concentracao_total < 76.20
THEN  class = um  [16 0 0]

IF    concentracao < 13.15  AND class_B > 7.00  AND HP < 70.00
THEN  class = um  [29 0 0]

IF    class_B > 34.00  AND HP < 77.50
THEN  class = um  [7.50 0 0]

IF    concentracao > 26.85  AND class_A > 30.00
  AND vitalidade < 53.00
THEN  class = um  [6.50 0 0]

IF    idade < 26.50 AND motilidade < 8.50
THEN class = um [4 0 0]

IF    class_C > 70.00
THEN  class = um  [18 0 0]

IF    idade < 30.50  AND volume > 7.30
THEN  class = um  [1 0 0]

IF    concentracao > 21.35  AND concentracao_total < 344.90
  AND num_leu > 5.74
THEN  class = dois  [0 7 0]
```

```
IF    volume < 1.75  AND concentracao < 94.00  AND class_A > 18.00
  AND Kruger < 10.00
THEN  class = dois  [0 8.50 0]

IF    idade > 49.00  AND volume > 3.10
THEN  class = dois  [0 3 0]

IF    idade < 25.50  AND volume < 2.05
THEN  class = dois  [0 2 0]

IF    idade < 34.00  AND class_D > 47.50  AND vitalidade > 91.50
THEN  class = dois  [0 1.50 0]

IF    idade > 27.50  AND concentracao > 89.90
  AND concentracao_total < 322.40  AND motilidade < 16.50
  AND num_leu < 1.35
THEN  class = dois  [0 7 0]

IF    processamento > 55.00  AND tempo_abs > 3.50
THEN  class = dois [0 1 0]

IF    idade > 34.00  AND class_C < 9.00
THEN  class = dois  [0 1 0]

IF    idade < 33.50  AND concentracao > 593.40
THEN  class = dois  [0 1 0]

IF    idade > 25.50  AND concentracao_total > 10.35
  AND motilidade > 41.00
THEN  class = tres  [0 0 17]

IF    concentracao_total > 339.40  AND Kruger > 4.50
THEN  class = tres  [0 0 16]

IF    129.27 < concentracao_total < 336.80  AND class_C < 18.50
THEN  class = tres  [0 0 18]

IF    concentracao_total > 305.10  AND class_C > 14.50
  AND class_D > 45.50
THEN  class = tres  [0 0 20]

IF    idade > 32.50  AND motilidade > 10.00
  AND Kruger > 14.00  AND HP > 53.50
THEN  class = tres  [0 0 14.50]

IF    24.65 < concentracao < 73.80  AND motilidade > 28.50
THEN  class = tres  [0 0 15]

IF    class_A > 63.50  AND 84.50 < vitalidade < 87.00
THEN  class = tres  [0 0 4]
```

```
IF    concentracao_total > 339.40  AND class_B > 16.00
  AND num_leu > 0.12
THEN  class = tres  [0 0 3]

IF    tempo_abs > 7.50  AND volume > 3.05
THEN  class = tres  [0 0 5]

IF    tempo_abs < 1.50  AND volume < 1.00
THEN  class = tres  [0 0 1]

IF    concentracao_total > 227.00  AND HP > 85.50
THEN  class = tres  [0 0 6]

IF    hora = manha  AND class_D > 61.00  AND HP > 83.50
THEN  class = tres  [0 0 10]

(DEFAULT) class = um  [103 53 84]

Executing rules...
        PREDICTED
ACTUAL     um      dois    tres    Accuracy
    um     103      0       0     100.0 %
  dois     21      32       0      60.4 %
  tres     10       0      74      88.1 %
Overall accuracy:  87.1 %
Default accuracy:  42,9 %
```

B Conjunto de Regras Induzidas pelo Algoritmo $\mathcal{S}$ee5 — h_2.

```
See5 INDUCTION SYSTEM [Release 1.09a]   Mon Jan 14 21:04:39 2002
-------------------------------------

    Options: Generating rules

Class specified by attribute class

Read 240 cases (22 attributes) from proc.see5.data

Extracted rules:

Rule 1: (cover 40)
        motilidade <= 15
        HP <= 56
    ->  class um  [0.929]

Rule 2: (cover 10)
        idade <= 28
        motilidade <= 15
        class_C > 18
    ->  class um  [0.917]

Rule 3: (cover 9)
        idade > 30
        idade <= 33
        cor = branco
        concentracao_total > 95.2
        concentracao_total <= 338.8
        motilidade <= 15
        class_C > 18
        vitalidade <= 79
        HP <= 83
    ->  class um  [0.909]

Rule 4: (cover 20)
        volume > 2.4
        cor = branco
        concentracao_total <= 95.2
        class_A <= 33
        HP <= 88
    ->  class um  [0.864]

Rule 5: (cover 73)
        concentracao_total <= 95.2
        motilidade <= 17
    ->  class um  [0.840]
```

```
Rule 6: (cover 4)
        concentracao_total <= 95.2
        motilidade > 17
        class_A > 33
    ->  class um  [0.833]

Rule 7: (cover 8)
        idade > 35
        idade <= 36
        cor = branco
        motilidade <= 15
    ->  class um  [0.800]

Rule 8: (cover 16)
        idade > 30
        idade <= 36
        volume <= 2.1
        motilidade <= 15
    ->  class um  [0.778]

Rule 9: (cover 5)
        volume <= 2.4
        cor = branco
        concentracao_total <= 95.2
        motilidade > 17
        class_A <= 33
        HP <= 88
    ->  class dois  [0.857]

Rule 10: (cover 5)
        idade > 33
        idade <= 35
        volume > 2.1
        cor = branco
        concentracao_total > 95.2
        concentracao_total <= 338.8
        motilidade <= 15
        class_C > 18
        HP > 56
    ->  class dois  [0.857]

Rule 11: (cover 7)
        idade > 28
        idade <= 30
        tempo_abs <= 7
        concentracao_total > 95.2
        motilidade <= 15
        class_C > 18
    ->  class dois  [0.778]
```

```
Rule 12: (cover 6)
        hora = manha
        cor = branco
        liquefacao = completa
        concentracao_total <= 46.2
        HP > 62
    ->  class dois  [0.750]

Rule 13: (cover 5)
        motilidade > 15
        num_leu > 4
    ->  class dois  [0.714]

Rule 14: (cover 4)
        idade > 36
        concentracao_total > 95.2
        motilidade <= 15
        class_B > 23
    ->  class dois  [0.667]

Rule 15: (cover 6)
        volume <= 1.4
        concentracao_total > 95.2
        motilidade <= 15
    ->  class dois  [0.625]

Rule 16: (cover 10)
        hora = manha
        cor = branco
        concentracao_total <= 46.2
        HP > 62
    ->  class dois  [0.583]

Rule 17: (cover 18)
        idade <= 36
        class_C <= 18
    ->  class tres  [0.950]

Rule 18: (cover 58)
        motilidade > 17
        class_A <= 33
    ->  class tres  [0.617]

Rule 19: (cover 128)
        concentracao_total > 95.2
        HP > 56
    ->  class tres  [0.562]

Default class: um
```

```
Evaluation on training data (240 cases):

        Decision Tree           Rules
      ----------------    ----------------
      Size      Errors      No      Errors

        35   26(10.8%)      19   37(15.4%)   <<


       (a)   (b)   (c)  <-classified as
      ----  ----  ----
        99     1     3  (a): class um
         7    25    21  (b): class dois
         2     3    79  (c): class tres


Time: 0.3 secs
```

C Conjunto de Regras h_1 na Sintaxe Padrão Prolog

```
rule(id_105,0001,[less(concentracao,66.70),less(vitalidade,49.50),less(hp,70.00),
class(um),[0.1200,0.0000,0.5733,0.3067,225],[0.3333,0.000000,0.5333,0.1333,15]).

rule(id_105,0002,[less(idade,46.00),greater(ph,7.70),less(motilidade,5.50),
greater(class_a,10.00),less(hp,58.50)],class(um),
[0.1022,0.0000,0.5733,0.3244,225],[0.0667,0.000000,0.5333,0.4000,15]).

rule(id_105,0003,[less(processamento,35.00),less(tempo_abs,6.00),
less(concentracao_total,18.55)],class(um),
[0.0708,0.0000,0.5708,0.3583,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0004,[less(ph,8.15),less(concentracao,72.45),greater(class_b,6.00),
greater(class_c,57.50)],class(um),
[0.0855,0.0000,0.5812,0.3333,234],[0.5000,0.000000,0.1667,0.3333,6]).

rule(id_105,0005,[less(tempo_abs,4.50),greater(concentracao_total,49.05),
less(concentracao_total,76.20)],class(um),
[0.0667,0.0000,0.5708,0.3625,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0006,[less(concentracao,13.15),greater(class_b,7.00),less(hp,70.00),
class(um),[0.1207,0.0000,0.5690,0.3103,232],[0.2500,0.000000,0.6250,0.1250,8]).

rule(id_105,0007,[greater(class_b,34.00),less(hp,77.50)],class(um),
[0.0302,0.0000,0.5690,0.4009,232],[0.1250,0.000000,0.6250,0.2500,8]).

rule(id_105,0008,[greater(concentracao,26.85),greater(class_a,30.00),
less(vitalidade,53.00)],class(um),
[0.0260,0.0000,0.5801,0.3939,231],[0.1111,0.000000,0.3333,0.5556,9]).

rule(id_105,0009,[less(idade,26.50),less(motilidade,8.50)],class(um),
[0.0168,0.0000,0.5672,0.4160,238],[0.0000,0.000000,1.0000,0.0000,2]).

rule(id_105,0010,[greater(class_c,70.00)],class(um),
[0.0750,0.0000,0.5708,0.3542,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0011,[less(idade,30.50),greater(volume,7.30)],class(um),
[0.0042,0.0000,0.5672,0.4286,238],[0.0000,0.000000,1.0000,0.0000,2]).

rule(id_105,0012,[greater(concentracao,21.35),less(concentracao_total,344.90),
greater(num_leu,5.74)],class(dois),
[0.0293,0.0000,0.7782,0.1925,239],[0.0000,0.000000,1.0000,0.0000,1]).

rule(id_105,0013,[less(volume,1.75),less(concentracao,94.00),
greater(class_a,18.00),less(kruger,10.00)],class(dois),
[0.0343,0.0000,0.7768,0.1888,233],[0.1429,0.000000,0.8571,0.0000,7]).

rule(id_105,0014,[greater(idade,49.00),greater(volume,3.10)],class(dois),
[0.0126,0.0000,0.7773,0.2101,238],[0.0000,0.000000,1.0000,0.0000,2]).
```

```
rule(id_105,0015,[less(idade,25.50),less(volume,2.05)],class(dois),
[0.0084,0.0000,0.7773,0.2143,238],[0.0000,0.000000,1.0000,0.0000,2]).

rule(id_105,0016,[less(idade,34.00),greater(class_d,47.50),
greater(vitalidade,91.50)],class(dois),
[0.0044,0.0000,0.7773,0.2183,229],[0.0909,0.000000,0.8182,0.0909,11]).

rule(id_105,0017,[greater(idade,27.50),greater(concentracao,89.90),
less(concentracao_total,322.40),less(motilidade,16.50),less(num_leu,1.35)],
class(dois),[0.0295,0.0000,0.7764,0.1941,237],[0.0000,0.000000,1.0000,0.0000,3]).

rule(id_105,0018,[greater(processamento,55.00),greater(tempo_abs,3.50)],
class(dois),[0.0042,0.0000,0.7792,0.2167,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0019,[greater(idade,34.00),less(class_c,9.00)],class(dois),
[0.0042,0.0000,0.7773,0.2185,238],[0.0000,0.000000,1.0000,0.0000,2]).

rule(id_105,0020,[less(idade,33.50),greater(concentracao,593.40)],class(dois),
[0.0042,0.0000,0.7773,0.2185,238],[0.0000,0.000000,1.0000,0.0000,2]).

rule(id_105,0021,[greater(idade,25.50),greater(concentracao_total,10.35),
greater(motilidade,41.00)],class(tres),
[0.0714,0.0000,0.6555,0.2731,238],[0.0000,0.000000,0.0000,1.0000,2]).

rule(id_105,0022,[greater(concentracao_total,339.40),greater(kruger,4.50)],
class(tres),[0.0687,0.0000,0.6438,0.2876,233],[0.0000,0.000000,0.8571,0.1429,7]).

rule(id_105,0023,[greater(concentracao_total,129.27),less(concentracao_total,336.80),
less(class_c,18.50)],class(tres),
[0.0750,0.0000,0.6500,0.2750,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0024,[greater(concentracao_total,305.10),greater(class_c,14.50),
greater(class_d,45.50)],class(tres),
[0.0833,0.0000,0.6500,0.2667,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0025,[greater(idade,32.50),greater(motilidade,10.00),
greater(kruger,14.00),greater(hp,53.50)],class(tres),
[0.0625,0.0000,0.6429,0.2946,224],[0.0625,0.000000,0.7500,0.1875,16]).

rule(id_105,0026,[greater(concentracao,24.65),less(concentracao,73.80),
greater(motilidade,28.50)],class(tres),
[0.0625,0.0000,0.6500,0.2875,240],[0.000,0.000,0.000,0.000,0]).


rule(id_105,0027,[greater(class_a,63.50),greater(vitalidade,84.50),
less(vitalidade,87.00)],class(tres),
[0.0173,0.0000,0.6407,0.3420,231],[0.0000,0.000000,0.8889,0.1111,9]).

rule(id_105,0028,[greater(concentracao_total,339.40),greater(class_b,16.00),
greater(num_leu,0.12)],class(tres),
[0.0126,0.0000,0.6485,0.3389,239],[0.0000,0.000000,1.0000,0.0000,1]).
```

```
rule(id_105,0029,[greater(tempo_abs,7.50),greater(volume,3.05)],class(tres),
[0.0208,0.0000,0.6500,0.3292,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0030,[less(tempo_abs,1.50),less(volume,1.00)],class(tres),
[0.0042,0.0000,0.6500,0.3458,240],[0.000,0.000,0.000,0.000,0]).

rule(id_105,0031,[greater(concentracao_total,227.00),greater(hp,85.50)],
class(tres),
[0.0259,0.0000,0.6422,0.3319,232],[0.0000,0.000000,0.8750,0.1250,8]).

rule(id_105,0032,[equal(hora,manha),greater(class_d,61.00),
greater(hp,83.50)],class(tres),
[0.0431,0.0000,0.6422,0.3147,232],[0.0000,0.000000,0.8750,0.1250,8]).

rule(id_105,0033,[],class(um),[],[]).
```

D Conjunto de Regras h_2 na Sintaxe Padrão Prolog

```
rule(id_104,0001,[lessORequal(motilidade,15),lessORequal(hp,56)],class(um),
[0.1638,0.0086,0.5603,0.2672,232],[0.3750,0.2500,0.3750,0.0000,8]).

rule(id_104,0002,[lessORequal(idade,28),lessORequal(motilidade,15),
greater(class_c,18)],class(um),
[0.0420,0.0000,0.5672,0.3907,238],[0.0000,0.0000,1.000,0.0000,2]).

rule(id_104,0003,[greater(idade,30),lessORequal(idade,33),equal(cor,branco),
greater(concentracao_total,95.2),lessORequal(concentracao_total,338.8),
lessORequal(motilidade,15),greater(class_c,18),lessORequal(vitalidade,79),
lessORequal(hp,83)],class(um),
[0.0409,0.0000,0.5772,0.3818,220],[0.0000,0.0000,0.5000,0.5000,20]).

rule(id_104,0004,[greater(volume,2.4),equal(cor,branco),
lessORequal(concentracao_total,95.2),lessORequal(class_a,33),
lessORequal(hp,88)],class(um),
[0.0786,0.0087,0.5677,0.3450,229],[0.3636,0.0909,0.3636,0.1818,11]).

rule(id_104,0005,[lessORequal(concentracao_total,95.2),
lessORequal(motilidade,17)],class(um),
[0.2583,0.0458,0.525,0.1708,240],[0.0000,0.0000,0.0000,0.0000,0]).

rule(id_104,0006,[lessORequal(concentracao_total,95.2),greater(motilidade,17),
greater(class_a,33)],class(um),
[0.01667,0.0000,0.5708,0.4125,240],[0.0000,0.0000,0.0000,0.0000,0]).

rule(id_104,0007,[greater(idade,35),lessORequal(idade,36),equal(cor,branco),
lessORequal(motilidade,15)],class(um),
[0.0299,0.0043,0.5726,0.3931,234],[0.0000,0.0000,0.3333,0.6667,6]).

rule(id_104,0008,[greater(idade,30),lessORequal(idade,36),lessORequal(volume,2.1),
lessORequal(motilidade,15)],class(um),
[0.0546,0.0126,0.5546,0.3781,238],[0.0000,0.0000,1.0000,0.0000,2]).

rule(id_104,0009,[lessORequal(volume,2.4),equal(cor,branco),
lessORequal(concentracao_total,95.2),greater(motilidade,17),
lessORequal(class_a,33),lessORequal(hp,88)],class(dois),
[0.0218,0.0000,0.7860,0.1921,229],[0.0909,0.0000,0.6363,0.2727,11]).

rule(id_104,0010,[greater(idade,33),lessORequal(idade,35),
greater(volume,2.1),equal(cor,branco),greater(concentracao_total,95.2),
lessORequal(concentracao_total,338.8),lessORequal(motilidade,15),
greater(class_c,18),greater(hp,56)],class(dois),
[0.0220,0.0000,0.7841,0.1938,227],[0.0000,0.0000,0.6923,0.3077,13]).

rule(id_104,0011,[greater(idade,28),lessORequal(idade,30),
lessORequal(tempo_abs,7),greater(concentracao_total,95.2),
lessORequal(motilidade,15),greater(class_c,18)],class(dois),
```

```
[0.0252,0.0042,0.7731,0.1974,238],[0.0000,0.0000,1.0000,0.0000,2]).

rule(id_104,0012,[equal(hora,manha),equal(cor,branco),equal(liquefacao,completa),
lessORequal(concentracao_total,46.2),greater(hp,62)],class(dois),
[0.0043,0.0000,0.7860,0.2096,229],[0.0000,0.0000,0.6363,0.3636,11]).

rule(id_104,0013,[greater(motilidade,15),greater(num_leu,4)],class(dois),
[0.0167,0.0042,0.7741,0.2050,239],
[0.0000,0.0000,1.0000,0.0000,1]).

rule(id_104,0014,[greater(idade,36),greater(concentracao_total,95.2),
lessORequal(motilidade,15),greater(class_b,23)],class(dois),
[0.0126,0.4202,0.7731,0.2101,238],[0.0000,0.0000,1.0000,0.0000,2]).

rule(id_104,0015,[lessORequal(volume,1.4),greater(concentracao_total,95.2),
lessORequal(motilidade,15)],class(dois),[0.0125,0.0041,0.7750,0.2083,240],
[0.0000,0.0000,0.0000,0.0000,0]).

rule(id_104,0016,[equal(hora,manha),equal(cor,branco),
lessORequal(concentracao_total,46.2),greater(hp,62)],class(dois),
[0.0087,0.0000,0.7860,0.2052,229],[0.0000,0.0000,0.6363,0.3636,11]).

rule(id_104,0017,[lessORequal(idade,36),lessORequal(class_c,18)],class(tres),
[0.0714,0.0000,0.6554,0.2731,238],[0.5000,0.0000,0.0000,0.5000,2]).

rule(id_104,0018,[greater(motilidade,17),lessORequal(class_a,33)],class(tres),
[0.1500,0.0250,0.6250,0.200,240],[0.0000,0.0000,0.0000,0.0000,0]).

rule(id_104,0019,[greater(concentracao_total,95.2),greater(hp,56)],class(tres),
[0.3017,0.0689,0.5732,0.0560,232],[0.1250,0.0000,0.8750,0.0000,8]).

rule(id_104,0020,[],class(um),[],[]).
```

Referências

Baranauskas, J. A. (2001). Extração automática de conhecimento por múltiplos indutores. Tese de Doutorado, ICMC-USP.

Baranauskas, J. A. and Batista, G. E. A. P. A. (2001). O projeto Discover: Idéias iniciais. (Comunicação Pessoal).

Batista, G. E. A. P. A. (1997). Um ambiente de avaliação de algoritmos de aprendizado de máquina utilizando exemplos. Dissertação de Mestrado, ICMC-USP.

Batista, G. E. A. P. A. (2000). Pré-processamento de dados em aprendizado de máquina supervisionado. Monografia do Exame de Qualificação de Doutorado, ICMC-USP.

Batista, G. E. A. P. A. (2001). Sintaxe padrão do arquivo de exemplos do projeto Discover. http://www.icmc.sc.usp.br/ gbatista/SintaxePadraoFinal.htm.

Bernardini, F. C. (2001). Combinação de classificadores para melhorar o poder preditivo e descritivo de ensembles. Monografia do Exame de Qualificação de Mestrado, ICMC-USP.

Bernardini, F. C. and Monard, M. C. (2002a). Projeto e implementação do módulo de combinação e explicação de classificadores simbólicos do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Technical report, ICMC-USP. Em elaboração.

Bernardini, F. C. and Monard, M. C. (2002b). Um estudo de caso utilizando o módulo de explicação de ensembles e combinação de regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Technical report, ICMC-USP. Em elaboração.

Caulkins, C. W. (2000). Aquisição de conhecimento utilizando aprendizado de máquina relacional. Dissertação de Mestrado, ICMC-USP.

Clark, P. and Niblett, T. (1989). The $\mathcal{CN}2$ induction algorithm. *Machine Learning*, 3(4):261–283.

Esteves, S. and Bento, F. C. (1998). Sperm processing for assisted reproductive technology (art): effects of pentoxifyline on recovery of motile sperm in asthenozoospermic men. In *FertilSteril*, page 5196.

Esteves, S., Lee, H., Monard, M., and Lopes, L. (2001). Inteligência artificial aplicada à andrologia: um estudo de caso do processamento de sêmen diagnóstico. In *XXVIII Congresso Brasileiro de Urologia*, Fortaleza, CE, Brasil.

Esteves, S. C., Sharma, R. K., Thomas, A. J. J., and Agarwal, A. (2000). Improvement in motion characteristics and acrosome status in cryopreserved human spermatozoa by swim-up processing before freezing. In *HumReprod*, pages 2173–2179.

Félix, L. C. M. (1998). Data mining no processo de extração de conhecimento de bases de dados. Dissertação de Mestrado, ICMC-USP.

Ferro, M., Lee, H. D., and Esteves, S. C. (2002). Intelligent data analysis: A case study of the diagnostic sperm processing. In *International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CSITeA-2002)*. (submetido).

Freitas, A. A. (1998a). A multi-criteria approach for the evaluation of rule interestingness. In *Proceedings of the International Conference on Data Mining*, pages 7–20, Rio de Janeiro, RJ.

Freitas, A. A. (1998b). On objective measures of rule surprisingness. In *Proceedings of the Second European Symp. LNAI*, volume 1510, pages 1–9.

Gomes, A. K. (2001). Análise de regras utilizando medidas de avaliação e de interessabilidade do conhecimento simbólico. Exame de Qualificação de Mestrado, ICMC-USP.

Gomes, A. K., Bernardini, F. C., and Monard, M. C. (2002). Uma sintaxe padrão Prolog para classificadores simbólicos. Technical Report 154, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_154.ps.zip.

Gomes, A. K. and Monard, M. C. (2002). Descrição do módulo de análise de regras do $\mathcal{R}_{ule}\mathcal{S}_{ystem}$. Technical Report 155, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_155.ps.zip.

Hilderman, R. J. and Hamilton, H. J. (1999). Knowledge discovery and interestingness measures: A survey. Technical Report s4s, University of Regina.

Horst, P. S. (1999). Avaliação do conhecimento adquirido por algoritmos de aprendizado de máquina utilizando exemplos. Dissertação de Mestrado, ICMC-USP.

Imamura, C. Y. (2001). Pré-processamento para extração de conhecimento de bases textuais. Dissertação de Mestrado, ICMC-USP.

Lavrac, N., Flach, P., and Zupan, B. (1999). Rule evaluation measures: a unifying view. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming. LNAI*, volume 1634, pages 74–185.

Lee, H. D. (2000). Seleção e construção de features relevantes para o aprendizado de máquina. Dissertação de Mestrado, ICMC-USP.

Lee, H. D. and Monard, M. C. (2000a). Indução construtiva guiada pelo conhecimento: Um estudo de caso do processamento de sm̂en diagnóstico. In *Proceedings IBERAMIA-SBIA 2000, Open Discussion Track*, pages 167–176, Atibaia, SP, Brasil.

Lee, H. D. and Monard, M. C. (2000b). A practical approach for knowledge-driven constructive induction. In *Procedings Argentine Symposium on Artificial Intelligence, ASAI'2000, 29th International Conference SADIO*, pages 71–86, Argentina.

Martins, C. A. (2001). Clustering conceitual em aprendizado de máquina. Exame de Qualificação de Doutorado, ICMC-USP.

Milaré, C. R. (2000). Extração de conhecimento de redes neurais. Exame de Qualificação de Doutorado, ICMC-USP.

Nagai, W. A. (2000). Avaliação do conhecimentos extraído de problemas de regressão. Dissertação de Mestrado, ICMC-USP.

Pila, A. D. (2001). Seleção de atributos relevantes para aprendizado de máquina utilizando a abordagem de Rough Sets. Dissertação de Mestrado, ICMC-USP.

Prati, R. C., Baranauskas, J. A., and Monard, M. C. (1999). BibView: Um sistema para auxiliar a manutenção de registros para o BibTeX. Technical Report 95, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_95.ps.zip.

Prati, R. C., Baranauskas, J. A., and Monard, M. C. (2001a). Extração de informações padronizadas para a avaliação de regras induzidas por algoritmos de aprendizado de máquina simbólico. Technical Report 145, ICMC-USP. `ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_145.ps.zip`.

Prati, R. C., Baranauskas, J. A., and Monard, M. C. (2001b). Uma proposta de unificação da linguagem de representação de conceitos de algoritmos de aprendizado de máquina simbólicos. Technical Report 137, ICMC-USP. `ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_137.ps.zip`.

Pugliesi, J. B. (2001). O pós-processamento em extração de conhecimento de bases de dados. Exame de Qualificação de Doutorado, ICMC-USP.