Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Aprendizado de Máquina Proposicional e Relacional para Extração de Conhecimento: Um Estudo de Caso

Chandler Wellington Caulkins

Maria Carolina Monard/ILTC

Nº 122

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos

Novembro/2000

Aprendizado de Máquina Proposicional e Relacional para Extração de Conhecimento: Um Estudo de Caso [1]

Chandler Wellington Caulkins

Maria Carolina Monard/ILTC

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

Resumo Sistemas de aprendizado proposicional têm sido amplamente aplicados em problemas de "Data Mining", obtendo um êxito razoável em várias aplicações. No entanto, há limitações nesses sistemas devido à representação utilizada ter o poder expressivo de lógica proposicional. Dessa forma, não é possível representar objetos estruturados complexos e/ou relações entre objetos e seus componentes. Neste trabalho discute-se o uso de aprendizado proposicional, através de um experimento realizado com o sistema $\mathcal{S}$ee5 e de aprendizado relacional, através do sistema Progol, em uma base de dados do Programa de Melhoramento Genético da Raça Nelore (PMGRN). PLI é utilizada no sentido de encontrar relações que permitam distinguir características dos bois mais rentáveis e melhorar a produção dos animais da raça Nelore no País. Os resultados obtidos mostram que é possível fazer uso de PLI para induzir regras a partir da base de dados do PMGRN.

Palavras Chaves: aprendizado de máquina proposicional, aprendizado de máquina relacional, $\mathcal{S}$ee5, Progol, PMGRN

Este documento encontra-se no site da Biblioteca do ICMC

ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_122.ps.zip

Novembro 2000

[1] Trabalho realizado com auxílio parcial de CAPES e FINEP.

Sumário

1 Introdução **1**

2 Histórico do PMGRN **2**

3 Contexto do Problema **2**

4 Processo para Aquisição de Conhecimento **4**

4.1 Definição do Problema 5

4.2 Coleta, Limpeza e Preparação dos Dados 5

4.3 Aquisição de Conhecimento 7

5 Experimentos Realizados e Resultados **8**

5.1 Experimento com $\mathcal{S}$ee5 8

5.2 Transformação dos Dados de $\mathcal{S}$ee5 para Progol 13

5.3 Progol - EXPERIMENTO P1A 15

5.4 Progol - EXPERIMENTO P1B 17

5.5 Progol - EXPERIMENTO P2 19

6 Conclusões **21**

Lista de Figuras

1 Célula Animal 3

2 Passos do Processo de Aquisição de Conhecimento 5

3 Quantidade de Progênie Gerada pelas Matrizes Mais Produtivas 9

4 Algumas Regras Geradas no Experimento TipoMt_$\mathcal{S}$ee5 10

5 Alguns Touros que Determinam a Classe menoq0 na Sua Progênie 11

6 Quantidade de Progênie Gerada pelos Touros Mais Produtivos 11

7 Alguns Touros que Determinam a Classe maioq0 na Sua Progênie 12

8 Alguns Touros em que o Atributo MGT Determina a Classe menoq0 na Progênie . 13

9 Regras com o Atributo MTDNA 13

Lista de Tabelas

1 Base de Dados Inicial e Reduzida: Descrição dos Atributos Considerados 6

2 Características do Conjunto de Dados para os Dois Problemas Tratados 7

3 Resumo dos Resultados Obtidos com $\mathcal{S}$ee5 9

4 EfMat_$\mathcal{S}$ee5 – Quantidade de Regras por Classe e Atributos Considerados nessas Regras 10

5 Distribuição da Progênie de Alguns Touros na BD Completa do PMGRN para a Classe menoq0 12

6 Distribuição da Progênie de Alguns Touros na BD Completa do PMGRN para a Classe maioq0 12

7 Predicados Criados pelo Método MB e os Atributos Correspondentes para Cada Problema 14

8 Resumo de Resultados Obtidos Utilizando Progol e o Identificador de Exemplo . 15

9 Resumo de Resultados Obtidos Utilizando Progol e o Identificador de Animal . . 17

10 Número e Porcentagem de Exemplos Positivos não Generalizados pelas Regras em cada Experimento 18

11 Resumo dos Resultados Obtidos Utilizando Progol e o Identificador do Animal — sem o MGT 19

12 Resumo dos Resultados Obtidos com Progol Utilizando Conhecimento de Fundo sobre Ancestral (Progenitor) e Identificador do Animal 20

13 Número e Porcentagem de Exemplos Positivos não Generalizados pelas Regras em cada Experimento 20

14 Resumo dos Resultados Obtidos sem Considerar o Atributo MGT 20

1 Introdução

O Programa de Melhoramento Genético da Raça Nelore — PMGRN — da Universidade de São Paulo em Ribeirão Preto, busca a aplicação de novas tecnologias na pecuária no sentido de distinguir características dos bois mais rentáveis a fim de melhorar a produção dos animais da raça Nelore. Para isso, mantém uma base de dados sobre os diversos animais bovinos cadastrados no programa, permitindo obter informações sobre pesos e medidas importantes dos animais, além de suas linhagens paterna e materna. A partir de uma parceria entre o PMGRN e o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo em São Carlos, técnicas de Inteligência Artificial têm sido aplicadas a essa base de dados para extrair conhecimento.

Uma área de pesquisa sendo explorada dentro do PMGRN é a relação entre a linhagem materna de um animal e suas características de produção de leite e de carne, em função do conteúdo do DNA mitocondrial do animal. Com isso é possível separar o gado em dois grupos, *Bos indicus* e *Bos taurus*, através de seu tipo mitocondrial.

Sistemas de Aprendizado de Máquina — AM — proposicional têm sido amplamente aplicados em problemas de "Data Mining", obtendo um êxito razoável em várias aplicações. Assim, uma forma de extrair conhecimento da base de dados do PMGRN pode ser através do uso de AM proposicional. No entanto, há limitações nesses sistemas devido ao fato da representação utilizada para expressar o conceito induzido ter o poder expressivo da lógica proposicional. Dessa forma, não é possível representar objetos estruturados complexos e/ou relações entre objetos e seus componentes. Além disso, conhecimento prévio do domínio só pode ser utilizado de uma forma muito restrita nos sistemas de AM proposicional. Programação Lógica Indutiva – PLI – é uma área na interseção de aprendizado de máquina indutivo com programação lógica, que permite uma representação mais expressiva, além de uma flexibilidade maior na incorporação de conhecimento prévio do domínio, através do uso de uma forma de lógica de primeira ordem. Com isso, torna-se possível encontrar relações entre objetos e atributos, inclusive relações de ancestral entre indivíduos, o que não é possível descrever em lógica proposicional.

Neste trabalho são mostrados e descritos uma série de experimentos realizados para extrair conhecimento de uma base de dados do PMGRN. Esses experimentos são realizados com aprendizado proposicional, utilizando o sistema $\mathcal{S}$ee5 [Rulequest-Research, 1999], e aprendizado relacional, utilizando o sistema de PLI Progol [Muggleton, 1995]. O trabalho está organizado da seguinte forma: na Seção 2 é feito um breve histórico do PMGRN e sua parceria com o LABIC. Na Seção 3 o contexto do problema a ser tratado é colocado. Na Seção 4 um processo de aquisição de conhecimento é esboçado e detalhado por meio de sua aplicação nos experimentos. Na Seção 5 os experimentos realizados com $\mathcal{S}$ee5 e Progol são mostrados e analisados. Finalmente, na Seção 6, algumas conclusões e direções para trabalhos futuros são apresentadas.

2 Histórico do PMGRN

O PMGRN foi criado em junho de 1988, a partir de uma iniciativa conjunta de criadores de gado Nelore e pesquisadores do Departamento de Genética da Faculdade de Medicina – DGFM – da USP em Ribeirão Preto. Com o intuito de buscar tecnologias novas que pudessem ser aplicadas na pecuária, foi implantada a metodologia do Modelo Animal para avaliação de touros, matrizes e animais jovens [Lôbo et al., 1999]. Essa metodologia permite determinar Diferenças Esperadas nas Progênies — DEPs — de características econômicas dos animais como peso e perímetro escrotal, conseguindo assim melhoras na criação desses animais.

A base de dados é composta por uma série de informações sobre os mais de 177.000 animais no programa, tais como: registro genealógico do animal, data de nascimento, informações referentes ao pai (ou touro) e mãe (ou matriz) do animal, manejo do animal, pesagens, DEPs, etc. Cada informação é representada como um atributo dentro da base de dados, a qual possui mais de 80 atributos cadastrados para cada animal. Assim, se tem um material bastante rico para extrair conhecimento desses dados, com vários atributos relevantes para serem estudados separadamente.

Em 1997, o DGFM iniciou um projeto de pesquisa em parceria com o LABIC da USP – São Carlos para analisar as informações nas bases de dados do PMGRN, a fim de descobrir relações entre essas informações que pudessem auxiliar o trabalho do grupo de pesquisa. Alguns resultados interessantes, utilizando Redes Bayesianas, já foram obtidos e publicados [da Rocha, 1999, Lôbo et al., 1999]. Esses resultados iniciais geraram interesse em realizar outros estudos nas bases de dados. Um desses estudos envolve um novo atributo introduzido na base, chamado de MTDNA ou "tipo mitocondrial". Na seção seguinte, mostra-se a motivação da criação desse atributo e o interesse que tornou-o objeto de análise no estudo por nós realizado.

3 Contexto do Problema

As raças bovinas domesticadas existentes no mundo hoje podem ser divididas, basicamente, em dois grupos ou sub-espécies: os taurinos, conhecidos cientificamente como *Bos taurus*, e os Zebu, conhecidos como *Bos indicus*. O segundo grupo é composto de três raças, a Gyr, a Brahma e a Nelore, sendo esta última objeto de pesquisa do PMGRN.

Os taurinos foram trazidos ao continente sul-americano no século XVI, importados da Espanha e de Portugal. Os animais Zebu, dentre os quais exemplares da raça Nelore, foram importados da Índia. A população Nelore brasileira é oriunda das importações de 1930 e, notadamente, das de 1960 e 1962. Embora estivessem inclusas algumas fêmeas nos lotes de Zebu importados no País, a maioria dos touros Zebu foram cruzados com fêmeas taurinas "nativas", descendentes dos animais originários da Espanha e de Portugal.

Com esses cruzamentos, ocorreu uma miscigenação das genes das duas sub-espécies, principalmente no núcleo de suas células. A célula animal pode ser dividida, a grosso modo, em duas partes: o núcleo e o citoplasma. Na Figura 1, a composição do núcleo é indicada pelos números 6, 10 e 11 e o citoplasma é todo o resto da célula. No núcleo, a herança genética se dá por meio do pai e da mãe de um indivíduo. Já a herança citoplásmatica é controlada essencialmente através da expressão de genes contidas na mitocôndria, indicada pelo número 9 na Figura 1. Essas genes são herdadas, nos mamíferos, quase que exclusivamente pela linhagem materna. Com isso, um animal pode ter "tipo mitocondrial" *Bos taurus* ou *Bos indicus*, de acordo com o conteúdo da mitocôndria de suas células. É interessante observar que, de acordo com experimentos realizados e publicados na literatura, somente um terço da população considerada Nelore no País apresenta o tipo mitocondrial *Bos indicus* [Meirelles et al., 1999].

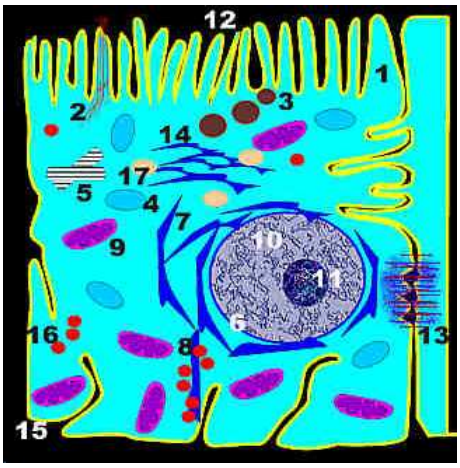

Figura 1: Célula Animal

Já foi mostrado que características bovinas de produção de carne e de leite variam em diferentes linhagens maternas. Dessa forma, evidencia-se a influência do tipo *Bos taurus* nos cruzamentos do rebanho brasileiro. Isso resultou num maior interesse por parte do PMGRN em pesquisar a influência que o tipo mitocondrial pode ter nas características citadas anteriormente.

A identificação, ou tipagem, do animal é realizada em laboratório através de uma série de testes relativamente custosos, analisando o sêmen dos machos e os leucócitos das fêmeas. Tendo em vista esse custo, foram selecionados 212 tourinhos, ou bezerros, advindos de duas fazendas do PMGRN, para formar um grupo inicial de tipagem. A partir desses testes, 179 bezerros foram identificados como *Bos taurus* e 33 como *Bos indicus*, informação essa que foi inserida na base de dados.

Com esses dados, e baseando-se na informação fornecida pelo especialista de que o material genético da mitocôndria é totalmente herdado pela linhagem materna, é possível descobrir o tipo

mitocondrial de vários animais e gerações dentro da base do PMGRN. O tipo de uma matriz pode ser conhecido através de um dos bezerros tipados em laboratório, ou o tipo de um animal pode ser conhecido desde que na sua árvore genealógica haja uma matriz em comum na árvore genealógica de algum dos bezerros já tipados. Para tanto, é necessário também conhecer a matriz dos animais cadastrados. Analisando os 177.379 registros de animais da base de dados, constatou-se que 32.753 não possuem essa informação, de modo que foram eliminados da base a ser tratada. Assim, restou um total de 144.626 registros de animais para os quais existe a possibilidade de deduzir seu tipo.

Para ampliar o número de animais com tipo conhecido, sem recorrer aos testes em laboratório, foi por nós implementado um programa, chamado de Tipagem, na linguagem de programação lógica Prolog. Nesse programa, a informação obtida a partir dos 212 bezerros tipados em laboratório é propagada através das linhagens maternas dos outros 144.414 registros de animais. Como resultado dessa computação foi possível tipar 3.299 novos animais, os quais, junto com os 212 bezerros tipados em laboratório, totalizam uma base de 3.511 registros de animais tipados. Desses 3.511 animais, 1.404 foram tipados como *Bos indicus* e 2.107 como *Bos taurus*. Essa base compõe o subconjunto inicial de dados utilizado neste trabalho.

A seguir é apresentado o processo genérico de aquisição de conhecimento baseado em dados, o qual serviu como base para a realização da pesquisa.

4 Processo para Aquisição de Conhecimento

Um processo de aquisição de conhecimento pode ser dividido basicamente em quatro passos [Caulkins, 2000], ilustrados na Figura 2.

1. Definição e compreensão do problema;
2. Coleta, preparação e limpeza dos dados;
3. Aquisição de conhecimento e
4. Análise dos resultados.

O processo se inicia com a definição do problema (passo 1) em entrevista com o especialista do domínio, buscando compreender melhor o domínio do problema a ser tratado. A partir das informações fornecidas pelo especialista, os dados são coletados, limpados e preparados (passo 2) de modo que possam ser fornecidos a um sistema de aprendizado. Em seguida, os dados são apresentados a um ou mais sistemas de aprendizado de modo a realizar a aquisição de conhecimento

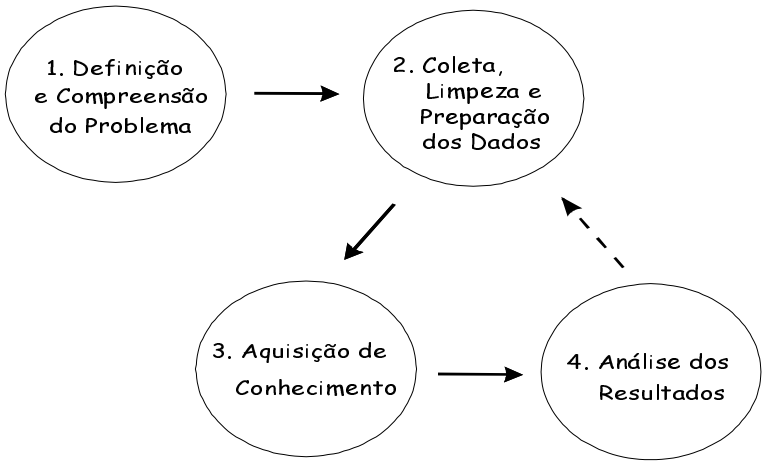

Figura 2: Passos do Processo de Aquisição de Conhecimento

propriamente dita (passo 3). O resultado dessa aquisição é um conhecimento que deve ser analisado em conjunto com o especialista (passo 4). Se a análise mostrar que os resultados não foram satisfatórios, geralmente retoma-se o processo a partir do passo 2. Pode ser que seja necessário realizar esse ciclo várias vezes até que se obtenha um conhecimento que seja de proveito para o especialista.

A seguir esses passos são detalhados com sua aplicação no estudo por nós realizado.

4.1 Definição do Problema

Em conversas com o Prof. Raysildo Lôbo, coordenador do PMGRN, e o analista de sistemas Luiz Bezerra, responsável pela administração de toda a base de dados do PMGRN, foi levantada a possibilidade da influência do tipo mitocondrial na Diferença Esperada na Progênie — DEP — para efeito materno aos 120 dias, ou DMPP120. Essa DEP prediz a diferença esperada em peso dos produtos, ou filhotes, das filhas do animal avaliado, devido às diferenças na habilidade maternal apresentadas por elas. Sendo assim, decidiu-se por representar dois problemas para serem analisados: um para verificar se existe a influência do tipo mitocondrial na DMPP120, e um outro relacionado à influência que a DMPP120 possa ter no tipo mitocondrial. Em ambos os casos, consideram-se informações sobre o animal e seus progenitores.

4.2 Coleta, Limpeza e Preparação dos Dados

A base de dados inicial para este estudo de caso, como já mencionado, possui 3.511 animais tipados e mais de 80 atributos. Entre esses registros há casos onde um atributo possui um valor

inválido e outros onde ele não possui um valor, isto é, o valor é desconhecido. Para manter a base completa, esses registros foram eliminados. Assim, foram retirados da base em torno de 400 registros que possuíam MGT (Mérito Genético Total) com o valor inválido -9.99, utilizado pelos analistas do PMGRN para indicar que não foi possível calcular alguma das DEPs para um determinado animal. Foram removidos ainda alguns registros para os quais o pai não estava registrado no atributo G_PAI (valor desconhecido). Com isso, a base de dados inicial foi reduzida a um total de 3.079 registros.

Também foram selecionados alguns atributos considerados mais relevantes. Tendo por base as indicações dos especialistas, foram selecionados sete atributos: G_ANIM, G_PAI, G_MAE, SX, MGT, DMPP120 e MTDNA. Esses atributos são descritos na Tabela 1, mostrando os valores possíveis e encontrados, tanto para a base de dados inicial de 3.511 animais quanto para a base de dados reduzida de 3.079 animais.

			#Valores Distintos			
Nome do	Descrição	Tipo	BD Inicial		BD Reduzida	
Atributo	do Atributo	do Atributo	possíveis	encontr.	possíveis	encontr.
G_ANIM	Código do animal	discreto	3511	3511	3079	3079
G_PAI	Código do pai do animal	discreto	393	394	362	362
G_MAE	Código da mãe do animal	discreto	1140	1140	1125	1125
SX	Sexo: 0 - indefinido (morto ao nascer), 1 - macho, 2 - fêmea	discreto	3	4	3	2
MGT	Mérito Genético Total do animal: soma ponderada das DEPs mais importantes	contínuo	-	262	-	261
DMPP120	DEP para efeito materno aos 120 dias	contínuo	-	482	-	482
MTDNA	Tipo Mitocondrial: i - Bos indicus, t - Bos taurus	discreto	2	2	2	2

Tabela 1: Base de Dados Inicial e Reduzida: Descrição dos Atributos Considerados

Tendo obtido essa base de dados reduzida, delineou-se o objetivo do trabalho como sendo utilizar um algoritmo de aprendizado de máquina proposicional e outro relacional e analisar, em igualdade de condições, o conhecimento induzido por ambos os algoritmos. Considerando as indicações dos especialistas, foram escolhidos dois atributos da Tabela 1 para serem utilizados como classe:

1. MTDNA, ou Tipo Mitocondrial, e
2. DMPP120, ou DEP de Efeito Materno aos 120 dias.

No primeiro caso, o atributo MTDNA, que pode assumir somente dois valores (i e t), é considerado como atributo-classe. No segundo caso, o atributo considerado, DMPP120, é contínuo e necessita ser discretizado [Molina Félix et al., 1999]. Seguindo as indicações dos especialistas, esse atributo, quando utilizado como classe, pode separar os casos em aqueles com valor DMPP120 > 0 (classe maioq0) e DMPP120 ≤ 0 (classe menoq0). Assim, tem-se dois problemas diferentes, um quando o tipo mitocondrial é utilizado para classificar os dados, o qual é denominado TipoMT, e o outro

quando a DEP do efeito materno é utilizado para esse fim, o qual é denominado de EfMat, como descrito na Tabela 2.

Conjunto de Dados	#Instâncias	#Duplicados ou Conflitantes (%)	#Atributos (cont.,discr.)	Classe	% Classe	Erro Majoritário	Valores Desconhecidos
TipoMt	3.079	0 (0%)	6 (2,4)	i t	39,79% 60.21%	39,79% em t	Não
EfMat	3.079	0 (0%)	6 (1,5)	menoq0 maioq0	35,08% 64,92%	35,08% em maioq0	Não

Tabela 2: Características do Conjunto de Dados para os Dois Problemas Tratados

Nessa tabela, descreve-se o número de instâncias, ou registros, considerado em cada problema, e quantas são duplicadas ou conflitantes. Além disso, mostra-se quantos dos atributos utilizados são contínuos e quantos são discretos. Para o problema TipoMt, são dois atributos contínuos, pois o DMPP120 é considerado como contínuo quando o atributo MTDNA é utilizado como classe. No problema EfMat há apenas um atributo contínuo, o MGT, pois o atributo DMPP120 está discretizado e é considerado como classe.

4.3 Aquisição de Conhecimento

Com relação aos sistemas de aprendizado, foram utilizados especificamente os seguintes sistemas nos dois conjuntos de dados da Tabela 2:

1. o sistema de aprendizado proposicional $\mathcal{S}$ee5 [Rulequest-Research, 1999], e
2. o sistema de Programação Lógica Indutiva Progol [Muggleton, 1995].

Com relação ao sistema Progol, decidiu-se realizar dois tipos de experimentos utilizando cada conjunto de dados. No primeiro experimento, denominado experimento P1, Progol é utilizado em ambos conjuntos de dados sem inserir informação adicional como conhecimento de fundo. Ou seja, somente os dados correspondentes aos exemplos de cada um dos dois conjuntos de dados são fornecidos ao Progol. Esses dados são apresentados ao Progol de duas formas diferentes: primeiramente com cada exemplo identificado por um identificador criado através de uma ferramenta de conversão por nós implementada [Caulkins e Monard, 2000], constituindo o experimento P1a, e em seguida com cada exemplo identificado pelo identificador G_ANIM, constituindo o experimento denominado P1b.

No segundo experimento, denominado experimento P2, além do conhecimento utilizado nos experimentos P1, decidiu-se fornecer ao Progol conhecimento de fundo extra, na forma intensional. Especificamente, esse conhecimento consiste das duas seguintes relações:

```
progenitorMat(X, Y) :- mae(X, Y).
progenitorMat(X, Y) :- mae(X, Z),
                progenitorMat(Z, Y).
```

```
progenitorPat(X, Y) :- pai(X, Y).
progenitorPat(X, Y) :- pai(X, Z),
                progenitorPat(Z, Y).
```

A introdução desse conhecimento de fundo extra tem como objetivo verificar se PROGOL consegue induzir relações não somente com os progenitores do animal, como no experimento P1, mas também considerando outros ancestrais existentes em sua linhagem paterna e/ou materna.

5 Experimentos Realizados e Resultados

Nesta seção são detalhados cada um dos experimentos realizados e os resultados obtidos utilizando os dois conjuntos de dados da Tabela 2.

5.1 Experimento com $\mathcal{S}$ee5

Como já mencionado, decidiu-se utilizar como sistema de aprendizado proposicional o $\mathcal{S}$ee5, pois consideramos que é um dos melhores sistemas para induzir regras de decisão. Na realidade, $\mathcal{S}$ee5 é um produto comercial que inclui melhorias dos algoritmos $\mathcal{C}4.5$ e $\mathcal{C}4.5$-rules [Quinlan, 1993], que induzem árvores e regras de decisão, respectivamente [Baranauskas e Monard, 2000].

Primeiramente, os conjuntos de dados TipoMt e EfMat — Tabela 2 — foram armazenados em uma tabela de base de dados utilizando o Microsoft Access^{TM}. Em seguida foram transformados para o formato requerido por $\mathcal{S}$ee5 através de um comando disponível no próprio Access^{TM}. $\mathcal{S}$ee5 foi executado, obtendo um conjunto de regras de conhecimento para cada experimento, aos quais denominamos:

1. TipoMt_$\mathcal{S}$ee5 para o conjunto de dados TipoMt, e
2. EfMat_$\mathcal{S}$ee5 para o conjunto de dados EfMat.

A Tabela 3 mostra os resultados desses experimentos, em que o erro aparente refere-se ao erro obtido utilizando todos os 3.079 exemplos como conjunto de treinamento e conjunto de teste. O erro verdadeiro refere-se ao erro obtido através de *10-fold cross validation*. O "# de Regras" refere-se ao número de regras induzidas utilizando todos os 3.079 exemplos e o tempo de execução (tempo de CPU) refere-se ao tempo utilizado para gerar essas regras em um microcomputador Pentium II de 300MHz, com o sistema operacional Windows NT^{TM}.

É interessante observar que no experimento TipoMt_$\mathcal{S}$ee5, no qual a classe é a tipagem do animal, $\mathcal{S}$ee5 induziu uma regra para cada uma das 1.125 matrizes na base de dados, e cada regra induzida

Experimento	Erro Aparente (%)	Erro (10 CV) Verdadeiro (%)	# de Regras	Tempo de Execução
TipoMt_$\mathcal{S}$ee5	0,0	4,5 ± 0,5	1.125	4,3 seg.
EfMat_$\mathcal{S}$ee5	10,3	16,6 ± 0,6	220	2,5 seg.

Tabela 3: Resumo dos Resultados Obtidos com $\mathcal{S}$ee5

consiste somente de um antecedente, que é a matriz. A Figura 3 mostra o número de progênie gerada pelas matrizes mais produtivas enquanto na Figura 4 são mostradas algumas regras geradas por $\mathcal{S}$ee5. Por exemplo, a primeira regra na Figura 4, `Rule 1`, simplesmente especifica que

Se `G_MAE = 56402` **então** `class = I`,

com um fator de confiança de 0.960 cobrindo 23 exemplos, que são justamente os 23 exemplos referentes à progênie gerada por essa matriz, como mostra a Figura 3. Em outras palavras, $\mathcal{S}$ee5 descobriu o óbvio, que a tipagem da matriz determina a tipagem do animal por ela gerado.

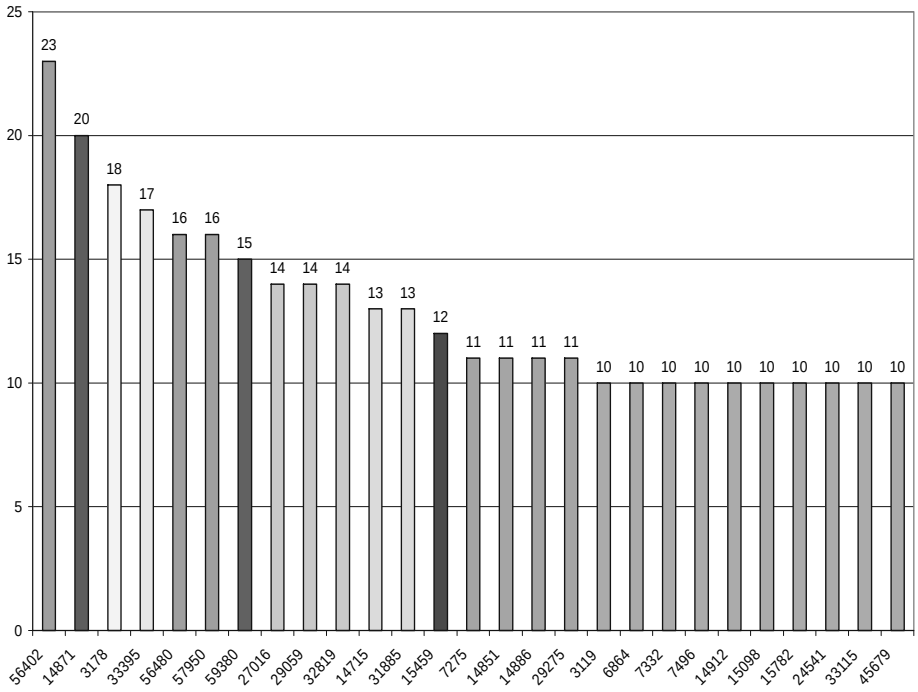

Figura 3: Quantidade de Progênie Gerada pelas Matrizes Mais Produtivas

No experimento EfMat_$\mathcal{S}$ee5, no qual a classe está relacionada com o efeito materno aos 120 dias, foram geradas 220 regras, cinco vezes menos regras que no experimento anterior, com um erro verdadeiro aproximadamente quatro vezes maior — Tabela 3. Ainda assim, o erro verdadeiro, 16,6 ± 0,6, é bem menor que o erro majoritário de 35,08% — Tabela 2.

A Tabela 4 mostra os atributos considerados nas regras induzidas por $\mathcal{S}$ee5, bem como o número

```
Rule 1: (cover 23)
     G_MAE = 56402
->  class I  [0.960]

Rule 2: (cover 18)
     G_MAE = 3178
->  class I  [0.950]
```

```
Rule 361: (cover 20)
     G_MAE = 14871
->  class T  [0.955]

Rule 362: (cover 17)
     G_MAE = 33395
->  class T  [0.947]
```

Figura 4: Algumas Regras Geradas no Experimento TipoMt_$\mathcal{S}$ee5

de regras em que esses atributos aparecem no antecedente das regras. Segue uma análise das 220 regras induzidas por $\mathcal{S}$ee5:

- A mãe do animal (atributo G_MAE) não é relevante para distinguir a habilidade materna de nenhum animal, já que esse atributo não aparece em nenhuma das regras induzidas para esse problema.

- Em todas as regras induzidas o atributo G_PAI aparece, exceto nas seguintes regras

```
Rule 120: (cover 374)
     MGT <= -0.34
->  class menoq0  [0.665]
```

```
Rule 180: (cover 2020)
     MGT > 0.03
->  class maioq0  [0.761]
```

  Deve ser observado que ainda que o fator de confiança de ambas as regras não seja muito alto, elas cobrem um grande número de animais. Essas regras estabelecem que certos animais podem ser classificados considerando somente seu mérito genético total (MGT). No caso do MGT > 0.03, há 2020 animais com boa habilidade materna. No caso do MGT $<=$ -0.34, há 374 animais sem boa habilidade materna. Esses dois valores de MGT deveriam ser melhor analisados pelos especialistas para verificar quais as características dos progenitores de um animal para gerar uma progênie com MGT > 0.03.

Atributo Presente na Regra				Número de	Regras por Classe	
G_PAI	SX	MGT	MTDNA	Regras	menoq0	maioq0
•				168	104	64
		•		2	1	1
•		•		44	29	15
•	•			2	1	1
•			•	1	–	1
•	•	•		1	1	–
•		•	•	1	1	–
•	•	•	•	1	1	–
Número Total de Regras				220	138	82

Tabela 4: EfMat_$\mathcal{S}$ee5 – Quantidade de Regras por Classe e Atributos Considerados nessas Regras

- Existem 104 touros cuja progênie é toda da classe menoq0 e 64 cuja progênie é toda da classe maioq0, independentemente de outros fatores. Como exemplo dessas regras, são mostradas na Figura 5 duas da classe menoq0.

```
Rule 1: (cover 70)                                  Rule 2: (cover 22)
     G_PAI = 2993                                        G_PAI = 7434
->  class menoq0  [0.972]                           ->  class menoq0  [0.958]
```

Figura 5: Alguns Touros que Determinam a Classe menoq0 na Sua Progênie

Elas estabelecem que toda a progênie do G_PAI = 2993 e toda a progênie do G_PAI = 7434 são da classe menoq0, como pode ser confirmado na Figura 6. Ou seja, o número de instâncias coberto por cada uma dessas regras é exatamente o número total da progênie de cada um desses touros, na base de dados utilizada no experimento.

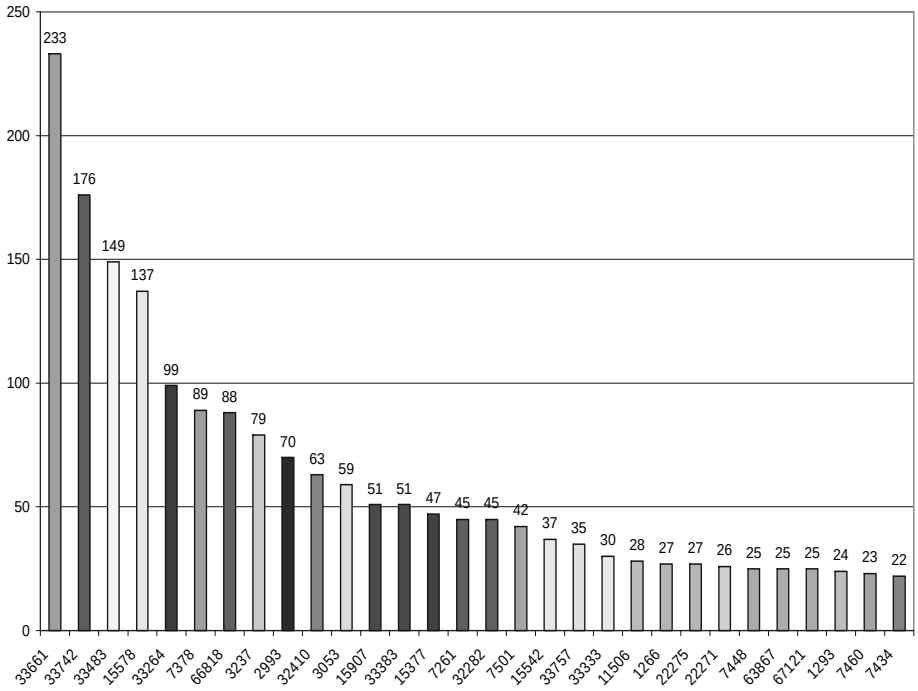

Figura 6: Quantidade de Progênie Gerada pelos Touros Mais Produtivos

Entretanto, a base de dados por nós considerada neste estudo de caso consiste de 3.079 registros de animais. Ou seja, menos de 2% da base de dados completa do PMGRN que tem 177.379 registros. Para os dois touros mencionados, na base de dados completa do PMGRN tem-se o número total de progênie mostrado na Tabela 5. Pode ser observado que nessa base completa, a cobertura para a classe menoq0 não é de 100% para esses dois animais, mas cobre mais de 90% em ambos os casos.

Os melhores reprodutores são aqueles que somente geram animais da classe maioq0 e $\mathcal{S}$ee5 induziu 64 regras para touros diferentes, independentemente de outros fatores. A Figura 7 mostra regras que identificam alguns desses animais, pois cobrem toda a progênie deles,

G_PAI	# de Progênie	Progênie menoq0	% menoq0
7434	254	253	99,6
2993	686	629	91,7

Tabela 5: Distribuição da Progênie de Alguns Touros na BD Completa do PMGRN para a Classe menoq0

como pode ser verificado através da Figura 6. O número total de progênie desses mesmos touros na base de dados completa do PMGRN é mostrada na Tabela 6, onde pode ser observado que a cobertura não é 100% para todos os touros, mas é sempre maior que 86%.

```
Rule 139: (cover 137)
     G_PAI = 15578
->  class maioq0  [0.993]

Rule 140: (cover 79)
     G_PAI = 3237
->  class maioq0  [0.988]

Rule 141: (cover 99)
     G_PAI = 33264
->  class maioq0  [0.970]

Rule 142: (cover 30)
     G_PAI = 33333
->  class maioq0  [0.969]

Rule 143: (cover 25)
     G_PAI = 7448
->  class maioq0  [0.963]

Rule 145: (cover 47)
     G_PAI = 15377
->  class maioq0  [0.959]

Rule 146: (cover 63)
     G_PAI = 32410
->  class maioq0  [0.954]

Rule 154: (cover 37)
     G_PAI = 15542
->  class maioq0  [0.923]

Rule 155: (cover 59)
     G_PAI = 3053
->  class maioq0  [0.902]
```

Figura 7: Alguns Touros que Determinam a Classe maioq0 na Sua Progênie

G_PAI	# de Progênie	Progênie maioq0	% maioq0
33333	156	156	100,0
63867	52	51	98,1
33264	339	332	97,9
15578	798	772	96,7
7448	313	288	92,0
15542	130	118	90,8
3237	3014	2706	89,8
32410	115	101	87,8
15377	113	98	86,7
33757	134	116	86,6

Tabela 6: Distribuição da Progênie de Alguns Touros na BD Completa do PMGRN para a Classe maioq0

- Em geral, quando o mérito genético total (atributo MGT) aparece nas regras induzidas, serve para "separar" a progênie de um mesmo touro em classes diferentes. A Figura 8 mostra algumas dessas regras para o touro 14901 que tem 12 filhotes de sua progênie no conjunto de dados utilizado, e o touro 33386 que tem 15 filhotes nessa base.

  Deve ser observado que o valor de MGT que separa a progênie de cada um desses touros é diferente. Esse valor é -0.05 para o primeiro e -0.15 para o segundo touro. Como esses valores são diferentes para outros touros que aparecem nesse tipo de regra, pode-se concluir que os valores de MGT são utilizados por $\mathcal{S}$ee5 para especializar essas regras de modo a cobrir a progênie de cada touro individualmente, e não para representar um valor geral.

```
Rule 16: (cover 5)
     G_PAI = 14901
     MGT <= -0.05
->  class menoq0  [0.857]

Rule 160: (cover 7)
     G_PAI = 14901
     MGT > -0.05
->  class maioq0  [0.889]
```

```
Rule 57: (cover 2)
     G_PAI = 33386
     MGT <= -0.15
->  class menoq0  [0.750]

Rule 152: (cover 13)
     G_PAI = 33386
     MGT > -0.15
->  class maioq0  [0.933]
```

Figura 8: Alguns Touros em que o Atributo MGT Determina a Classe menoq0 na Progênie

- $\mathcal{S}$ee5 conseguiu induzir somente três regras nas quais o tipo mitcondrial (atributo MTDNA) aparece nos antecedentes – Tabela 4. Essas três regras são mostradas na Figura 9. Elas cobrem poucos casos e não parecem acrescentar qualquer conhecimento.

```
Rule 14: (cover 5)
     G_PAI = 1305
     SX = 2
     MTDNA = T
     MGT <= 0.34
->  class menoq0  [0.857]
```

```
Rule 67: (cover 5)
     G_PAI = 22271
     MTDNA = I
     MGT <= 0.28
->  class menoq0  [0.714]
```

```
Rule 182: (cover 2)
     G_PAI = 3195
     MTDNA = T
->  class maioq0  [0.750]
```

Figura 9: Regras com o Atributo MTDNA

Como pode ser observado, as regras proposicionais induzidas por $\mathcal{S}$ee5 consideram cada touro separadamente, não conseguindo realizar uma boa generalização. Um dos objetivos deste trabalho, como já foi mencionado, é comparar o conhecimento obtido com um sistema de aprendizado proposicional e aquele obtido com um sistema de aprendizado relacional, pois pode haver alguma relação entre animais de uma mesma linhagem. Nas próximas seções são descritos os experimentos realizados utilizando o sistema Progol.

5.2 Transformação dos Dados de $\mathcal{S}$ee5 para Progol

Primeiramente, os dados utilizados com $\mathcal{S}$ee5 devem ser transformados para o formato requerido por Progol. Deve ser ressaltado que os arquivos de definição de tipos e de dados de $\mathcal{S}$ee5 são praticamente os mesmos que os utilizados por $\mathcal{C}$4.5, de modo que os dados podem ser convertidos para os arquivos de Progol fazendo-se uso das ferramentas de conversão por nós implementadas e descritas em [Caulkins e Monard, 2000]. Os conjuntos de dados TipoMt e EfMat (Tabela 2), no formato requerido por $\mathcal{S}$ee5, foram transformados para um formato por nós chamado de MB para que pudessem ser utilizados no experimento P1a. Segundo o método utilizado nas ferramentas de conversão, para cada atributo é criado um predicado. Como nos dois problemas tratados há seis atributos sendo considerados, são criados seis nomes de predicado de a0 até a5, todos de aridade 2. Em todos esses predicados, o primeiro argumento representa um número (#id) que identifica

o exemplo considerado. Como segundo argumento de cada predicado é utilizado o valor de cada um dos atributos correspondentes. A Tabela 7 descreve os predicados gerados pelas ferramentas para os dois conjuntos de dados considerados. Deve ser observado que em todos os experimentos com PROGOL, os identificadores G_ANIM, G_PAI e G_MAE são acrescentados da letra "n" para que o sistema não os trate como números.

TipoMt	EfMat
a0(#id, <G_ANIM>)	a0(#id, <G_ANIM>)
a1(#id, <G_PAI>)	a1(#id, <G_PAI>)
a2(#id, <G_MAE>)	a2(#id, <G_MAE>)
a3(#id, <SX>)	a3(#id, <SX>)
a4(#id, <DMPP120>)	a4(#id, <MTDNA>)
a5(#id, <MGT>)	a5(#id, <MGT>)

Tabela 7: Predicados Criados pelo Método MB e os Atributos Correspondentes para Cada Problema

Para o experimento P1b, o identificador de exemplo foi substituído pelo identificador único do animal (G_ANIM). Com isso, o predicado a0 do experimento anterior se torna redundante e é eliminado do conjunto de predicados considerados. Para realizar o experimento P2, como mencionado anteriormente, foram adicionadas ao conhecimento de fundo dos experimentos P1a e P1b as relações `progenitorPat/2` e `progenitorMat/2` definidas na Seção 4.3. PROGOL foi executado, obtendo um conjunto de regras de conhecimento para os três experimentos aos quais denominamos

1. TipoMt_Progol_X, para o conjunto de dados TipoMt e
2. EfMat_Progol_X, para o conjunto de dados EfMat

onde X = P1a indica o experimento P1a, X = P1b indica o experimento P1b e X = P2 indica o experimento P2. Cada experimento foi executado duas vezes: primeiro considerando os exemplos da classe majoritária como classe positiva e a classe minoritária como classe negativa e, em seguida, a classe minoritária foi apresentada como classe positiva e a classe majoritária como classe negativa. Dessa forma, se tem uma noção mais precisa da classificação como um todo, pois PROGOL induz regras apenas para a classe positiva.

Os resultados desses experimentos são descritos nas próximas seções, nas quais os erros aparente e verdadeiro apresentados foram obtidos para cada uma das execuções através do utilitário `test/1` do PROGOL.

Para cada cláusula encontrada por PROGOL, quatro medidas são calculadas:

- p – o número de exemplos positivos cobertos pela cláusula
- n – o número de exemplos negativos cobertos pela cláusula e

- h – o número de átomos necessários para completar a cláusula. Esse número é calculado inspecionando as variáveis de saída da cláusula e determinando se já foram definidas. Por exemplo, para um predicado `s(A, B)` sendo considerado como cabeça das cláusulas, h seria 3, se fossem necessários, no mínimo, três átomos no corpo de uma cláusula para ligar `A` a `B`. Nos experimentos realizados, a relação a ser aprendida é de aridade 1, portanto o valor de h é sempre 0.

- f – uma medida do poder preditivo e da compressão da cláusula dada pela equação 1

$$f = P(p - (n + c + h))/p \tag{1}$$

  onde c é o comprimento da cláusula, P é o número total de exemplos positivos e p, n e h são os mesmos números calculados anteriormente.

Com essas medidas, PROGOL escolhe a cláusula a ser retornada como sendo aquela que tem o maior valor para f. Se houver mais de uma cláusula com o maior valor, é retornada a primeira que for encontrada, desde que a sua cobertura de exemplos negativos não seja maior que a definida pelo parâmetro `noise`. Esse parâmetro seta o ruído permitido nas regras, ou seja, quantos exemplos negativos uma regra pode cobrir. Nos nossos experimentos esse parâmetro foi setado para 20%.

5.3 PROGOL - EXPERIMENTO P1A

A Tabela 8 mostra os resultados do experimento P1a. Deve ser observado que o número de regras, cláusulas e tempo de execução, dado em segundos, se referem aos valores obtidos executando PROGOL com todos os dados disponíveis. O tempo de execução (tempo de CPU) refere-se ao tempo, em segundos, utilizado para obter as regras num microcomputador Pentium II de 300MHz, rodando o sistema operacional Linux. O número de cláusulas refere-se a todas as cláusulas induzidas por PROGOL. Muitas dessas cláusulas não têm corpo, ou seja, consistem somente de uma cabeça. Na realidade, essas cláusulas simplesmente nomeiam os exemplos positivos que não puderam ser cobertos pelas cláusulas com corpo induzidas por PROGOL. As cláusulas com corpo são denominadas regras nas tabelas a seguir. Assim, a diferença entre o número de cláusulas e o número de regras mostra o número total de exemplos positivos não cobertos pelas regras.

Experimento	Classe Positiva	Erro Aparente (%)	Erro Verdadeiro (%)	No. de Regras	No. de Cláusulas	Tempo de Execução
TipoMt_PROGOL_P1a	t	17,2	42,3 ± 1,3	238	665	134131,49
TipoMt_PROGOL_P1a	i	8,1	24,3 ± 1,1	171	408	45824,00
EfMat_PROGOL_P1a	maioq0	17,9	27,8 ± 1,1	69	386	24919,23
EfMat_PROGOL_P1a	menoq0	15,8	23,0 ± 1,1	65	487	39817,47

Tabela 8: Resumo de Resultados Obtidos Utilizando PROGOL e o Identificador de Exemplo

O experimento TipoMt_PROGOL_P1a utilizando como classe positiva t, foi o mais demorado consumindo mais de 37 horas de tempo de CPU. Este experimento mostra um caso típico de "não aprendizado", considerando o excessivo número de regras induzidas, 238, quando todos os dados disponíveis foram utilizados, e o erro verdadeiro de 42,3 ± 1,3. Esse erro é maior que o erro majoritário de 39,79% — Tabela 2, página 7.

O experimento TipoMt_PROGOL_P1a utilizando como classe positiva i, também gera um número grande de regras, ainda que o erro verdadeiro é inferior ao erro majoritário. Entretanto, analisando as regras geradas, a grande maioria delas identifica os exemplos positivos simplesmente utilizando o identificador da mãe do animal, analogamente a $\mathcal{S}$ee5. Outras regras geradas são descritas a seguir, onde, para um melhor entendimento, foi utilizada a correspondência com os nomes dos atributos originais descritos na Tabela 7, ou seja, `g_pai:a1`, `dmpp120:a4`, `mgt:a5`.

```
f=284,p=9,n=2
exemplo(A) :- dmpp120(A,B), mgt(A,C), C =< 1.130, -2.040 < B.
f=472,p=8,n=0
exemplo(A) :- g_pai(A,n3053), mgt(A,B), B =< -0.580.
f=778,p=16,n=1
exemplo(A) :- dmpp120(A,1.380), mgt(A,B), B < 0.020.
```

O experimento EfMat_PROGOL_P1a utilizando como classe positiva maioq0, gerou um número bem menor de regras, 69, com erro verdadeiro menor que o erro majoritário. Entretanto, a grande maioria das regras também identifica os exemplos positivos utilizando somente o identificador da mãe ou do pai do animal. Segue uma descrição de algumas das regras geradas, onde `sexo:a3, mgt:a5`, nas quais participam outros atributos, e que merecem maior atenção.

```
f=1585,p=610,n=126
exemplo(A) :- sexo(A,sexo2), mgt(A,B), 0.300 =< B.
f=1742,p=66,n=4
exemplo(A) :- sexo(A,sexo1), mgt(A,B), 0.910 =< B.
```

A primeira regra diz que o animal é da classe maioq0 se é uma vaca (sexo2) e seu MGT é maior ou igual a 0.300. A segunda regra diz que o animal é da classe maioq0 se é um touro (sexo1) e seu MGT é maior ou igual a 0.910.

O experimento EfMat_PROGOL_P1a utilizando como classe positiva menoq0, gerou o menor número de regras com o menor erro verdadeiro. Analogamente ao experimento anterior, a grande maioria das regras identifica os exemplos positivos utilizando somente o identificador da mãe ou do pai do animal. Algumas regras geradas são mostradas a seguir, onde `pai:a1`, `sexo:a3`, `mgt:a5`.

```
f=703,p=39,n=9
exemplo(A) :- sexo(A,sexo2), mgt(A,B), B =< -0.600.
f=870,p=94,n=14
exemplo(A) :- pai(A,n33661), mgt(A,B), B =< 0.620.
```

5.4 PROGOL - EXPERIMENTO P1B

Após uma análise cuidadosa dos resultados do experimento P1a, foi possível observar que o PROGOL não conseguiu considerar as relações entre o animal e seus pais, devido ao fato de ter o número do identificador do exemplo, gerado pela ferramenta de conversão de dados para o formato MB. Assim, decidiu-se por utilizar a identificação única (atributo G_ANIM) de cada animal na base, suprimindo a identificação de cada exemplo.

A Tabela 9 mostra os resultados deste experimento.

Experimento	Classe Positiva	Erro Aparente (%)	Erro Verdadeiro (%)	No. de Regras	No. de Cláusulas	Tempo de Execução
TipoMt_PROGOL_P1b	t	23,2	36,1 ± 0,9	241	854	186608,01
TipoMt_PROGOL_P1b	i	5,4	15,5 ± 0,7	106	242	16216,40
EfMat_PROGOL_P1b	maioq0	15,8	24,2 ± 0,8	68	325	16742,79
EfMat_PROGOL_P1b	menoq0	15,7	22,6 ± 0,8	65	481	35534,25

Tabela 9: Resumo de Resultados Obtidos Utilizando PROGOL e o Identificador de Animal

A grande maioria das regras gerada no experimento TipoMt_PROGOL_P1b utilizando i como classe positiva levam a mãe do animal em consideração. Deve ser observado que o tipo mitocondrial dos animais não faz parte do conhecimento de fundo, a fim de que o sistema não induza relações óbvias tais como

```
exemplo(A) :- tipomt(A, t)
```

e/ou

```
exemplo(A) :- mae(A, B), tipomt(B, t)
```

quando `t` é utilizado como classe positiva, ou

```
exemplo(A) :- tipomt(A, i)
```

e/ou

```
exemplo(A) :- mae(A, B), tipomt(B, i)
```

quando `i` é utilizado como classe positiva. A seguir são mostradas algumas das regras obtidas:

```
f=1093,p=34,n=0
exemplo(A) :- mae(A,B), mae(B,n29275).
```

```
f=920,p=55,n=10
exemplo(A) :- mae(A,B), mgt(B,C), C=< -0.850.
```

No experimento EfMat_PROGOL_P1b utilizando maioq0 como classe positiva, a grande maioria das regras também levam a mãe ou pai do animal em consideração. A seguir são mostradas algumas das regras obtidas nesse experimento:

```
f=1585,p=610,n=126
exemplo(A) :- sexo(A,sexo2), mgt(A,B), 0.300=<B.
f=1839,p=134,n=8
exemplo(A) :- pai(A,B), mgt(B,C), -0.310 =<C.
```

Deve ser observado que a primeira regra é idêntica à primeira regra que foi obtida no experimento EfMat_PROGOL_P1a com maioq0 como classe positiva. A segunda regra diz que um animal possui habilidade materna maior que zero (classe maioq0) se seu pai tem MGT >= -0,31.

O número de regras gerado no experimento EfMat_PROGOL_P1b utilizando menoq0 como classe positiva foi o menor de todo o experimento P1b, como pode ser visto na Tabela 9. Suas regras também utilizam o pai ou a mãe do animal como determinante para a classificação.

Um outro aspecto interessante a ser observado tem a ver com o número de regras e exemplos positivos não generalizados por essas regras, os quais são listados como cláusulas sem corpo. Na Tabela 10, a coluna "Não generalizados" mostra a diferença entre o número de cláusulas e o número de regras (cláusulas com corpo) geradas por PROGOL. Em outras palavras, esse número dá o número de exemplos positivos que PROGOL não consegue generalizar. A coluna "% Não generalizados" mostra a porcentagem de exemplos da classe considerada como positiva durante o aprendizado que não foram generalizados pelas regras. Considerando os valores do erro verdadeiro na Tabela 9 e os valores da Tabela 10, pode ser observado que não há uma correlação aparente. Ou seja, a porcentagem de exemplos positivos que PROGOL não consegue generalizar não parece influir no erro verdadeiro de classificação.

Experimento	Classe Positiva	# Instâncias	No. de Regras	No. de Cláusulas	Não Generalizados '	% Não Generalizados
TipoMt_PROGOL_P1b	t	1854	241	854	613	33,06%
TipoMt_PROGOL_P1b	i	1225	106	242	136	11,10%
EfMat_PROGOL_P1b	maioq0	1999	68	325	257	12,86%
EfMat_PROGOL_P1b	menoq0	1080	65	481	416	38,52%

Tabela 10: Número e Porcentagem de Exemplos Positivos não Generalizados pelas Regras em cada Experimento

Um dos interesses do especialista é encontrar alguma relação entre a DEP de efeito materno aos 120 dias e o tipo mitocondrial do animal. Entretanto, como o atributo MGT contribui bastante

para a determinação da DEP de efeito materno, decidiu-se executar o experimento mais uma vez, ignorando o atributo MGT. Além disso, concentrou-se os esforços nos experimentos com o problema EfMat. Os resultados dessa execução podem ser vistos na Tabela 11.

Experimento	Classe Positiva	Erro Aparente (%)	Erro Verdadeiro (%)	No. de Regras	No. de Cláusulas	Tempo de Execução
EfMat_PROGOL_P1b	maioq0	17,3	27,6 ± 0,8	108	531	36937,27
EfMat_PROGOL_P1b	menoq0	20,7	27,3 ± 0,8	69	678	42781,35

Tabela 11: Resumo dos Resultados Obtidos Utilizando PROGOL e o Identificador do Animal — sem o MGT

Comparando os resultados com os mostrados na Tabela 9, pode ser observado que ambos os erros, tanto o aparente quanto o verdadeiro, são maiores neste experimento. Também, o número de exemplos positivos que PROGOL não consegue generalizar é consideravelmente maior que no experimento anterior, 423 exemplos versus 257 exemplos no experimento anterior quando usando maioq0 como classe positiva. No caso de usar menoq0 como classe positiva, 611 exemplos não foram generalizados enquanto que no experimento anterior foram 416 exemplos. O número de regras geradas também é maior, especialmente quando maioq0 é usado como classe positiva, 108 regras enquanto que no experimento anterior foram induzidas somente 68 regras.

Duas das regras obtidas utilizando maioq0 como classe positiva chamam a atenção por ambas incluírem o tipo mitocondrial t (Bos taurus) como determinante para a classe. Essas regras são:

```
f=1695,p=224,n=31
exemplo(A) :- pai(A,B), tipomt(A,t), mae(B,C).
f=820,p=10,n=1
exemplo(A) :- pai(A,n7460), tipomt(A,t).
```

Utilizando menoq0 como classe positiva não foram induzidas regras que chamassem a atenção.

5.5 PROGOL - EXPERIMENTO P2

Neste experimento, acresacentou-se dois conceitos recursivos, `progenitorPat` e `progenitorMat`, ao conhecimento de fundo, como já foi mencionado. Com isso, PROGOL pode, eventualmente, incluir estes conceitos no corpo das cláusulas por ele descobertas, separando os animais não somente através do pai e mãe mas também através de suas linhagens paternas e maternas. Os resultados obtidos são mostrados na Tabela 12. Deve ser observado que os tempos de execução são relativos à execução em um computador com dois processadores Pentium II de 300 MHz, rodando o sistema operacional Linux.

Pode ser observado na Tabela 13 que o número de regras e o número de exemplos positivos que PROGOL não consegue generalizar é maior que nos casos anteriores, 43,72% quando maioq0 é

Experimento	Classe Positiva	Erro Aparente (%)	Erro Verdadeiro (%)	No. de Regras	No. de Cláusulas	Tempo de Execução
EfMat_PROGOL_P2	maioq0	34,7	27,8 ± 1,1	191	1065	84229,71
EfMat_PROGOL_P2	menoq0	22,9	22,2 ± 0,7	145	813	59944,35

Tabela 12: Resumo dos Resultados Obtidos com PROGOL Utilizando Conhecimento de Fundo sobre Ancestral (Progenitor) e Identificador do Animal

usado como classe positiva e 61,85% quando menoq0 é usada como classe positiva. Assim, é possível concluir que a generalização realizada é muito pobre. Uma possível explicação é que os conceitos recursivos `progenitorPat` e `progenitorMat` necessitam, para serem provados, dos ancestrais dos animais no conjunto de exemplos. Entretanto, esse encadeamento não acontece nesta base pois somente são considerados aqueles animais que puderam ser tipados.

Experimento	Classe Positiva	# Instâncias	No. de Regras	No. de Cláusulas	Não Generalizados	% Não Generalizados
EfMat_PROGOL_P2	maioq0	1999	191	1065	874	43,72%
EfMat_PROGOL_P2	menoq0	1080	145	813	668	61,85%

Tabela 13: Número e Porcentagem de Exemplos Positivos não Generalizados pelas Regras em cada Experimento

A seguir, é mostrada uma das regras obtidas utilizando maioq0 como classe positiva que cobre um bom número de exemplos e mostra um encadeamento interessante:

```
f=1855,p=160,n=8
exemplo(A) :- progenitorPat(A,B), progenitorMat(B,C), mgt(B,D),
              -0.310 =<D.
```

Foi observado que também nesse experimento o atributo MGT contribui bastante para determinar a DEP de efeito materno do animal. Assim, analogamente ao experimento P1b, o atributo foi ignorado para tentar verificar se havia algum relacionamento da DEP com o tipo mitocondrial. Na Tabela 14, são mostrados os resultados obtidos nessa execução.

Experimento	Classe Positiva	Erro Aparente (%)	Erro Verdadeiro (%)	No. de Regras	No. de Cláusulas	Tempo de Execução
EfMat_PROGOL_P2	maioq0	35,8	39,0 ± 0,9	160	1433	144310,62
EfMat_PROGOL_P2	menoq0	25,9	29,9 ± 0,8	62	844	54495,16

Tabela 14: Resumo dos Resultados Obtidos sem Considerar o Atributo MGT

Ainda que na Tabela 14 o número de regras geradas tenha diminuído, o número de exemplos positivos que PROGOL não consegue generalizar incrementou consideravelmente. Eles representam 63,68% para classe maioq0 (enquanto que no experimento anterior foram 43,72%) e 72,41% (enquanto que no experimento anterior foram 61,85%). O erro verdadeiro também incrementou

consideravelmente. Para a classe maioq0 o erro verdadeiro de 39,0 ± 0,9 é ainda maior que o erro majoritário de 35,08% — Tabela 2, página 7.

Algumas regras obtidas utilizando maioq0 como classe positiva, e nas quais o tipo mitocondrial é considerado, são mostradas a seguir:

```
f=1067,p=21,n=5
exemplo(A) :- progenitorPat(A,B), tipomt(A,t), tipomt(B,t).
f=1799,p=101,n=7
exemplo(A) :- progenitorPat(A,B), tipomt(B,t).
f=620,p=10,n=2
exemplo(A) :- progenitorPat(A,n7501), tipomt(A,i).
f=820,p=10,n=1
exemplo(A) :- progenitorPat(A,n7460), tipomt(A,t).
```

6 Conclusões

Neste trabalho apresentou-se alguns experimentos para extração de conhecimento de uma base de dados disponibilizada pelo Programa de Melhoramento Genético da Raça Nelore — PMGRN — da Universidade de São Paulo em Ribeirão Preto. Para a realização dos experimentos foram utilizados dois sistemas bastante conhecidos na comunidade de Aprendizado de Máquina — AM: um de aprendizado proposicional, o sistema $\mathcal{S}$ee5, e outro de aprendizado relacional, o sistema Progol. A partir dos resultados foi possível observar as limitações estruturais da lógica proposicional, as quais não permitem ao $\mathcal{S}$ee5 "enxergar" as relações necessárias para extrair conhecimento interessante.

Como esperado, o tempo de execução é muito maior nos experimentos realizados com Progol que naqueles feitos com $\mathcal{S}$ee5. Isto pode ser explicado pelo tamanho do espaço de busca considerado pelos algoritmos de aprendizado relacional. Para que seja possível realizar Data Mining utilizando algoritmos relacionais, o tempo de execução desses algoritmos necessita ser reduzido bastante. Um aspecto positivo do uso de Progol em contraste com outros algoritmos de PLI, por exemplo Foil, é que ele permite a inclusão de conhecimento de fundo na forma intensional, o que facilita a definição de novas relações a serem consideradas como conhecimento de fundo.

Uma solução imediata para contornar o problema de tempo de execução excessivo requerido pelo Progol poderia ser utilizar um número limitado de exemplos de treinamento, e refinar o conhecimento obtido passo a passo, generalizando e especializando as relações induzidas automaticamente pelos sistemas relacionais. Esse refinamento pode ser realizado através de um "workbench" com ferramentas tais como as que se encontram no sistema MOBAL [Morik et al., 1993]. Esse sistema permite modelar conhecimento de um domínio através de uma representação baseada em

lógica de primeira ordem, de modo que é adequado para trabalhar em conjunto com os algoritmos relacionais. Trabalhos futuros estão sendo planejados nesse sentido.

Os resultados obtidos nos experimentos mostram que AM relacional é apropriado para induzir regras a partir da base de dados do PMGRN. Mostram também a possibilidade de geração de regras individuais que possam ter relevância para o problema sendo tratado. Entretanto, se considerarmos a hipótese induzida, ou seja, o conjunto total de cláusulas induzidas em cada um dos experimentos, os resultados deixam a desejar do ponto de vista de classificação. Como mencionado anteriormente, um dos motivos pode ser a falta de encadeamento de progenitores na base reduzida considerada neste trabalho, a qual somente considera animais tipados. Como novos animais estão sendo tipados em laboratório, pretende-se incluí-los nesta base de dados, aumentando assim tanto o número de exemplos de treinamento quanto as possibilidades de maiores e novos encadeamentos que esses animais podem possibilitar. Depois de avaliar junto com os especialistas envolvidos as regras obtidas nos primeiros experimentos, pretende-se desenvolver novos experimentos para uma nova extração de conhecimento.

Referências

[Baranauskas e Monard, 2000] Baranauskas, J. A. e Monard, M. C. (2000). Reviewing some Machine Learning Concepts and Methods. Relatório Técnico 102, ICMC – USP, São Carlos, SP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_102.ps.

[Caulkins, 2000] Caulkins, C. W. (2000). Aquisição de Conhecimento Utilizando Aprendizado de Máquina Relacional. Dissertação de Mestrado, ICMC-USP.

[Caulkins e Monard, 2000] Caulkins, C. W. e Monard, M. C. (2000). Ferramentas para Conversão de Dados de Entrada dos Algoritmos de Aprendizado de Máquina Proposicional CN2 e C4.5 para os Algoritmos de Aprendizado de Máquina Relacional Foil e Progol. Relatório Técnico 116, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_116.ps.zip.

[da Rocha, 1999] da Rocha, C. A. J. (1999). Redes Bayesianas para Extração de Conhecimento de Bases de Dados, Considerando a Incorporação de Conhecimento de Fundo e o Tratamento de Dados Incompletos. Dissertação de Mestrado, ICMC-USP. Dissertação de Mestrado.

[Lôbo et al., 1999] Lôbo, R. B., Bezerra, L. A. F., e de Oliveira, H. N. (1999). *Avaliação Genética de Animais Jovens, Touros e Matrizes.* GEMAC – Departamento de Genética – FMRP – USP, Ribeirão Preto.

[Meirelles et al., 1999] Meirelles, F. V., Rosa, A. J. M., Lôbo, R. B., Garcia, J. M., Smith, L. C., e Duarte, F. A. M. (1999). Is the American Zebu Really Bos Indicus? *Genetics and Molecular Biology*, 22(4):543–546.

[Molina Félix et al., 1999] Molina Félix, L. C., Rezende, S. O., Monard, M. C., e Caulkins, C. W. (1999). Some Experiences with the Discretization Process: from Regression to Classification. In *Proceedings of the Argentine Symposium on Artificial Intelligence*, pages 155–166, Buenos Aires.

[Morik et al., 1993] Morik, K., Wrobel, S., Kietz, J.-U., e Emde, W. (1993). *Knowledge Acquisition and Machine Learning*. Academic Press.

[Muggleton, 1995] Muggleton, S. (1995). Inverse Entailment and Progol. *New Generation Computing*, 13:245–286.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California.

[Rulequest-Research, 1999] Rulequest-Research (1999). Data Mining Tools $\mathcal{S}$ee5 and $\mathcal{C}$5.0. http://www.rulequest.com/see5-info.html.