Instituto de Ciências Matemáticas e de Computação
ISSN - 0103-2569

Uma Medodologia para Auxiliar o Processo de Interpretação Semi-Automática de *Clusters*

Claudia Aparecida Martins

Maria Carolina Monard

Ana Silvia Haedo

Noemí Lorena Matsudo

Nº 133

RELATÓRIOS TÉCNICOS DO ICMC

USP – São Carlos
Março de 2001

Uma Metodologia para Auxiliar o Processo de Interpretação Semi-Automática de *Clusters*[*]

Claudia Aparecida Martins[†‡]
Maria Carolina Monard[†]
Ana Silvia Haedo[§]
Noemí Lorena Matsudo[§]

[†]Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
C.P. 668, 13560-970 - São Carlos, SP - Brasil
e-mail: {cam, mcmonard}@icmc.sc.usp.br

[‡]Universidade Federal de Mato Grosso
Instituto de Ciências Exatas e da Terra
Departamento de Ciência da Computação
Av. Fernando Corrêa da Costa s/n
78060-900, Cuiabá, MT - Brasil

[§]Universidade de Buenos Aires
Departamento de Computação
Ciudad Universitaria - Pabellón I
1428 Buenos Aires - Argentina
e-mail: {haedo}@qb.fcen.uba.ar
e-mail: {nmatsudo}@dc.uba.ar

Resumo

Algoritmos de Aprendizado de Máquina são freqüentemente utilizados para extrair conhecimento de bases de dados. Quando os exemplos da base de dados não estão rotuladas com o atributo classe, é possível utilizar algoritmos de Aprendizado de Máquina não supervisionado para descobrir padrões nos dados, denominados *clusters*. Muitas vezes, é de interesse do especialista tentar "interpretar" ou "explicar" os *clusters* que foram encontrados. Neste trabalho propomos uma metodologia, guiada pelo especialista, para auxiliar o processo de interpretação de clusters, utilizando algoritmos de Aprendizado de Máquina Simbólicos. É apresentado um processo utilizando esta metodologia e um estudo de caso utilizando dados do mundo real.

Palavras Chaves: Aprendizado de Máquina, Mineração de Dados, *Clustering*

Março 2001

[*]Trabalho realizado com auxílio parcial da CAPES e FINEP.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

1 Introdução

O processo de mineração de dados em grandes bases de dados pode ser bastante útil em aplicações práticas. Assim, cada vez mais, cresce o interesse em algoritmos de Aprendizado de Máquina (AM) para extração de conhecimento de base de dados. AM pode ser supervisionado ou não supervisionado, e a escolha de qual deles usar depende dos exemplos contidos na base, geralmente no formato atributo-valor, estarem ou não rotuladas com o atributo classe. Quando os dados estão rotulados com classes conhecidas é possível utilizar algoritmos de AM supervisionados, os quais induzem conceitos dos dados.

Por outro lado, no aprendizado não supervisionado os exemplos não estão rotulados e é necessário de alguma forma descobrir algum agrupamento desses exemplos. Muitas vezes, utilizar esse tipo de aprendizado pode ser uma tarefa cara em termos de custo e tempo. Assim, no caso dos dados não estarem explicitamente rotulados com uma classe, é possível utilizar algoritmos de AM não supervisionado, os quais procuram por padrões nos dados a partir de alguma caracterização de regularidade (Decker & Focardi, 1995). Esses padrões são denominados *clusters* (McCallum et al., 2000) sendo que os exemplos contidos em um mesmo *cluster* são mais similares, segundo alguma medida de similaridade, que aqueles contidos em *clusters* diferentes.

Mas, apenas agrupar dados de acordo com alguma medida de similaridade, conceitualmente, pode ser pouco representativo. Muitas vezes, é de interesse do especialista do domínio tentar encontrar uma "interpretação" ou "explicação" para os dados contidos em cada *cluster* bem como reconhecer se dois ou mais *clusters* agrupam exemplos que podem ser considerados de uma mesma classe.

Neste trabalho propomos uma metodologia utilizando algoritmos de AM simbólicos, para auxiliar o especialista na tarefa de interpretação de *clusters.* É apresentado um processo utilizando esta metodologia e um estudo de caso que usa dados do mundo real. Resultados preliminares encontram-se publicados em (Martins & Monard, 2000).

O trabalho está organizado da seguinte forma: na Seção 2 é descrita uma visão geral dessa metodologia. Na Seção 3 é mostrado um processo que utiliza essa metodologia, bem como os algoritmos de AM utilizados e uma ferramenta computacional por nós implementada, presentes nesse processo. Na Seção 4, um estudo de caso utilizando o processo proposto é realizado com dados do mundo real. A descrição da base de dados utilizada, a limpeza dos dados, os experimentos e resultados obtidos também são mostrados. Finalmente, na Seção 5 são apresentadas algumas conclusões.

2 Metodologia

A metodologia que propomos para auxiliar a tarefa de interpretação de *clusters* de forma semi-automática é ilustrada na Figura 1. Basicamente, a metodologia é composta pelas seguintes quatro etapas:

1. Uma base de dados já pré-processada com exemplos não rotuladas, no formato atributo-valor, é submetida a um processamento realizado por algum algoritmo de AM não supervisionado. Esse algoritmo é o responsável por descobrir *clusters* presentes na base de dados;

2. O resultado obtido (*clusters* encontrados) é processado por uma ferramenta computacional por nós implementada, que rotula os exemplos da base original, ou um subconjunto desses exemplos, com o *cluster* ao qual pertencem. Assim, é gerada uma base de dados com uma dimensão adicional, a qual é considerada como sendo o atributo classe desses exemplos;

3. A nova base de dados gerada na etapa anterior possui as características necessárias para ser utilizada como entrada para algoritmos de AM supervisionado. Como o interesse é tentar explicar os *clusters* previamente encontrados, a linguagem de descrição de conceitos (ou hipóteses) utilizada pelo algoritmo de AM supervisionado escolhido deve ser uma linguagem simbólica, tal como regras ou árvores de decisão. Assim, os *clusters* podem ser descritos, simbolicamente, através dessas linguagens;

4. Finalmente, o conhecimento do especialista do domínio é de fundamental importância ao se tentar dar uma interpretação semântica aos *clusters*, agora descritos utilizando outro formalismo. Com a interpretação do especialista é possível, então, ter uma compreensão e uma "explicação" para os dados pertencentes a cada *cluster* encontrado.

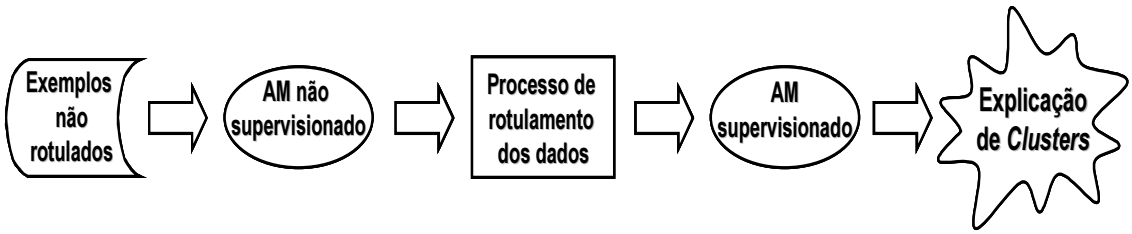

Figura 1: Metodologia

Como pode ser observado, o usuário (especialista) participa do processo após a contrução dos *clusters*, ou previamente, fixando o número de *clusters* a serem encontrados pelo algoritmo de AM não supervisionado. Existe uma outra abordagem, que considera que "é mais fácil criticar do que construir" (Cohn et al., 2000), a qual permite que o usuário iterativamente forneça *feedback* ao algoritmo de *clustering* através de restrições.

3 Clusters *versus* Classes

É importante observar que dois ou mais *clusters* podem agrupar exemplos que referem-se ao mesmo conceito. Isto é ilustrado claramente na Figura 2 que mostra graficamente um conjunto de exemplos de treinamento rotulados com as classes "+" e "–".

Após submeter esses exemplos a um algoritmo de AM supervisionado que induz o conceito através de uma árvore de decisão, por exemplo, as regras geradas seriam do tipo

if $x < a$ and $y < b$ then classe "–"
if $x \geq a$ and $y \geq b$ then classe "–"
if $x < a$ and $y \geq b$ then classe "+"
if $x \geq a$ and $y < b$ then classe "+"

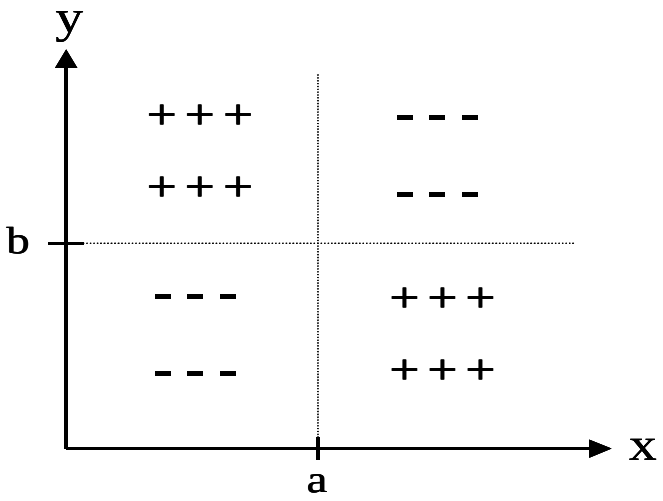

Figura 2: Conjunto de Exemplos de Treinamento

Entretanto, no caso de desconhecer a classe, os exemplos de treinamento são vistos pelo algoritmo de *clustering* como mostra a Figura 3, isto é, apenas como pontos no espaço de busca que podem ser agrupados de acordo com algum critério de similaridade.

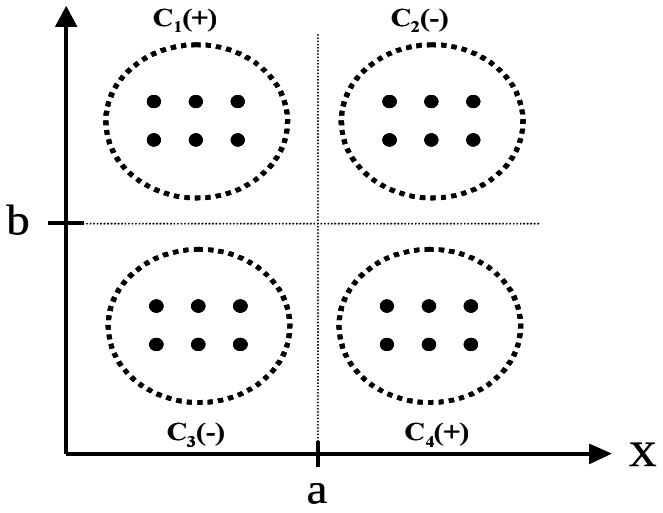

Figura 3: Conjunto de Exemplos de Treinamento Vistos pelo Algoritmo de *Clustering*

Neste caso, o algoritmo de *clustering* encontra 4 *clusters* distintos — C_1, C_2, C_3 e C_4, Figura 4. Porém, neste conjunto de treinamento há apenas dois conceitos que representam as classes "+" (*clusters* 1 e 4) e "–" (*clusters* 2 e 3)

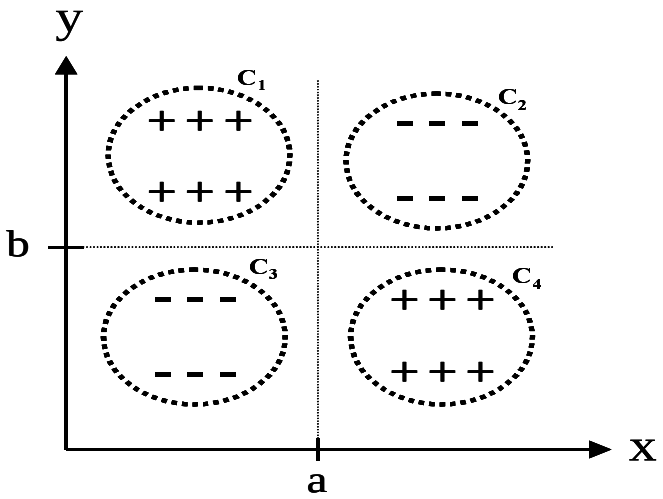

Figura 4: Conjunto de Exemplos de Treinamento Agrupados em *Clusters*

4 Processo Proposto para Explicação de *Clusters*

Nesta seção, é proposto um processo, ilustrado na Figura 5, que utiliza a metodologia descrita na Seção 2, usando o *AutoClass* como algoritmo de AM não supervisionado. *AutoClass* é um algoritmo de *clustering* em que o número de *clusters* pode ser especificado *a priori* ou encontrado automaticamente pelo próprio algoritmo. A saída gerada consiste de vários relatórios com descrições dos *clusters* encontrados e a probabilidade parcial dos exemplos nesses *clusters*. Esses relatórios são utilizados na etapa de rotulamento dos exemplos. O processo de rotulamento dos dados é realizado através de uma ferramenta computacional, por nós implementada, denominada InClass. Essa ferramenta utiliza como entrada um dos relatórios gerados pelo *AutoClass* e o conjunto de exemplos originais (não rotulados). A saída gerada pela ferramenta consiste de uma base de dados que contém os mesmos exemplos que foram processados pelo *AutoClass*, todos eles também listados no relatório gerado, acrescentados de uma dimensão adicional, a qual consiste do rótulo relacionado ao *cluster* ao qual pertence cada exemplo.

Nesse processo foi escolhido o $\mathcal{S}$ee5 como algoritmo de AM supervisionado. $\mathcal{S}$ee5 utiliza tanto regras quanto árvores de decisão como linguagem de descrição de conceitos. Assim, a base de dados gerada por InClass, no formato requerido por $\mathcal{S}$ee5, é processada por este para induzir regras de conhecimento. Tomando como base as regras geradas pelo $\mathcal{S}$ee5, análise de erro e diversas estatísticas, o especialista pode realizar uma análise apurada para tentar explicar o agrupamento dos exemplos nos *clusters* encontrados.

Como descrito na metodologia proposta, o processo finaliza com a análise do especialista sobre as regras de conhecimento geradas. É através dessa análise que pode-se verificar se o conhecimento gerado é importante, desconhecido e útil.

É importante salientar que este é um processo interativo e iterativo. Assim, poderá ser repetido, em qualquer etapa, caso o resultado não seja satisfatório ou mesmo utilizando–se outros algoritmos de AM. A seguir, são descritas resumidamente as características principais dos algoritmos *AutoClass* e $\mathcal{S}$ee5, bem como da ferramenta InClass, utilizados nesse processo.

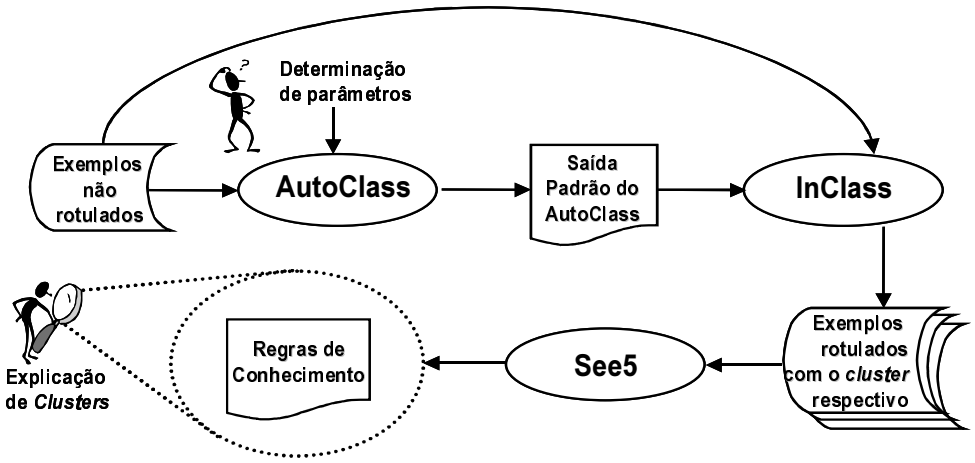

Figura 5: Processo Proposto para Explicação de *Clusters*

4.1 *AutoClass*

AutoClass é um algoritmo de aprendizado não supervisionado baseado na teoria Bayesiana, desenvolvido pelo grupo de Bayes no Ames Research Center. Basicamente, *AutoClass* descreve *clusters* a partir da distribuição probabilística sobre os atributos dos exemplos, considerando que existe independência condicional nos dados.

AutoClass tem sido usado e testado em muitos conjuntos de dados, pela NASA e pela indústria, meio acadêmico e outras agências. Foram encontradas e mostradas algumas classificações surpreendentes, que mostram padrões nos dados muitas vezes desconhecidos para o especialista da área. *AutoClass* é um algoritmo robusto e de domínio público, que apresenta, basicamente, as seguintes características (Cheeseman & Stutz, 1990):

- determina o número de *clusters* automaticamente ou permite que seja definido pelo usuário;
- os valores dos atributos podem ser tanto contínuos quanto discretos;
- manipula valores ausentes e desconhecidos;
- tempo de processamento é robustamente linear;
- gera relatórios descrevendo os *clusters* encontrados e prediz o *cluster* ao qual pertencem novos exemplos.

AutoClass procura a melhor classificação que possa encontrar nos dados. Uma classificação poderá ser a descoberta de um conjunto de *clusters*, descrevendo qual a porcentagem provável dos exemplos estarem em cada *cluster,* e uma denominação probabilística dos exemplos para esses *clusters.* Isto é, para cada exemplo, a probabilidade relativa de ser membro de cada *cluster.* A entrada para o algoritmo consiste de um conjunto de dados na forma atributo-valor, definição de modelos e de parâmetros de busca. O *AutoClass* procura um conjunto de *clusters* que seja altamente provável com os dados e modelos especificados. A saída desse algoritmo é a descrição probabilística dos *clusters* identificados e dos exemplos pertencentes a esses *clusters.* O próprio algoritmo não impõe nenhum limite específico no número de exemplos, mas bases de dados

com mais de 100.000 valores (considerando o número de elementos na tabela atributo-valor que descreve os exemplos) podem necessitar de tempo de execução excessivo.

4.2 $\mathcal{S}$ee5

$\mathcal{S}$ee5 é um produto comercial para plataforma Windows™ que inclui melhorias dos algoritmos $\mathcal{C}$4.5 e $\mathcal{C}$4.5rules (Quinlan, 1993), os quais têm sido usados, freqüentemente, para comparar seu desempenho com outros algoritmos de AM. O $\mathcal{S}$ee5 foi projetado para trabalhar com bases de dados relativamente grandes. Como seus precursores, manipula atributos com valores discretos ou contínuos, induzindo conceitos expressos como árvores de decisão ou conjuntos de regras não ordenadas **if-then** (Baranauskas & Monard, 2000b). Seu desempenho tem se mostrado muito bom na maioria dos casos.

4.3 InClass

Para que o processo de rotulamento dos exemplos seja feito de forma automática, implementamos uma ferramenta computacional denominada InClass, na linguagem de programação PERL[1] (Wall et al., 1996). Como mencionado anteriormente, os dados de entrada para a ferramenta consistem da base de dados original e um dos relatórios gerados pelo *AutoClass*. As informações dos *clusters* encontrados, contidas nesse relatório, estão no formato apresentado na Tabela 1. O cabeçalho que o relatório contém foi omitido nessa tabela.

Case #	Class	Prob	Case #	Class	Prob	Case #	Class	Prob
1	5	0,999	47	7	0,986	93	2	0,992
2	5	0,996	48	21	1,000	94	0	1,000
3	2	0,996	49	4	1,000	95	10	1,000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
46	1	1,000	92	4	0,998	128	0	0,997

Tabela 1: Relatório Gerado pelo *AutoClass*

O relatório é subdividido em 3 colunas, e cada coluna representa a tripla

<Case #, Class, Prob>

onde:

- Case # é um número inteiro que identifica o exemplo;
- Class é o *cluster* ao qual esse exemplo pertence;
- Prob é a probabilidade relativa do exemplo pertencer àquele *cluster*.

Por exemplo, num total de 128 exemplos, na primeira linha do relatório descrito na Tabela 1, estão representados as 3 colunas da seguinte forma:

- o exemplo 1 que pertence ao *cluster* 5 com probabilidade 0,999;
- o exemplo 47 que pertence ao *cluster* 7 com probabilidade 0,986;

[1] Pratical Extraction and Report Language

- o exemplo 93 que pertence ao *cluster* 2 com probabilidade 0,992.

O InClass está programado para processar *strings* através de expressões regulares, que é a característica principal da linguagem PERL. Com base nesse relatório do *AutoClass* e na base de dados original (exemplos não rotulados), o InClass cria uma nova base de dados com os exemplos originais rotulados com o *cluster* respectivo. A Tabela 2 mostra o arquivo final com a nova base de dados gerada pelo InClass, no qual At_i refere–se ao atributo i dos exemplos e $v_{i,j}$ refere–se ao valor do atributo i do exemplo j. Na realidade, o InClass pode considerar, na construção da nova base, todos os exemplos ou apenas exemplos que pertencem, com uma dada probabilidade, aos *clusters* encontrados por *AutoClass*, dependendo dos parâmetros especificados pelo usuário. O Algoritmo 1 apresenta o algoritmo InClass em uma linguagem de alto nível.

Case #	At_1	At_2	At_3	...	At_m	Classe
1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	...	$v_{1,m}$	5
2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	...	$v_{2,m}$	5
3	$v_{3,1}$	$v_{3,2}$	$v_{3,3}$	...	$v_{3,m}$	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
128	$v_{128,1}$	$v_{128,2}$	$v_{128,3}$	...	$v_{128,m}$	0

Tabela 2: Arquivo Final Gerado pelo InClass

Algoritmo 1 InClass

Require: o conjunto original de exemplos e
o relatório gerado pelo *AutoClass* utilizando esse conjunto de exemplos

```
1: procedure InClass
2: Entre com a Probabilidade ou 0 para considerar todos os exemplos
3: for all cluster encontrado in relatório AutoClass do
4:   if Probabilidade ok then
5:     insere a classe como um novo atributo no exemplo correspondente do conjunto original
6:     grave este novo exemplo em um novo arquivo FINAL
7:   end if
8: end for
9: return FINAL
10: end InClass
```

5 Estudo de Caso

O estudo de caso realizado neste trabalho utiliza uma base de dados do mundo real. A base de dados foi fornecida pela Profª Ana Haedo e a Licenciada Noemí Matsudo do grupo de estatística da Universidade de Buenos Aires, Argentina. Essa base de dados foi também analisada utilizando outros métodos (Matsudo, 2000).

5.1 Descrição da Base de Dados

A base de dados contém dados de uma Pesquisa Permanente de Domicílios (EPH[2]), que é um programa nacional da Argentina, conjuntamente com a Administração Estadual de Estatística

[2]Encuesta Permanente de Hogares

(DPE[3]) desse país. Dessa pesquisa, retira-se informação sócio-econômica de aglomerados urbanos do país.

Os dados são coletados através de um questionário familiar com dados da moradia e características demográficas do domicílio, além de um questionário individual com dados de trabalho, renda, educação e migração de cada um dos componentes do domicílio. Os objetivos gerais que sustentam a EPH consistem, sinteticamente, em conhecer e caracterizar a população desde a sua inserção sócio-econômica. Nesse sentido, pretende-se conhecer a situação das pessoas e dos domicílios, por serem estes os núcleos básicos de convivência onde os indivíduos se associam segundo seu lugar na estrutura social.

Em função dos objetivos gerais, a EPH resgata como temática um conjunto de dimensões básicas que pretendem dar respostas aos seguintes conceitos:

- caracterizar a população do ponto de vista demográfico;
- caracterizar a população por sua participação na produção de bens e serviços;
- caracterizar a população por sua participação na distribuição do produto social.

A EPH é uma pesquisa por amostragem, desenvolvida em aglomerados urbanos, ou seja, para conhecer as diversas características do total dos domicílios de um aglomerado urbano se pesquisa uma pequena fração representativa do mesmo. A aplicação rigorosa de técnicas estatísticas permite garantir a precisão dos dados assim obtidos. Por tal motivo, as conclusões que surgem da análise de dados de um aglomerado urbano não devem validar outras áreas geográficas não cobertas pela pesquisa.

Assim, através dessa pesquisa, uma base de dados foi construída originalmente contendo 4.648 exemplos os quais utilizam 35 atributos com informações dos domicílios e 150 atributos com características das pessoas que habitavam os domicílios, totalizando 185 atributos.

5.2 Limpeza nos Dados

Num primeiro momento, em uma limpeza realizada conjuntamente com os especialistas, foram retirados vários atributos irrelevantes. Também foram discretizados e combinados alguns atributos. Para discretizar os atributos, os especialistas realizaram uma análise de maneira que a discretização respeitasse a distribuição dos mesmos. Nesta primeira etapa, dos 35 atributos com informação dos domicílios foram selecionados apenas 3, descritos na Tabela 3. Dos 150 atributos com informações pessoais foram selecionados 37, descritos na Tabela 4. Assim, dos 185 atributos iniciais restaram somente 40 atributos, sendo 9 com valores contínuos (C) e 31 com valores discretos (D).

Campo	Tipo	Descrição
P01	D	Tipo de moradia; 1=casa; 2=apartamento; 3=moradia no lugar que trabalho; 4=inquilinato; 5=hotel ou pensão; 6=moradia não destinada a fins de habitação; 7=moradia em vila; 8=outros
P03	D	Habitações de uso exclusivo do domicílio
P07	D	Regras (regimento) de moradia 1=proprietário da moradia e do terreno; 2=propietário da moradia somente 3=inquilino ou arrendatário da moradia; 4=ocupante com relação de dependência 5=ocupante gratuito; 8=outros

Tabela 3: Dicionário de Dados do Domicílio

[3]Direcciones Provinciales de Estadistica

Campo	Tipo	Descrição
COMPONENTE	D	Número de componentes do domicílio
H08	D	Relação de parentesco 0l=chefe; 02=cônjuge; 03=filho; 04=genro/nora; 05=irmão; 06=neto; 07=cunhado; 08=pai ou sogro; 09=outros familiares; l0=serviço doméstico; ll=outros componentes
H12	C	Anos cumpridos -1=menos de 1 ano; 98=noventa e oito ou mais; 99=ns/nr idade
H13	D	Sexo 1=homem; 2=mulher
H14	D	Estado Civil 1=solteiro; 2=concubinato; 3=casado; 4=separado ou divorciado; 5=viúvo
P01	D	Tem trabalhado durante a semana? 1=sim; 2=não
P02	D	Recebe algum pagamento por seu trabalho? 2=não
P12	D	Quantas ocupações tem
P15T	C	Total hs trab + horas extras sem.de ref
P17	D	É você... 1=patrão ou empregado; 2=trabalhador por conta própria; 3=obreiro ou empregado; 4=trabalhador sem salário
P18	C	O que se faz ou o que se produz no estabelecimento onde trabalha
P18B	D	Tipo de estabelecimento 1=público; 2=privado; 3=outros
P20	C	Nome da ocupação e o tipo de tarefa realizada (ver Clasif. Nac. Ocupac.INDEC)
P21	C	Quanto ganha nessa ocupação
P21D	C	Quantidade de dias que recebe pagamento
P22	C	Quanto tempo faz que está nessa ocupação (anos)...
P22M	C	Quanto tempo faz que está nessa ocupação (meses) ...
P23	D	Nessa ocupação goza dos seguintes benefícios 32=indenização por ser despedido; 08=férias; 04=13o salário; 02=aposentadoria; 16=seguro de trabalho; 0l=outras inclusive obra social; 63=todos os benefícios; 64=sem benefícios
P24	D	Essa ocupação é 1=permanente; 2=um trab.temporário(por prazo fixo, tarefa ou obra); 3=um “bico”; 4=de duração desconhecida(instável); 9=sem especificar
P24_2_M	D	Para p24=2 por quantos meses
P24_D	D	Para p24=2 e p24=3 por quantos dias
P29	D	Busca outra ocupação 1=sim; 2=não
P30	D	Busca trabalho 1=porque ganha pouco; 2=porque está insatisfeito com sua tarefa; 3=porque a relação com o empregador é ruim; 4=porque acredita que não será despedido; (assal.) 5=porque o trabalho que tem vai se acabar; 6=porque tem pouco trabalho (não assal.); 7=por outras causas trabalhistas; 8=por motivos pessoais
P47	D	Recodificação de Rendas de Fonte Trabalhista 1=tem rendas e declara total; 2=não tem rendas; 9=tem rendas e não declara total ou declara parcialmente
P47CAT	D	Totais de rendas
P54	D	Sabe ler e escrever 1=sim; 2=não
P55	D	Freqüentou ou freqüenta a escola 1=frequüenta; 2=freqüentou; 3=nunca freqüentou
P56	D	Que estudo cursa ou cursou (indique somente o nível mais alto alcançado) branco branco=pré-escolar; 01=primário; 02=nacional; 03=comercial; 04=normal; 05=técnico; 06=outro ensino médio; 07=superior; 08=universitário
P58	D	Finalizou o estudo? 1=sim; 2=não
P58B	D	Qual é o último grau ou ano aprovado nesse estudo - pré-escolar se ingressa como branco; - 8vo. do EGB (Educação Geral Básica) se ingressa como lro.nacional; - 9no. do EGB se ingressa como 2do.nacional
P59	D	Onde nasceu

continua na próxima página

continuação da página anterior

Campo	Tipo	Descrição
		1=nesta cidade; 2=em outro lugar desta Província (Estado); 3=em outra Província; 4=em outro país
P59COD	C	Código da província ou país
ITFCAT	D	Quantidade de Rendas Totais Familiares
IPCFCAT	D	Quantidade de Rendas *Per Capita* Familiar
BENEF2	D	ll=somente aposentadoria; 12=combinações com aposentadoria; 13=combinações sem aposentadoria; 14=todos os benefícios; 15=sem benefícios (*); 16=ocupações não assalariados(sem corresp.benef); 17=não tem ocupação (estado#l); (*) Total sem aposentadoria códigos: (13+15)
RAMA	D	1=atividades primárias; 2=Ind.alimentos, bebidas e tabaco; 3=Ind.Têxteis, confecções e calçados; 4=Ind.Prod.químicos e de refinação de petróleo e combustível nuclear; 5=Ind.Prod.metálicos, maquinarias e equipamentos; 6=Outras indústrias manufaturadas; 7=Fornecedor de eletricidade, gás e água; 8=Construção; 9=Comércio grande quantidade; 10=Comércio de pequena quantidade; ll=Restaurantes e Hotéis; 12=Transporte - Serviços Conexos de Transporte e comunic.; 14=Intermediação financeira; 15=Atividades imobiliárias, empresariais e de aluguel; 16=Administração Pública e Defesa; 17=Ensino; 18=Serviços Sociais e de Saúde; 19=Outras Atividades de Serv.Comunitários e sociais 20=Serviços de Reparação; 21=Domicílios privados com serv.doméstico; 22=Outros Serviços pessoais; 89=Novos Trabalhadores; 99=Sem especificar
FUENTE	D	1=somente de trabalho assalariado; 2=somente de trabalho por conta própria; 3=somente de utilidades e benefícios; 4=somente de aluguéis, interesses e dividendos; 5=somente de aposentadoria ou pensão; 6=somente de outras rendas; 7=de trabalho assalariado e de trabalho por conta própria ; 8=de trabalho assalariado e de trabalho por conta própria; 9=de trabalho assalariado e aluguéis, interesses e dividendos; 10=de trabalho assalariado e aposentadoria ou pensão; ll=de trabalho assalariado e outras rendas; 12=de trabalho por conta própria e utilidades e benefícios; 13=de trabalho por conta própria e aluguéis, interesses e dividendos ou pensão; 15=de trabalho por conta própria e outras rendas; 16=de utilidades e benefícios e aluguéis, interesses e dividendos; 17=de utilidades e benefícios e aposentadoria ou pensão; 18=de utilidades e benefícios e outras rendas; 19=de aluguéis, interesses e benefícios e aposentadoria ou pensão; 20=de aluguéis, interesses e benefícios e outros rendas; 21=de aposentadoria, pensão e outras rendas; 22=qualquer combinação de três ou mais fontes; 23=não tem rendas

Tabela 4: Dicionário de Dados Pessoais

Após essa primeira etapa, foi por nós realizada uma última limpeza, para retirar dados com ruídos. Considerando que o *AutoClass* trabalha com distribuição probabilística e muitos atributos com valores contínuos apresentavam valores sem significado, representados por um valor específico, geralmente 9 ou 99, foi necessário substituir estes valores pelo símbolo "?", representando que o valor do atributo está ausente ou é desconhecido.

5.3 Experimentos Realizados

A base de dados foi submetida ao *AutoClass* sem fixar o número de *clusters,* tendo sido encontrados 25 *clusters*, um número relativamente alto. Para uma análise inicial exploratória foi decidido realizar o experimento fixando o número de *clusters* em 2. Logo após, a saída padrão do *AutoClass* foi submetida ao InClass, criando duas bases de dados rotuladas. Uma das bases contém todos os exemplos da base original, enquanto a outra considera apenas os exemplos que pertencem a um desses *clusters* com probabilidade 1. Ambas as bases foram submetidas ao *See5*. A Tabela 5 mostra um resumo dos resultados obtidos, nos quais:

- Clusters - representa os *clusters*, denotado por C(0) e C(1) representando respectivamente os *clusters* 0 e 1;
- P(InClass) - probabilidade que o InClass considera ao criar a nova base de dados com os exemplos rotulados. Valores ≥ 0 indicam que todos os exemplos são considerados. Valores $= 1$ consideram apenas exemplos com probabilidade 1 de pertencer ao *cluster*;
- # Exemplos - número de exemplos na nova base de dados criada por InClass;
- % Classe - porcentagem de exemplos que pertencem a cada *cluster;*
- Erro Aparente - erro aparente de $\mathcal{S}$ee5, isto é, utilizando toda a base de dados como treinamento e teste;
- Erro Verdadeiro (10CV) - erro verdadeiro de $\mathcal{S}$ee5 obtido através de 10k–fold cross–validation;
- Erro CM - erro da classe majoritária;
- # Regras - número de regras induzidas por $\mathcal{S}$ee5 utilizando todos os exemplos;
- # Médio de Regras - número médio de regras induzidas por $\mathcal{S}$ee5 obtido através de 10k–fold cross–validation.

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(0) C(1)	≥ 0	4.648	68,1 31,9	0,2%	$0,4\% \pm 0,2\%$	31,9%	5	$4,8 \pm 0,1$
C(0) C(1)	$= 1$	4.475	67,01 32,98	0,1%	$0,1\% \pm 0,0\%$	32,99%	5	$4,6 \pm 0,2$

Tabela 5: Experimento com 2 *Clusters* - Resumo dos Resultados

Pode-se observar que o poder preditivo utilizando 2 *clusters* é muito bom, apresentando um erro verdadeiro muito baixo e um número pequeno de regras induzidas. Do número total de 4.648 exemplos considerados, somente 273 exemplos ($4.648 - 4.475$, na Tabela 5) pertencem a algum *cluster* com probabilidade menor que 1, o que representa somente 5,9% do total de exemplos. As regras extraídas pelo $\mathcal{S}$ee5 considerando P(InClass) ≥ 0 são mostradas a seguir.

```
Rule 1: (cover 2929)
        P22M <= 0
        P47 = 1
    ->  class classe_0  [0.999]

Rule 2: (cover 236)
        P22M <= 0
        P47 = 9
    ->  class classe_0  [0.992]

Rule 3: (cover 1393)
        P22M > 0
    ->  class classe_1  [0.999]
```

```
Rule 4: (cover 179)
        P47 = 2
    ->  class classe_1  [0.978]

Rule 5: (cover 23)
        P24 = 9
    ->  class classe_1  [0.960]

Default class: classe_1
```

Pode-se observar que a primeira regra cobre mais de 50% dos casos e apenas 3 atributos (P22M, P47, P24) de um total de 40 atributos, foram suficientes para representar todo o conjunto de regras. O número total de exemplos cobertos por essas regras é 4.935 (2.929 + 236 + 1.393 + 179 + 177 + 21). Vale lembrar que $\mathcal{S}$ee5, induz regras não ordenadas, as quais podem cobrir mais de um exemplo. Isto indica que, no máximo, 287 exemplos (4.935 − 4.648) são cobertos por mais de uma regra. Esse número corresponde a 6,2% do total de exemplos considerados. O número que aparece dentro do símbolo "[ ]" é um valor entre 0 e 1 que indica a confiança com a qual essa predição é feita. Uma possível interpretação do conjunto dessas regras é a seguinte:

- todos os exemplos cujo valor do atributo P22M for menor ou igual a 0 **e** cujo valor do atributo P47 é igual a 1 **ou** 9, pertencem a um mesmo *cluster*, nesse caso denominado classe_0. Mais especificamente, todos os indivíduos que não tem ocupação e tem rendas (declarado ou não) pertencem a classe_0 — (Rules 1 e 2);
- todos os exemplos cujo valor do atributo P22M é maior que 0 pertencem ao *cluster* denominado classe_1. Mais especificamente, todos os indivíduos que tem ocupação pertencem a classe_1 — (Rule 3);
- todos os exemplos cujo valor do atributo P47 é igual a 2 pertencem ao *cluster* denominado classe_1. Mais especificamente, todos os indivíduos que não tem rendas pertencem a classe_1 — (Rule 4);
- todos os exemplos cujo valor do atributo P24 é igual a 9 pertencem ao *cluster* denominado classe_1. Mais especificamente, todos os indivíduos que não especificaram o tipo de ocupação pertencem a classe_1 — (Rule 5).

Como pode ser observado, na classe_0 encontram-se todos os indivíduos sem ocupação, porém com rendas. Entretanto, na classe_1 é difícil caracterizar os indivíduos pois eles também podem não ter ocupação, não ter renda, etc. A seguir é mostrado o conjunto de regras para 2 *clusters* com P(InClass) = 1

```
Rule 1: (cover 2784)
        P22M <= 0
        P47 = 1
    ->  class classe_0  [0.999]

Rule 2: (cover 216)
        P22M <= 0
        P47 = 9
```

```
    ->  class classe_0  [0.995]

Rule 3: (cover 1393)
        P22M > 0
    ->  class classe_1  [0.999]

Rule 4: (cover 177)
        P47 = 2
    ->  class classe_1  [0.989]

Rule 5: (cover 21)
        P24 = 9
    ->  class classe_1  [0.957]

Default class: classe_0
```

É interessante observar que no conjunto de regras induzidas pelo $\mathcal{S}$ee5 são semelhantes, nos 2 experimentos para P(InClass) ≥ 0 e P(InClass) $= 1$, diferindo apenas no número de exemplos cobertos e a confiança em cada regra.

Considerando que o objetivo deste trabalho é "clusterizar" os dados de forma que estes reflitam alguma informação semântica, decidiu–se realizar o experimento novamente, agora fixando o número de *clusters* em 3. A Tabela 6 apresenta um resumo dos resultados obtidos.

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(0)			45,65					
C(1)	≥ 0	4.648	31,84	0, 2%	$0,2\% \pm 0,1\%$	54, 35%	8	$7,4 \pm 0,2$
C(2)			22,50					
C(0)			45,49					
C(1)	$= 1$	4.588	32,17	0,4%	$0,3\% \pm 0,1\%$	54, 51%	5	$5,1 \pm 0,1$
C(2)			22,34					

Tabela 6: Experimento com 3 *Clusters* - Resumo dos Resultados

Pode ser observado que o número de exemplos que pertencem a um *cluster* com probabilidade 1 é ainda maior que no experimento anterior com 2 *clusters*. Isto significa que os *clusters* estão melhor separados. Na realidade, somente um exemplo pertence ao *cluster* 0 e 1 simultaneamente, denotado por C(0,1), e dois exemplos pertencem aos *clusters* 1 e 2 simultaneamente, denotada por C(1,2) na Tabela 7 e mostrado graficamente na Figura 6. *Clusters* que não tem sobreposição não são apresentados na tabela, ou seja, o *cluster* 0 e 2 não contém nenhum exemplo sobreposto. Observa–se também que o número de regras geradas por $\mathcal{S}$ee5 é pequeno, com um erro verdadeiro muito baixo. Assim, pode–se afirmar que o poder preditivo utilizando 2 ou 3 *clusters* é muito bom.

Clusters Sobrepostos	# Exemplos
C(0,1)	1
C(1,2)	2

Tabela 7: Sobreposição de *Clusters* - 3 *Clusters*

Ainda com o intuito de auxiliar o especialista, é possível obter a informação resumida apresentada na Tabela 8, que mostra o conjunto de atributos presentes nas regras induzidas pelo $\mathcal{S}$ee5

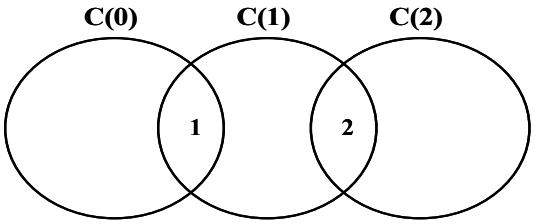

Figura 6: Experimento com 3 *Clusters*: Número de Exemplos Sobrepostos

para descrever as três classes (*clusters*). O símbolo "•" refere-se à presença do atributo na regra para P(InClass) ≥ 0, enquanto que o símbolo "◦" refere-se à P(InClass) $= 1$.

	# Atributos				
	01	02	03	04	05
C(0)	• ◦	• ◦	•		
C(1)	• ◦	•	• ◦	•	• ◦
C(2)	• ◦	• ◦			

Tabela 8: Conjunto de Atributos Presentes nas Regras - 3 *Clusters*

As regras extraídas pelo $\mathcal{S}$ee5 para 3 *clusters* são mostradas a seguir.

```
Rule 1: (cover 1940)
        P22M <= 0
        P47 = 1
        P59COD <= 0
    ->  class classe_0  [0.999]

Rule 2: (cover 180)
        P22M <= 0
        P47 = 9
        P59COD <= 0
    ->  class classe_0  [0.995]

Rule 3: (cover 1393)
        P22M > 0
    ->  class classe_1  [0.999]

Rule 4: (cover 53)
        P47 = 2
        P59COD > 0
    ->  class classe_1  [0.982]

Rule 5: (cover 179)
        P47 = 2
    ->  class classe_1  [0.978]
```

```
Rule 6: (cover 23)
        P29 = 9
    ->  class classe_1  [0.960]

Rule 7: (cover 23)
        P24 = 9
    ->  class classe_1  [0.920]

Rule 8: (cover 1059)
        P22M <= 0
        P59COD > 0
    ->  class classe_2  [0.985]

Default class: classe_1
```

Nota–se que foram induzidas 8 regras para cobrir os exemplos e novos atributos foram utilizados no conjunto de regras para 3 *clusters* – P59COD e P29. Porém, quando é utilizado P(InClass) = 1 o número de regras induzidas diminui de 8 para 5, as quais são listadas a seguir.

```
Rule 1: (cover 2132)
        P22M <= 0
        P59COD <= 0
    ->  class classe_0  [0.978]

Rule 2: (cover 1393)
        P22M > 0
    ->  class classe_1  [0.999]

Rule 3: (cover 178)
        P47 = 2
    ->  class classe_1  [0.983]

Rule 4: (cover 21)
        P24 = 9
    ->  class classe_1  [0.957]

Rule 5: (cover 1040)
        P22M <= 0
        P59COD > 0
    ->  class classe_2  [0.985]

Default class: classe_1
```

A mesma experiência foi realizada fixando o número de *clusters* em 5. O resumo dos resultados obtidos encontra–se nas Tabelas 9, 10 e 11. Pode ser observado que o erro verdadeiro de $\mathcal{S}$ee5 para 5 *clusters* continua baixo, mas o número de regras para exemplos com probabilidade ≤ 1 pertencentes aos *clusters* incrementa consideravelmente.

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(0)			35,20					
C(1)			24,18					
C(2)	≥ 0	4.648	20,25	2,3%	$4,3\% \pm 0,4\%$	64,80%	36	$37,3 \pm 1,2$
C(3)			10,46					
C(4)			9,92					
C(0)			24,64					
C(1)			35,02					
C(2)	$= 1$	3.007	13,57	0,1%	$0,6\% \pm 0,1\%$	64,98%	17	$15,4 \pm 0,6$
C(3)			12,70					
C(4)			14,07					

Tabela 9: Experimento com 5 *Clusters* - Resumo dos Resultados

Clusters Sobrepostos	# Exemplos
C(0,1)	168
C(0,1,2)	88
C(0,2)	282
C(1,2)	21
C(1,4)	2
C(2,3)	215
C(2,3,4)	2
C(2,4)	11
C(3,4)	10

Tabela 10: Sobreposição de *Clusters* - 5 *Clusters*

	# Atributos																				
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
C(0)	• ○	• ○	• ○		•	• ○	•	•	•	•	• ○	•	•							○	
C(1)	• ○	• ○	• ○		•	•			•		○	•		•	○					○	○
C(2)	• ○	• ○	• ○		•	• ○	•	•	•		• ○	•	•		• ○	•	• ○	•	• ○	•	○
C(3)	• ○	• ○	•								•				•	•	• ○	•	• ○		
C(4)		• ○	• ○																• ○		

Tabela 11: Conjunto de Atributos Presentes nas Regras - 5 *Clusters*

A Tabela 10 descreve detalhadamente os *clusters* que estão sobrepostos e número de exemplos comuns, mostrado graficamente na Figura 7. Vale ressaltar que a Figura 7 não representa intersecção de conjuntos, portanto, a sobreposição dos *clusters* 0 e 1, por exemplo, vai além de 168 exemplos se considerarmos que existe a sobreposição dos mesmos nos *clusters* 0, 1 e 2. Em outras palavras, 88 exemplos pertencem à intersecção entre os *clusters* 0, 1 e 2 e, além disso, mais 168 exemplos pertencem ao mesmo tempo aos *clusters* 0 e 1, mas não ao *cluster* 2.

Considerando os *clusters* com os maiores números de exemplos sobrepostos — C(0,1), C(0,2) e C(2,3) — os quais representam uma proporção aproximada[4] de 6,08% para C(0,1), 10,94% para C(0,2) e 15,06% para C(2,3), uma tentativa válida é agrupar esses *clusters* em um único. Assim, a sobreposição dos *clusters* com a maior proporção de exemplos comuns, que neste caso é o *cluster* 2 e 3, foram agrupados em um único *cluster* denotado por C(2-3). Após renomear os

[4]Por exemplo, a proporção aproximada para C(0,1) é obtida considerando que os *clusters* 0 e 1 contém 59,38% dos exemplos — 35,20% para C(0) + 24,18% para C(1). Estes 59,38% representam aproximadamente 2.760 exemplos dos 4.648 exemplos do conjunto de treinamento. Assim, 168 exemplos sobrepostos representam proporcionalmente 6,08% dos exemplos para os clusters 0 e 1.

clusters 2 e 3 para um único *cluster* — C(2-3) — a base de dados foi fornecida ao $\mathcal{S}$See5 obtendo os resultados mostrados nas Tabelas 12 e 13.

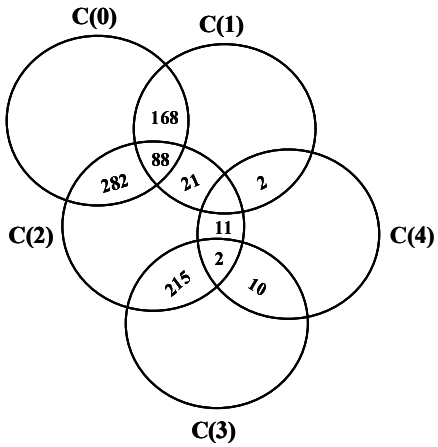

Figura 7: Experimento com 5 *Clusters*: Número de Exemplos Sobrepostos

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(2-3)	≥ 0	4.648	30,71	2,3%	$3,6\% \pm 0,4\%$	64,80%	21	$26,6 \pm 1,8$

Tabela 12: Experimento com C(2-3) para os 5 *Clusters* originais - Resumo dos Resultados

Pode ser observado que agrupando os *clusters* 2 e 3, não foi necessário utilizar todos os atributos para descrever as regras, como mostrado na Tabela 13. No caso, os atributos 15, 16, 18 e 21, não estão presentes nas regras, e o número de regras diminuiu.

	# Atributos																				
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
C(0)	•	•	•			•	•	•	•	•	•	•	•								
C(1)	•	•	•			•			•			•		•							
C(2-3)	•	•	•		•	•	•	•	•		•	•					•			•	
C(4)	•		•																•		

Tabela 13: Conjunto de Atributos com C(2) e C(3) Agrupados - 5 *Clusters*

Novamente, decidiu–se realizar novo experimento, agora fixando o número de *clusters* em 10. Os resultados obtidos encontram–se nas Tabelas 14, 15 e 16. O erro verdadeiro obtido pelo $\mathcal{S}$See5, apesar de ter aumentado, ainda continua baixo, como pode ser observado na Tabela 14. Porém, o número de regras aumentou significativamente.

Mesmo assim, no geral, o resultado pode ser considerado bom. Nota–se, porém, que a sobreposição dos *clusters* também aumentou, bem como o número de atributos utilizados na descrição das regras mostrados nas Tabelas 15 e 16.

Considerando também para 10 *clusters* a maior proporção aproximada dos exemplos sobrepostos (14,59% para C(5,7), 13,49% para C(0,2) e 12,74% para C(3,4)), foi realizado novo experimento cujos resultados são mostrados nas Tabelas 17 e 18. Esses *clusters* com a maior proporção de exemplos sobrepostos foram agrupados e considerados como classes C(0-2), C(3-4), C(5-7). A base de dados para $\mathcal{S}$See5 foi alterada com essas novas classes e a seguir foram realizados os

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(0)			22,35					
C(1)			12,84					
C(2)			9,85					
C(3)			9,92					
C(4)			9,32					
C(5)	≥ 0	4.648	8,43	$3,6\%$	$7,1\% \pm 0,5\%$	$77,65\%$	96	$90,5 \pm 2,5$
C(6)			8,18					
C(7)			7,04					
C(8)			6,26					
C(9)			5,81					
C(0)			18,62					
C(1)			15,34					
C(2)			9,62					
C(3)			8,48					
C(4)			10,99					
C(5)	$= 1$	2.712	7,60	$1,1\%$	$2,7\% \pm 0,5\%$	$81,38\%$	43	$41,2 \pm 0,8$
C(6)			13,20					
C(7)			6,86					
C(8)			1,14					
C(9)			8,15					

Tabela 14: Experimento com 10 *Clusters* - Resumo dos Resultados

Clusters Sobrepostos	# Exemplos	Clusters Sobrepostos	# Exemplos
C(0,2)	202	C(3,9)	5
C(0,7)	121	C(4,6)	3
C(0,7,8)	23	C(4,6,9)	1
C(0,8)	99	C(4,9)	9
C(1,2)	122	C(5,7)	105
C(1,7)	12	C(5,7,8)	3
C(0,2,8)	11	C(5,8)	66
C(1,2,7)	1	C(5,9)	11
C(2,9)	19	C(6,9)	10
C(3,4)	114	C(7,8)	24
C(3,4,9)	4	C(7,8,9)	1

Tabela 15: Sobreposição de *Clusters* - 10 *Clusters*

experimentos, listados abaixo, que correspondem a respectiva linha da Tabela 17, que mostra o resumo dos resultados. Os experimentos realizados com as novas classes foram:

1. a base de dados foi modificada considerando somente os *clusters* 0 e 2 como um único *cluster* denotado por C(0-2) — total 9 *clusters*;
2. idem ao anterior mas considerando somente os *clusters* 3 e 4 como um único *cluster* denotado por C(3-4) — total 9 *clusters*;
3. idem ao anterior mas considerando somente os *clusters* 5 e 7 como um único *cluster* denotado por C(5-7) — total 9 *clusters*;
4. idem aos anteriores mas agora considerando os três *clusters* , C(0-2), C(3-4) e C(5-7) — total 7 *clusters*;

Observa–se nessa tabela que foram quatro experimentos realizados: um experimento para cada novo agrupamento de classe e considerando todos os agrupamentos de classes em um único experimento — C(0-2),C(3-4),C(5-7).

No último experimento com 7 *Clusters*, o erro verdadeiro e o número de regras diminuem consideravelmente. A Tabela 18 mostra os atributos que estão presentes nas regras induzidas

	# Atributos																													
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
C(0)	• ○	• ○	• ○			○			•		•		•			•	• ○					•	•	•	•	•				
C(1)	• ○	• ○	• ○			•	•		○		• ○	•		•	•	•	• ○	•				•				•	• ○			
C(2)	• ○	• ○	• ○		○	• ○			•		• ○	•	•	•	○	•	• ○			•	•	•			•	•				
C(3)	• ○	• ○	•						• ○	•			•					• ○	• ○											
C(4)	• ○	• ○	•							•		•	•		• ○			• ○	• ○			•						•		
C(5)	• ○	• ○	•	•		• ○				•	•	•		•		• ○	• ○			•		•					• ○			
C(6)	• ○	•							•						• ○				○											
C(7)	• ○	• ○	• ○	•	•	• ○			○		• ○				•	• ○	• ○			•	•		•				• ○		• ○	
C(8)	• ○	• ○	•			• ○	•		•		•			•		• ○	• ○			•							• ○			
C(9)	• ○	• ○	• ○		○	• ○			•			•	•		•	•	• ○	• ○	• ○	•		•				•			•	•

Tabela 16: Conjunto de Atributos Presentes nas Regras - 10 *Clusters*

Clusters	P (InClass)	# Exemplos	% Classe	Erro Aparente	Erro (10CV) Verdadeiro	Erro CM	# Regras	# Médio Regras
C(0-2)			33,20	3, 2%	$6,0\% \pm 0,3\%$		91	$78,5 \pm 3,2$
C(3-4)	≥ 0	4.648	19,24	3, 2%	$6,4\% \pm 0,3\%$	49, 82%	77	$69,8 \pm 1,6$
C(5-7)			15,47	3, 3%	$6,1\% \pm 0,1\%$		94	$82,6 \pm 2,5$
C(0-2),C(3-4),C(5-7)			50,18	2, 6%	$4,9\% \pm 0,4\%$		59	$55,0 \pm 1,8$

Tabela 17: Experimento Agrupando os 10 *Clusters* originais em 4 Conjuntos de *Clusters* - Resumo dos Resultados

nesse quarto experimento. Percebe–se que ainda que as novas regras induzidas utilizam os três novos atributos 31, 32 e 33 que não foram usados anteriormente por See5— Tabela 16 pg. 19, 8 atributos utilizados antes do agrupamento, especificamente os atributos 04, 05, 09, 23, 24, 25, 28 e 30, não foram necessários para descrever as novas regras.

	# Atributos																														
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
C(0-2)	•	•	•			•				•	•	•		•	•	•	•			•		•				•						
C(1)	•	•	•			•	•					•		•	•		•	•				•				•	•					
C(3-4)	•																															
C(5-7)	•	•	•			•					•			•	•	•	•			•	•						•				•	•	•
C(6)	•	•													•																		
C(8)	•	•	•								•				•	•	•			•							•						•
C(9)	•	•	•			•						•	•		•	•	•		•	•		•				•			•		•		

Tabela 18: Conjunto de Atributos com C(0) e C(2), C(3) e C(4), C(5) e C(7) Agrupados - 10 *Clusters*

Com esses resultados é possível para o especialista do domínio realizar uma análise semântica, gerando assim uma "explicação" para os *clusters* encontrados. É importante ressaltar que podem ser obtidas várias informações em cada etapa do processo, através dos relatórios gerados tais como os apresentados nas tabelas descritas anteriormente. Mas a análise manual dos resultados contidos nesses relatórios nem sempre é uma tarefa fácil e é importante o desenvolvimento de ferramentas computacionais para auxiliar o especialista (Baranauskas & Monard, 2000a).

6 Conclusões

Em um processo de mineração de dados, a interação com o especialista é fundamental, visto que este detém o conhecimento do domínio. Neste trabalho, propusemos uma metodologia que pode auxiliar o especialista no processo de descobrir conhecimento de dados não rotulados através de algoritmos de AM simbólicos. A tarefa do aprendizado não supervisionado no qual os dados são agrupados em *clusters* de acordo com algum critério de similaridade, é importante e bastante utilizada nesse contexto. Vale ressaltar que, conceitualmente, descrever *clusters* não significa, necessariamente, descrever conceitos. Portanto, muitas vezes, é de interesse do especialista dar uma explicação semântica aos *clusters* descobertos para melhor interpretação dos resultados.

A metodologia proposta consiste, basicamente, em encontrar *clusters* que são inseridos na base de exemplos como um novo atributo, rotulando os dados iniciais, possibilitando assim o uso de algoritmos de AM simbólicos para gerar regras que tentam "explicar" esses *clusters*. O processo que utiliza a metodologia apresentada neste trabalho gera informações que vão além de simplesmente encontrar *clusters* em dados não rotulados, tais como sobreposição de *clusters*, atributos presentes nas regras induzidas, número de regras, o erro verdadeiro, etc. O estudo de caso apresentado utilizando essa metodologia, revelou bons resultados, mostrando que é possível e interessante o método proposto. O conhecimento gerado poderá, então, auxiliar o especialista do domínio na "interpretação" ou "explicação" dos *clusters* encontrados bem como agrupar *clusters* diferentes mas que referem–se a um mesmo conceito.

Referências

Baranauskas, J. A. & Monard, M. C. (2000a). An environment for rule extraction and evaluation from databases. In *Proceedings of the International Joint Conference IBERAMIA/SBIA (2000)*, pages 187–96, Atibaia, SP, Brazil.

Baranauskas, J. A. & Monard, M. C. (2000b). Reviewing some machine learning concepts and methodos. Technical Report 102, ICMC-USP. Disponível em: `ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/Rt_102.ps.zip`.

Cheeseman, P. & Stutz, J. (1990). Bayesian classification (autoclass): Theory and results advances in knowledge discovery and data mining. Disponível em: `http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass-c-program.html`.

Cohn, D., Caruana, R., & McCallum, A. (2000). Semi-supervised clustering with user feedback. *AAAI*.

Decker, K. M. & Focardi, S. (1995). Technology overview: A report on data mining. Technical report, CSCS-ETH, Swiss Scientific Computing Center.

Martins, C. A. & Monard, M. C. (2000). Interpretação de clusters utilizando aprendizado de máquina simbólico. *21 Iberian Latin American Congress on Computational Methods in Engineering - CILAMCE2000*, page 13.

Matsudo, N. (2000). Data mining - desde una perspectiva informática y estadistica. Tesis de Licenciatura en Ciencias de la Computación.

McCallum, A., Nigam, K., & Ungar, L. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. *KDD*.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. MK, Los Altos, California,USA.

Wall, L., Christiansen, T., & Schwartz, R. L. (1996). *Programming in PERL*. O'Reilly,Inc.