PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-Graduação em Informática

RAIMA: Proposta de um Método baseado em Regras de Associação para Identificação do Mecanismo de Ausência em Bases de Dados

Luciana Otávia Silva

Belo Horizonte
2009

Luciana Otávia Silva

RAIMA: Proposta de um Método baseado em Regras de Associação para Identificação do Mecanismo de Ausência em Bases de Dados

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Informática.

Orientador: Prof. Luis Enrique Zárate

Belo Horizonte
2009

FICHA CATALOGRÁFICA

Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

S586r

Silva, Luciana Otávia
RAIMA: proposta de um método baseado em regras de associação para identificação do mecanismo de ausência em bases de dados. / Luciana Otávia Silva. – Belo Horizonte, 2009.
109f. : il.

Orientador: Luis Enrique Zárate.
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-graduação em Informática.
Bibliografia.

1.Banco de dados – Teses. 2. Mineração de dados (Computação). I. Zárate, Luís Enrique. II. Pontifícia Universidade Católica de Minas Gerais. III. Título

CDU: 681.3.011

Bibliotecário: Fernando A. Dias – CRB6/1084

PUC Minas
Programa de Pós-graduação em Informática

FOLHA DE APROVAÇÃO

"RAIMA: Proposta de um Método Baseado em Regras de Associação para a Identificação do Mecanismo de Ausência em Bases de Dados"

Luciana Otávia Silva

Dissertação defendida e aprovada pela seguinte banca examinadora:

Prof. Luis Enrique Zárate Galvez - Orientador (PUC Minas)
Doutor em Engenharia Metalúrgica e de Minas - UFMG

Prof. Clodoveu Augusto Davis Junior - UFMG
Doutor em Ciência da Computação - UFMG

Prof. Mark Alan Junho Song - Orientador (PUC Minas)
Doutor em Ciência da Computação - UFMG

Belo Horizonte, 15 de junho de 2009.

Anel Rodoviário – KM 23,5, Rua Walter Ianni, 255 – São Gabriel / Belo Horizonte – MG / CEP: 31980-110
Fone: (31) 3439-5204 – Fax: (31) 3439-5204 – e-mail: mestradoinfsec@pucminas.br

Ao meu amado filho Matheus, que teve que aceitar
a minha ausência para a realização deste sonho.

AGRADECIMENTOS

Agradeço a Deus, por ter permitido que eu chegasse até aqui. “Porque dele, e por ele, e para ele, são todas as coisas; glória, pois, a ele eternamente. Amém.” Romanos 11:36.

Agradeço aos meus amados pais, pelo amor incondicional e por tudo que me ensinaram.

Agradeço ao meu amado filho Matheus, pelo seu carinho e sua alegria. – Matheus, terminei o que parecia não ter fim. Desculpe-me pelas inúmeras horas que deixamos de compartilhar juntos.

Agradeço aos meus queridos irmãos: Sil, Vivi e Tavinho, pelo incentivo e apoio.

Agradeço a querida Tia Realina, pela presença e atenção.

Agradeço Prof. Zárate, pelas ponderações, pelos conselhos, pela paciência, confiança e principalmente pelo incentivo nas horas mais difíceis.

Agradeço a todos, que de alguma forma me ajudaram a chegar até aqui.

A todos, o meu muito obrigado!

RESUMO

Dado ausente é um dos problemas mais significativos encontrados nos projetos de mineração de dados e descoberta de conhecimento. Diversos métodos já foram propostos para lidar com dados ausentes. Esses métodos são relacionados ao mecanismo de ausência. Entretanto, a identificação do mecanismo de ausência não é trivial, pois em determinadas situações os mecanismos de ausência se confundem. Este trabalho apresenta um método baseado em regras de associação que auxilia na identificação do mecanismo de ausência. Uma identificação precisa do mecanismo de ausência é crucial para a seleção de métodos apropriados para lidar com dados ausentes e consequentemente obtenção de melhores resultados nos processos de mineração de dados e descoberta de conhecimento. Os resultados mostram que o método proposto é consistente e alcançou desempenho muito satisfatório em base de dados com baixa e média correlação de atributos.

Palavras-chave: Dados ausentes. Mecanismo de dados ausentes. Regras de associação.

ABSTRACT

Missing data is one of the most significant problems encountered in real world data mining and knowledge discovery projects. Many approaches have been proposed to deal with the missing values problem. These approaches are related to missing data mechanism. However, it can be difficult to detect these mechanisms since in some situations they are not clear. This work presents a method based on association rules that assists in identifying the missing data mechanism. Precise identification of the missing mechanism is crucial to the selection of appropriate methods for dealing with missing data and thus obtains better results in the processes of data mining and knowledge discovery. The results show that the proposed method is consistent and has achieved very satisfactory performance on databases with low and medium attributes correlation.

Key-words: Missing data. Missing data mechanism. Association rules.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo tem por objetivo introduzir o trabalho desenvolvido, apresentar as hipóteses, a justificativa para o desenvolvimento do mesmo, a motivação, os objetivos a serem alcançados e as principais contribuições.

1.1 O problema de valores ausentes em KDD

A descoberta de conhecimento em bancos de dados (KDD-*Knowledge Discovery in Databases*), onde a mineração de dados (*Data Mining*) está inserida, tem como objetivo extrair a partir de dados padrões implícitos, desconhecidos e potencialmente úteis (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996a).

Na etapa de preparação de dados é necessário um esforço considerável para projetar e implementar procedimentos tais como limpeza, análise de *outliers*, transformação de dados, redução de dimensionalidade e tratamento de dados ausentes. Um dos grandes problemas enfrentados na maioria dos bancos de dados sobre domínios reais é a ocorrência de dados ausentes que podem ocorrer em um atributo, em vários atributos, em várias instâncias ou de forma aleatória.

O valor ausente pode ser chamado também de valor não observado ou valor desconhecido. Este dado pode ser ocasionado por circunstâncias não controladas, em que seu valor não foi adicionado à base de dados, mas existe um valor no domínio sobre o qual está inserido, ou seja, existe um valor para o atributo no mundo real (PYLE, 1999). Em algumas situações, os atributos não foram informados devido a uma questão de privacidade ou recusa. Algumas vezes é apenas consequência de certas medidas não estarem disponíveis no momento da coleta. Já em outras situações, os valores se perderam devido às falhas que ocorrem em sistemas de medidas manuais ou automatizados. Dentro do processo KDD e do *Data Mining,* a ocorrência de dados ausentes é extremamente prejudicial, podendo levar a padrões, conclusões ou tomadas de decisões equivocadas.

A ocorrência de valores ausentes em bases de dados normalmente obedece a um mecanismo que aponta as condições de geração dos dados ausentes.

Uma importante contribuição, proposta em (RUBIN, 1976), foi a classificação dos mecanismos de valores ausentes*: Missing Completely at Random* (MCAR), *Missing at Random* (MAR) e *Not Missing at Random* (NMAR). Essa classificação é baseada nas condições que provocaram a ausência dos dados. No mecanismo MCAR, os valores ausentes estão distribuídos aleatoriamente, ou seja, a probabilidade de encontrar um valor ausente é a mesma para qualquer valor do atributo. Nos outros mecanismos, os valores ausentes não estão aleatoriamente distribuídos, implicando que a probabilidade de encontrar um valor ausente pode depender de outro valor de atributo, no caso MAR, ou ainda, depender do próprio valor ausente, no caso do mecanismo NMAR.

A formalização dos mecanismos de ausência é descrita na Seção 2.1. De forma a proporcionar um melhor entendimento e mostrar a dificuldade de lidar com dados ausentes, a próxima Seção apresenta um experimento que gerou artificialmente valores ausentes para ilustrar os mecanismos de ausência.

1.2 Identificação e caracterização do problema

De uma forma geral, não se pode afirmar de antemão que o mecanismo de ausência é MCAR, MAR ou NMAR. Não há técnicas específicas para determinar qual o mecanismo de ausência de um determinado conjunto de dados. Nesta Seção é apresentada de forma detalhada a definição do problema a ser tratado neste trabalho e demonstrado que os mecanismos de ausência se confundem em determinadas situações.

O experimento partiu do trabalho de SOARES (2007), o qual sugere uma técnica para verificação do mecanismo MCAR pelo estudo comparativo da média e desvio padrão de duas amostras, casos completos e casos incompletos.

O experimento utilizou uma amostra de 117649 registros, extraída da base de dados de um processo siderúrgico (processo de laminação a frio). A base de dados possui os seguintes atributos: espessura de entrada do material a ser laminado, espessura de saída, tensão a ré, tensão à frente, tensão de escoamento do material, coeficiente de atrito, e carga de laminação necessária para deformar o material. A variável dependente é o atributo carga de laminação, que para a amostra selecionada possui o valor mínimo de 976 kgf/mm, o valor máximo 2930 kgf/mm, média 1755,5 kgf/mm, e desvio padrão 320. Esta base de dados possui dados consistentes gerados a partir de um modelo matemático (ZÁRATE, 1998).

A distribuição do atributo carga de laminação com os dados completos é visualizada no Gráfico 1. Os dados ausentes foram gerados artificialmente para a simulação dos mecanismos MCAR, MAR e NMAR e são apresentados nos gráficos de 2 até 16. Os gráficos foram gerados através do software R (R Development Core Team, 2007) e representam os histogramas de frequência do atributo carga de laminação.

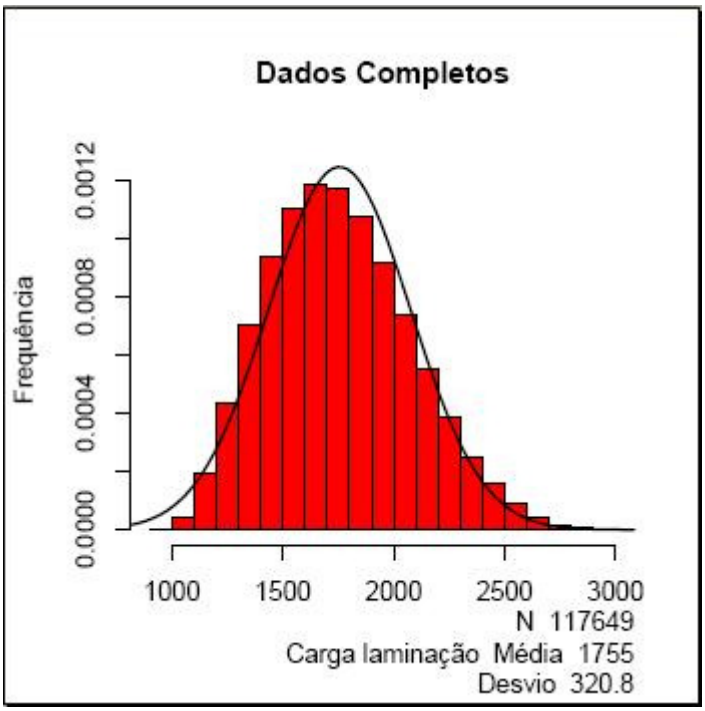


Gráfico 1 – Dados completos

1.2.1 Mecanismo MCAR

O mecanismo MCAR foi simulado pela exclusão aleatória dos dados. A primeira eliminação retirou 5% dos dados e a última 95%. No total 19 experimentos foram efetuados, e para cada experimento a amostra manteve suas características, conforme Tabela 1.

Dessa forma, para o mecanismo MCAR (distribuição aleatória dos valores ausentes), mesmo que o percentual de ausência seja alto (ver Gráfico 2), a amostra mantém suas características originais.

Tabela 1

Resultado Mecanismo MCAR

Número do Experimento	Ausência	Número de registros	Média	Desvio
1	5%	111767	1756	320,8
2	10%	105884	1755	320,5
3	15%	100002	1755	320,8
4	20%	94119	1756	321,2
5	25%	88237	1755	321,4
6	30%	82354	1756	320,9
7	35%	76472	1755	320,5
8	40%	70589	1754	320,2
9	45%	64707	1755	319,8
10	50%	58824	1757	321,6
11	55%	52942	1755	321,1
12	60%	47060	1755	321,5
13	65%	41177	1757	321,1
14	70%	35295	1755	320,2
15	75%	29412	1755	320,5
16	80%	23530	1757	320,5
17	85%	17647	1758	319,4
18	90%	11765	1758	319,7
19	95%	5882	1753	319

Total de registros dados completos: 117649, média 1755,5 kgf/mm

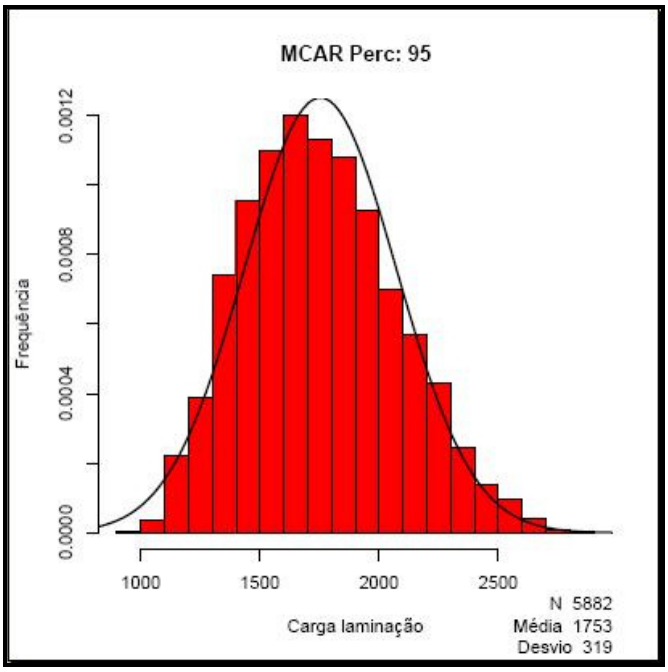


Gráfico 2 – MCAR 95% ausência

1.2.2 Mecanismo MAR

Para o mecanismo MAR, as amostras com dados ausentes foram simuladas através de condições impostas à variável atrito. As ausências na variável carga de laminação foram geradas artificialmente por quatro condições de eliminação de registros. Os quatro experimentos para o mecanismo MAR são sumarizados na Tabela 2.

Tabela 2

Resultado Mecanismo MAR

Número do Experimento	Condição	Número de registros	Média	Desvio
1	atrito >= 0,12	67228	1847	327,4
2	atrito < 0,12	50421	1634	266,9
3	0,1<= atrito <= 0,13	67228	1720	293,2
4	atrito <= 0,1 ou atrito >= 0,13	50421	1802	348,9

Total de registros dados completos: 117649, média 1755,5 kgf/mm

Os gráficos 3, 4, 5 e 6 representam as condições utilizadas para o mecanismo MAR. Para cada experimento MAR nota-se uma variação nas características da amostra, diferentemente dos experimentos observados para o mecanismo MCAR.

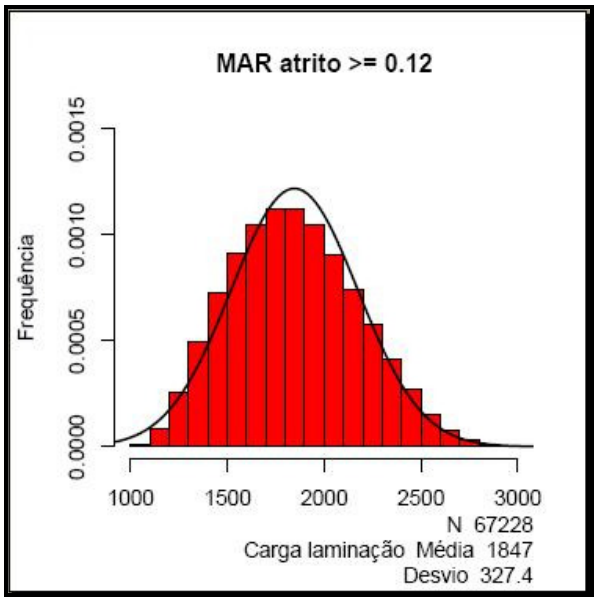


Gráfico 3 – MAR condição (atrito >= 0,12)

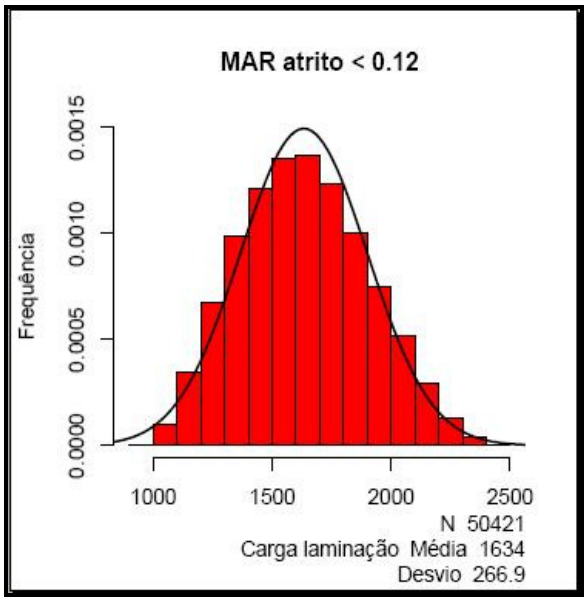


Gráfico 4 – MAR condição (atrito < 0,12)

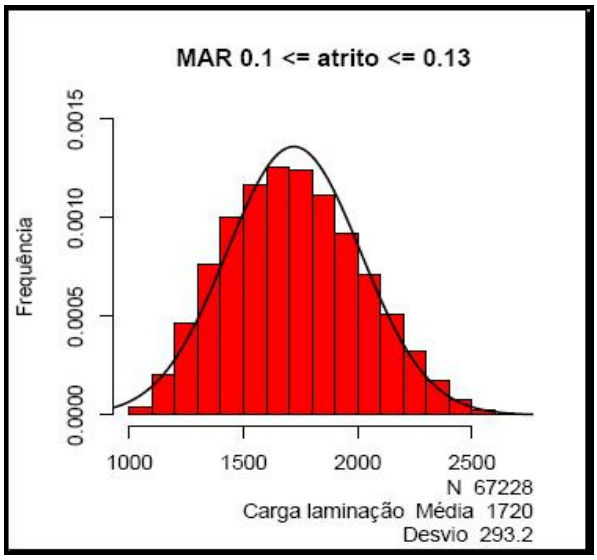


Gráfico 5 – MAR
condição (0,1 <= atrito <= 0,13)

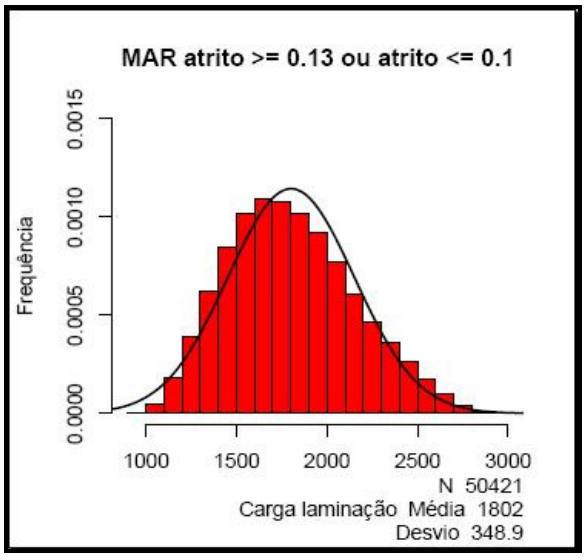


Gráfico 6 – MAR
condição (atrito<=0,1 ou atrito >= 0,13)

1.2.3 Mecanismo NMAR

O mecanismo NMAR é bem caracterizado quando todos os registros de uma determinada condição não estão presentes. As condições para simular dados ausentes para o mecanismo NMAR são apresentadas na Tabela 3.

Ao tornar ausentes todos os casos que obedecem a uma condição, nota-se claramente a influência do dado ausente na amostra, conforme Gráfico 16. Em problemas reais, as bases de dados não apresentam essa característica de ausência (totalidade de dados ausentes para a condição). A ausência pode ocorrer em porcentagens menores de dados ausentes. Para demonstrar essa situação o experimento utilizou percentuais diferentes de dados ausentes. O percentual inicial de dados ausentes foi de 10% e o percentual final de 100% para cada condição, conforme demonstrado na Tabela 3.

Os gráficos 7-8, 9-10, 11-12, 13-14, 15-16, mostram os histogramas para as condições de 1 a 5, com 30% e 90% de ausência, respectivamente.

Para cada condição nota-se a preservação das características da amostra até 30% de dados ausentes. Quando o percentual de dados ausentes ultrapassa 30% são visíveis as distorções nas características da amostra, conforme gráficos 8, 10, 12, 14, e 16.

Tabela 3

Resultado Mecanismo NMAR

%	Condição 1 Carga >= 1755,5			Condição 2 Carga < 1755,5			Condição 3 1518 <= carga <=1969			Condição 4 Carga >= 1969			Condição 5 Carga <= 1518		
	N	**μ**	**δ**	**N**	**μ**	**δ**	**N**	**μ**	**δ**	**N**	**μ**	**δ**	**N**	**μ**	**δ**
10	112070	1742	319,1	111464	1769	321,8	111773	1756	327,8	114701	1744	316,2	114708	1765	318,2
20	106490	1726	316,5	105278	1785	322,2	105897	1758	335,5	11753	1733	310,7	111768	1776	314,8
30	100911	1710	313,3	99092	1802	321,6	100020	1759	343,7	108805	1721	304,5	108827	1787	311,1
40	95331	1691	307,7	92907	1822	320	94144	1760	352,9	105857	1708	297	105886	1799	306,6
50	89752	1671	301,6	86721	1844	317,1	88268	1762	362,8	102909	1694	288,8	102945	1811	301,5
60	84173	1646	290,5	80536	1870	310,5	82392	1764	374	99961	1680	278,6	100005	1824	295,4
70	78593	1619	276,3	74350	1900	301,1	76516	1766	386,4	97013	1664	266,9	97064	1839	287,9
80	73014	1587	253,8	68165	1935	285,4	70639	1768	400,4	94065	1648	252,8	94123	1853	278,9
90	67434	1551	221,6	61979	1978	258,2	64763	1771	416,3	91117	1631	235,3	91183	1869	268,2
100	61855	1508	164,4	55794	2030	210	58887	1774	434,6	88169	1612	213,6	88242	1886	255,2

Total de registros dados completos: 117649, média 1755,5 kgf/mm, N= Número de Registros, μ = Média, δ = Desvio

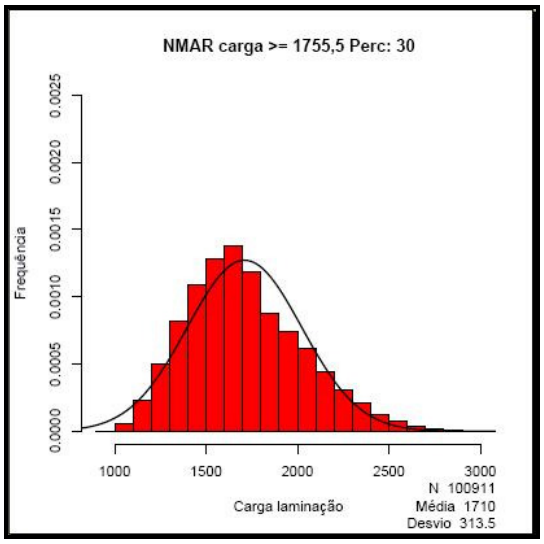


Gráfico 7 – NMAR
condição 1 com 30% ausência

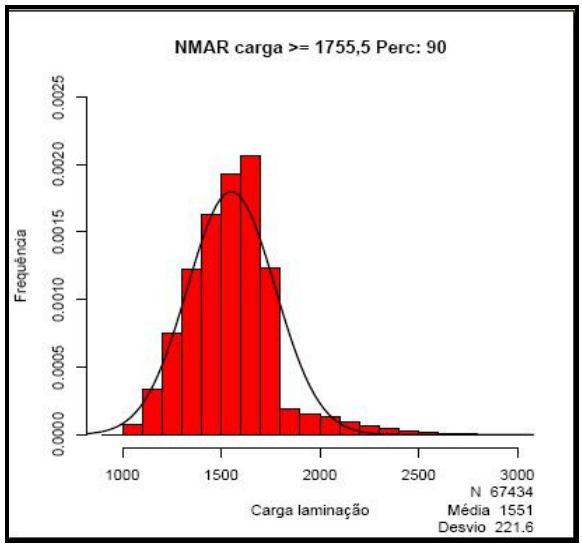


Gráfico 8 – NMAR
condição 1 com 90% ausência

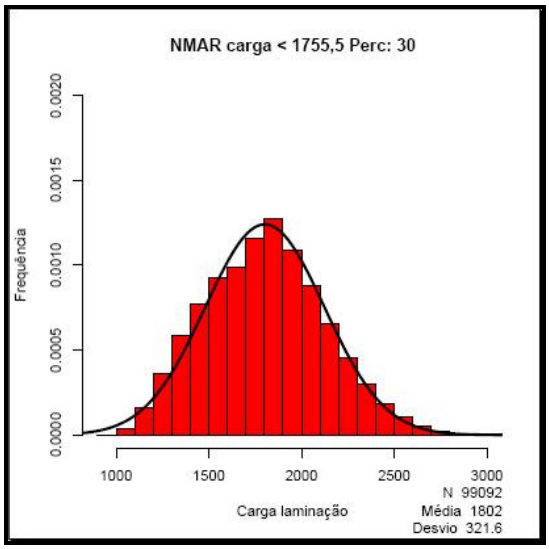


Gráfico 9 – NMAR
condição 2 com 30%

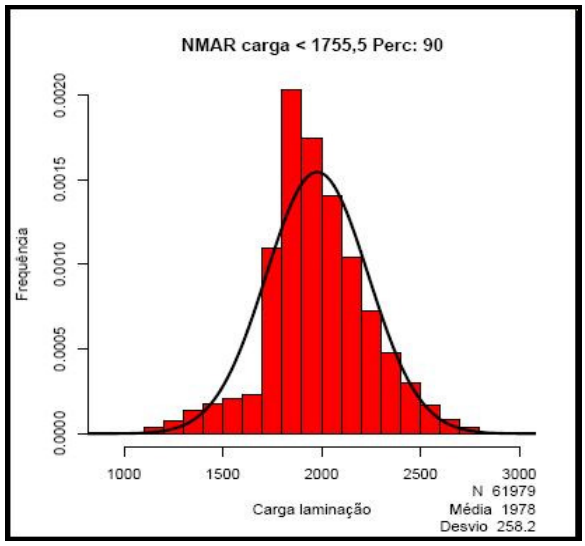


Gráfico 10 – NMAR
condição 2 com 90% ausência

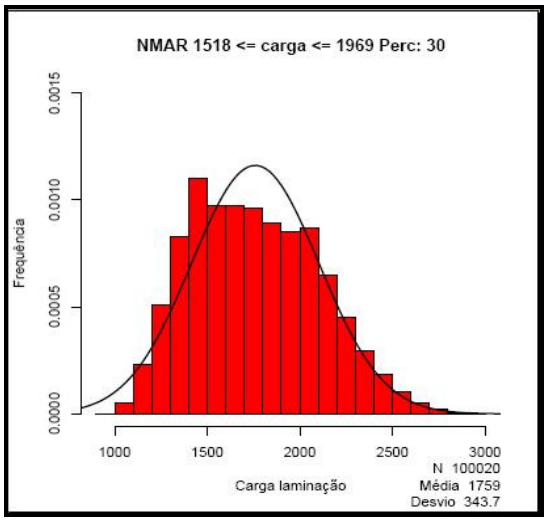


Gráfico 11 – NMAR
condição 3 com 30% ausência

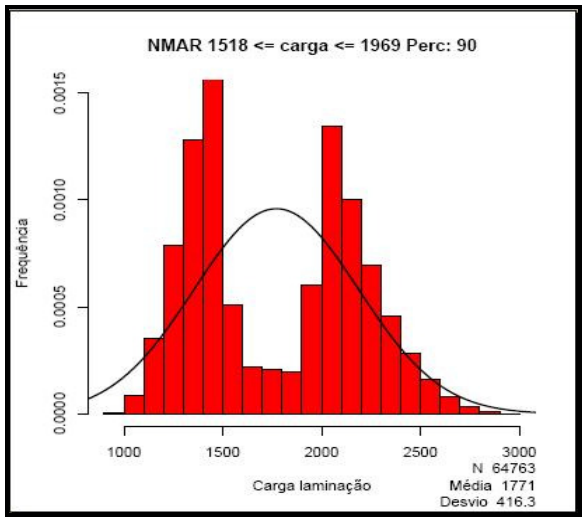


Gráfico 12 – NMAR
condição 3 com 90% ausência

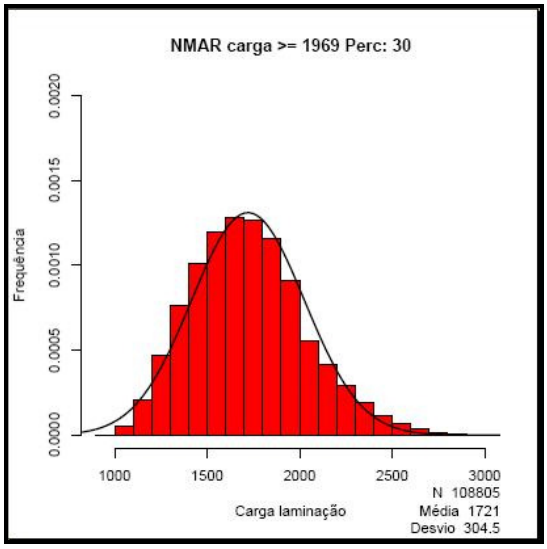


Gráfico 13 – NMAR
condição 4 com 30% ausência

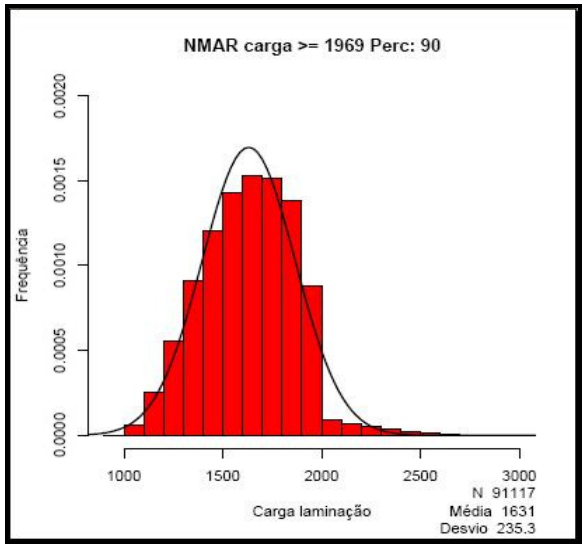


Gráfico 14 – NMAR
condição 4 com 90% ausência

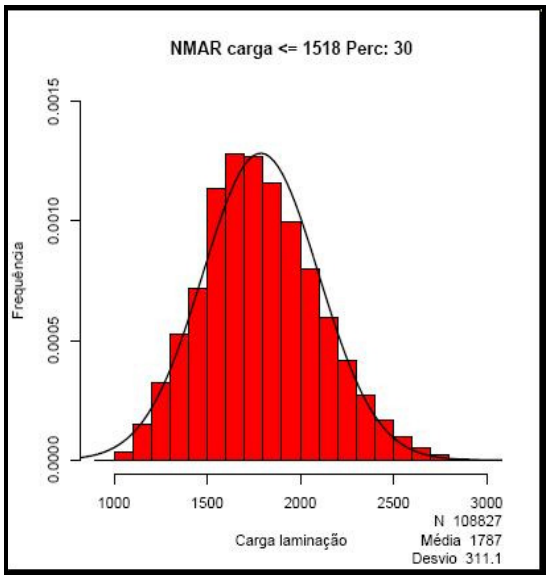


Gráfico 15 – NMAR
condição 5 com 30% ausência

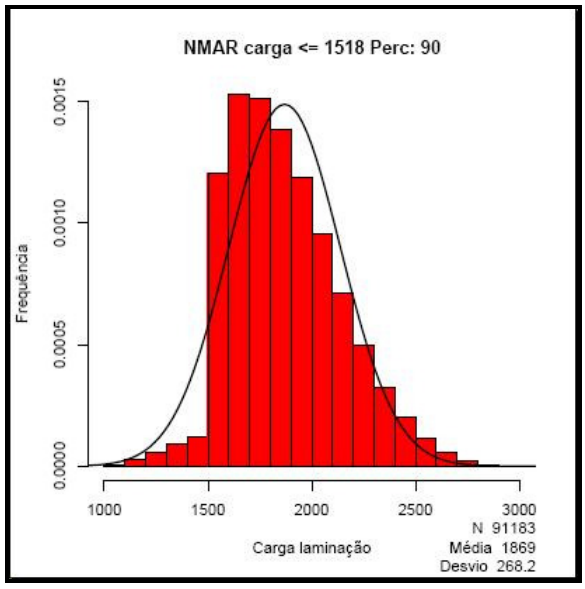


Gráfico 16 – NMAR
Condição 5 com 90% ausência

1.2.4 Conclusão

Assim, através deste experimento, é demonstrado que os mecanismos MCAR, MAR e NMAR (neste trabalho em torno de 30% de ausência) se confundem. Dado uma amostra com dados ausentes não é possível identificar o mecanismo de ausência, considerando apenas a distribuição da frequência, média e desvio padrão, conforme Tabela 4.

O mecanismo NMAR fica claro quando todos os registros de uma condição são eliminados e as características da população são conhecidas de antemão.

Desta forma fica caracterizado que a identificação do mecanismo de ausência não é uma tarefa simples. Geralmente, o dado contém pouca informação que auxilie na identificação do mecanismo de ausência. Pesquisadores têm-se esforçado para determinar porque algumas observações estão ausentes e outras estão presentes (HEITJAN, 1997).

O tratamento de valores ausentes deve ser cuidadosamente analisado, pois caso seja realizado de maneira inadequada, vários problemas poderão ocorrer. Talvez o problema mais sério seja a eliminação de dados ausentes com padrão NMAR. Este padrão não pode ser ignorado, pois causa severas distorções nos resultados. Outro problema surge quando o volume de dados ausentes é massivo, então a eliminação pode ocorrer sobre os registros ou

sobre os atributos que os contêm. De qualquer forma, a eliminação nesta situação pode ser um fator que afeta a qualidade final do conhecimento extraído.

Tabela 4

Comparação dos resultados mecanismos MCAR, MAR, NMAR

Mecanismo	Número do Experimento	Percentual	Número de registros	Média	Desvio
MCAR	6	30%	82354	1756	320,9
	19	95%	5882	1753	319
MAR	1	100%	67228	1847	327,4
	2	100%	50421	1634	266,9
	3	100%	67228	1720	293,2
	4	100%	50421	1802	348,9
NMAR	1	30%	100911	1710	313,5
	2	30%	99092	1802	321,6
	3	30%	100020	1759	343,7
	4	30%	108805	1721	304,5
	5	30%	108827	1787	311,1

Total de registros dados completos: 117649, média 1755,5 kgf/mm

Na literatura muitos métodos de tratamento de dados ausentes têm sido aplicados quando o mecanismo de ausência obedece ao padrão MCAR ou MAR. Porém, ainda estão em aberto métodos apropriados para o padrão NMAR (HEITAN, 1997 e PEARSON, 2005b). PEARSON (2005b) sugere a utilização de técnicas de mineração de dados para explorar as razões de ausência de um mecanismo não ignorável (NMAR), ou seja, procurar por padrões nos dados que indiquem a sistemática da ausência.

1.3 Hipóteses

As regras de associações descrevem padrões de relacionamentos de itens em bancos de dados, fornecendo um conjunto de regras que representam combinações de itens que ocorrem com determinada frequência. A técnica de regras de associação oferece medidas, como suporte e confiança, que permitem a avaliação da qualidade das regras geradas.

Este trabalho utiliza técnicas de regras de associação para procurar por padrões nos dados que indiquem a causa da ausência e consequemente auxiliem na identificação do mecanismo de ausência.

Para identificar os mecanismos de ausência através de regras de associação três hipóteses são propostas:

H1) “Para o caso MCAR, espera-se que todas as regras responsáveis pela ausência apresentem uma medida de confiança uniforme”.

H2) “Para o caso MAR, espera-se que regras responsáveis pela ausência apresentem alto valor na medida de interesse e o atributo causador da ausência seja o antecedente da regra”.

H3) “Para o caso NMAR, espera-se que regras responsáveis pela ausência apresentem alto valor na medida de interesse e o atributo classe seja o antecedente da regra”.

1.4 Justificativa

A ocorrência de dados ausentes em bancos de dados é um fato comum e pode gerar sérios problemas na análise de dados estatísticos, na extração do conhecimento e na aplicação dos algoritmos de mineração de dados. Por outro lado, eliminar instâncias e/ou atributos com dados ausentes pode causar perda de informação e a substituição por valores *default* pode introduzir distorções nos resultados. Dado ausente é um problema real, que necessita de tratamento. O não tratamento dos dados ausentes pode invalidar o resultado de um projeto de KDD. Sendo assim, tratamento de dados ausentes torna-se um ponto extremamente importante no processo de KDD.

1.5 Motivação

A crescente utilização dos computadores nas mais variadas áreas do conhecimento, tem proporcionado um aumento no volume de informações armazenadas em bases de dados nos últimos anos. Porém, um dos grandes problemas encontrados nestas bases de dados é o dado ausente, que impõe uma perda na qualidade dos resultados de análises baseadas nesses repositórios.

Muitos métodos já foram propostos para imputar, estimar ou lidar com os dados ausentes. Entretanto, os métodos propostos são fortemente relacionados ao mecanismo de

ausência e a identificação desse mecanismo não é uma tarefa simples. Além disso, os mecanismos se confundem em determinadas situações. Assim, métodos e técnicas que auxiliem na identificação dos mecanismos de ausência também contribuirão diretamente no processo de recuperação dos dados ausentes.

1.6 Objetivos

Este trabalho tem como objetivo geral propor um método para auxiliar a identificação do mecanismo de ausência. Também é considerado objetivo geral do trabalho:

- Definição de uma taxonomia dos métodos de tratamento de dados ausentes;
- Elaboração de experimentos que comprovem a usabilidade do método. Para isto três tipos de base de dados foram consideradas: base de dados com atributos com alta, média e baixa correlação.

1.7 Principais contribuições

A principal contribuição é apresentar um método de identificação do mecanismo de ausência, baseado em regras de associação que possa ser utilizado em base de dados com valores ausentes no atributo classe. Como trabalho futuro espera-se que esse método seja incorporado ao modelo de estimação e/ou imputação com a intenção de melhorar o desempenho do mesmo. Além disso, espera-se que a caracterização do mecanismo de ausência possa ser útil na proposta de novos modelos que se aplicam ao caso NMAR.

2 REVISÃO BIBLIOGRÁFICA E FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo apresentar uma revisão teórica fundamental acerca dos dados ausentes, uma taxonomia para os métodos, uma descrição dos principais métodos e técnicas disponíveis. Além disso, apresenta algumas considerações sobre o mecanismo de ausência e sobre a abordagem não convencional de tratamento de dados ausentes.

2.1 Revisão Teórica Fundamental

Os primeiros trabalhos relacionados aos dados ausentes surgiram nos anos 30 (ALLAN e WISHART, 1930). A primeira revisão de literatura relativa à análise de dados com valores ausentes foi apresentada em 1966 (AFIFI e ELASHOFF, 1966). Outros trabalhos identificados através de revisão de literatura incluem HARTLEY e HOCKING (1971), ORCHARD e WOODBURY (1972), DEMPSTER, LAIR e RUBIN (1977), LITTLE e RUBIN (1983), LITTLE e SCHENKER (1994) e LITTLE (1997), os quais são contribuições puramente estatísticas.

Um grande marco na literatura de dados ausentes foi a formalização dos mecanismos de ausência (RUBIN, 1976), até então ignorados. RUBIN (1976) conclui que para lidar com dados ausentes, o mecanismo que causa a ausência deve ser explicitamente considerado, e para isso, são necessários modelos que representem esse processo. RUBIN (1976) propôs uma taxonomia para os dados ausentes: MCAR (*missing completely at Random*), MAR (*missing at random*) e NMAR (*not missing at random*). O mecanismo MCAR ocorre quando a probabilidade de ausência para uma variável X não está relacionada ao próprio valor de X e também a nenhuma outra variável da base de dados. O mecanismo MAR ocorre quando a probabilidade de ausência para uma variável X depende de outras variáveis, mas não depende da própria variável X. Finalmente o mecanismo NMAR, também conhecido como não ignorável, ocorre quando a probabilidade de ausência para a variável X é relacionada ao próprio valor de X. Em BUHI, GOODSON e NEILANDS (2008), os autores simplificam a taxonomia RUBIN (1976) salientando as causas do dado ausente. Para o mecanismo MCAR a causa de ausência é completamente aleatória. Já o mecanismo MAR a causa obedece a uma

aleatoriedade condicional, enquanto o mecanismo NMAR apresenta razões sistemáticas para ausência do dado.

Seja Y um conjunto de dados, Y_{obs} um subconjunto de Y que contem os valores observados e Y_{aus} um subconjunto de Y que contem os valores ausentes, ou seja, $Y = Y_{obs} \cup Y_{aus}$, e R uma matriz indicadora de respostas para cada item de Y, indicando que $y_{i,j}$ está presente ($r_{i,j} = 0$) ou ausente ($r_{i,j} = 1$). O mecanismo de ausência é caracterizado pela distribuição condicional de R dado Y, seja $f(R|Y, \phi)$, onde ϕ representa parâmetros desconhecidos da função de probabilidade.

No mecanismo MCAR, a ausência não depende dos valores ausentes e nem dos valores completos:

$$f(R \mid Y, \phi) = f(R \mid \phi) \text{ para todo } Y, \phi \tag{1}$$

Para o mecanismo MAR, a ausência depende somente dos valores observados Y_{obs} e não depende dos valores ausentes Y_{aus}:

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, \phi) \text{ para todo } Y_{aus}, \phi \tag{2}$$

O mecanismo é chamado NMAR se a ausência de R depende dos valores ausentes de Y. Considere $Y = (y_1,...,y_n)^T$ onde y_i significa o valor de uma variável aleatória para o item i, e seja $R = (R_1,...R_n)$ onde $R_i = 0$ para itens completos e $R_i = 1$ para itens ausentes. Suponha que a distribuição conjunta de (y_i, R_i) é independente para os itens, então

$$f(Y, R \mid \theta, \phi) = f(Y \mid \theta) f(R \mid Y, \phi) = \prod_{i=1}^{n} f(y_i \mid \theta) \prod_{i=1}^{n} f(R_i \mid y_i, \phi) \tag{3}$$

onde $f(y_i|\theta)$ representa a densidade de y_i indexada por parâmetros desconhecidos θ, e $f(R_i|y_i, \phi)$ é a densidade de uma distribuição de Bernoulli para o indicador binário R_i com probabilidade $Pr(R_i=1| y_i, \phi)$, na qual y_i é ausente.

LITTLE e RUBIN (2002) definem o padrão de ausência que descreve que valores são observados na matriz R e quais valores estão ausentes. Basicamente os padrões resumem-se a geral (aleatório) e específico. No padrão geral ou aleatório os dados ausentes são encontrados em quaisquer registros do conjunto de dados. Já o padrão específico é univariado, ou seja, está

restrito a uma única variável da base de dados. O padrão específico pode ser também monotônico, quando a quantidade de valores ausentes é sempre crescente de uma amostra para outra. Como exemplo, HRUSCHKA JR. (2003) apresenta uma pesquisa com coleta de dados efetuada em três etapas. Na primeira etapa, não se tem dados ausentes; na segunda etapa, realizada alguns anos após a primeira com as mesmas pessoas, nem todos os participantes foram encontrados, tem-se assim alguns dados ausentes. Por fim, na última etapa, somente os participantes da segunda etapa foram procurados, da mesma forma nem todos os participantes foram encontrados, elevando o número de valores ausentes.

Outra contribuição importante foi a formalização do algoritmo EM-*Expectation-Maximization* (DEMPSTER, LAIRD e RUBIN, 1977), um método computacional para estimação eficiente de dados incompletos. O algoritmo EM representou um marco na maneira como os estatísticos tratavam o dado ausente. A partir desse trabalho os estatísticos passaram a ver o dado ausente como fonte de variabilidade a ser considerada nas análises estatísticas.

Os primeiros autores a afirmar que a melhor maneira de lidar com dados ausentes é não tê-los foram ORCHARD e WOODBURY (1972). Os autores fazem uma analogia dos dados ausentes com os acidentes, que não são planejados, porém medidas de prevenção são tomadas para evitá-los. Assim, precauções devem ser tomadas na coleta de dados para evitar a ocorrência de dados ausentes. Posteriormente, ALLISON (2000) e BUHI, GOODSON e NEILANDS (2008) reiteram a afirmação de ORCHARD e WOODBURY (1972). Trata-se de uma recomendação utópica, pois dado ausente é um problema real em base de dados de várias áreas do conhecimento e tem sido foco de muitas pesquisas nos últimos anos. LYNCH (2003) chega citar a dificuldade de publicação de trabalhos empíricos em Sociologia sem a discussão de como os dados ausentes foram tratados. Por sua vez, SHADISH (2002), na área de Psicologia, admite a rejeição de artigos que não explicitam o tratamento de dados ausentes. Em diversas áreas do conhecimento, artigos de revisão e tutoriais são publicados periodicamente, com objetivo de apoiar os pesquisadores no tratamento de dados ausentes. SCHAFER e GRAHAM (2002), apresentam uma ampla revisão dos métodos existentes para lidar com dados ausentes. Outros trabalhos de revisão são encontrados em MYERS (2000) e TSIKRIKTSIS (2005). Por sua vez, SCHEFFER (2002), ACOCK (2005) e BUHI, GOODSON e NEILANDS (2008), comparam as técnicas disponíveis nos principais pacotes estatísticos.

2.2 Padronização de termos

No contexto estatístico os procedimentos são aplicados a bases de dados retangulares (tabelas). As linhas representam as unidades do experimento (algumas vezes chamada de sujeitos, casos ou observações ou indivíduos diferentes de alguma população de interesse) enquanto as colunas contém as variáveis que representam as características coletadas ou observadas.

No contexto de mineração de dados comumente os procedimentos também são aplicados à base de dados retangulares. Uma base de dados consiste de uma matriz de valores de dados. As linhas representam os registros (algumas vezes chamados de itens, *itemsets*, tuplas, objetos ou instâncias) enquanto as colunas representam os atributos.

De forma a padronizar a análise e as contribuições em relação aos dados ausentes alguns termos são padronizados neste trabalho.

Dada uma base de dados com "*n*" registros e $X_1..X_k$ atributos, os atributos são representados por letras maiúsculas e os seus valores por letras minúsculas. Um vetor X_1 é aquele que contem $\{x_{11}, ..., x_{1n}\}$ ocorrências do atributo X_1. A matriz X representa as ocorrências dos atributos $X_1...X_k$.

2.3 Taxonomia dos métodos

Do ponto de vista prático, a literatura classifica os métodos para lidar com dados ausentes de diversas formas. Definitivamente não existe consenso entre os autores sobre a classificação das técnicas de dados ausentes existentes. Os trabalhos publicados propõem uma classificação própria com um grau de generalização maior ou menor (SOARES, 2007).

BATISTA e MONARD (2003) dividem os métodos em: ignorar e descartar dados ausentes, descartar registros com atributos que têm muitos dados ausentes, estimar parâmetros e imputar. Segundo SOARES (2004) os métodos estatísticos para analisar dados incompletos são agrupados em procedimentos de imputação e procedimentos de modelos. No primeiro grupo os procedimentos visam completar os dados ausentes, enquanto no segundo grupo utilizam-se modelos probabilísticos.

Já HRUSCHKA, HRUSCHKA JR, EBECKEN (2003a, 2003b) resumem o tratamento de valores ausentes em: ignorar os registros com valores ausentes, preenchê-los manualmente, substituir o valor ausente por uma constante, usar a média ou moda, ou ainda atribuir o valor mais provável. Esses métodos podem distorcer as características dos registros, pois alteram o cenário no qual os dados foram gerados.

MYRTVEIT, STENSRUD e OLSSON (2001) descrevem os seguintes métodos para lidar com dados ausentes: omitir todos os registros incompletos da análise de dados, técnicas baseadas em imputação, técnicas baseadas na designação de pesos e técnicas baseadas em modelos.

Por sua vez a classificação proposta por LITTLE e RUBIN (2002) restringe-se aos métodos estatísticos. Segundo eles, os procedimentos são agrupados em quatro categorias: procedimentos baseados nos casos completos, procedimentos baseados nos casos disponíveis, procedimentos de imputação e procedimentos de ponderação.

MAGNANI (2004) propõe uma classificação mais didática e refinada, buscando contemplar todas as possibilidades de técnicas para tratar o dado ausente. São elas: métodos convencionais, imputação, estimativa de parâmetros e gerenciamento direto dos dados ausentes. Em SOARES (2007) uma modificação na classificação de MAGNANI (2004) é proposta.

No presente trabalho a taxonomia proposta por SOARES (2007) é também alterada. Assim, a taxonomia adotada neste trabalho obedece a seguinte categorização:

1) Procedimentos baseados nos casos completos
    a. Remoção de registros incompletos
    b. Remoção em pares
    c. Remoção de colunas com valores ausentes
    d. Remoção com ponderação
2) Procedimentos baseados em imputação
    a. Global baseada nos atributos ausentes
    b. Global baseada nos atributos não ausentes
    c. Local
    d. Múltipla
    e. Composta
3) Procedimentos baseados em modelos
    a. Métodos de verossimilhança
    b. Modelos bayesianos

c. Modelos de seleção e modelos de mistura
d. Modelos neurais

4) Procedimentos baseados na manipulação direta dos dados ausentes

Cada um desses grupos de categorias será apresentado a seguir.

2.3.1 Procedimentos baseados nos casos completos

Os procedimentos baseados nos casos completos são simples e usualmente são opções *default* dos pacotes estatísticos (LITTLE e RUBIN, 2002).

A remoção de registros incompletos é chamada de *Listwise deletion*, que é o método mais simples de obter uma base de dados completa. Somente registros completos são mantidos (NIE et al., 1975). Este método mostra-se satisfatório quando a quantidade de dados ausentes é pequena. Por outro lado, quando aplicado em situações com muitos dados ausentes implica em grande perda de dado e informação. Além disso, este método deve ser aplicado somente se o mecanismo de ausência obedece ao padrão MCAR. No contexto de mineração de dados, a remoção de registros diminui a habilidade de encontrar padrões consistentes, implicando em piores resultados, uma vez que poucos registros são utilizados no processo de treinamento ou no ajuste de modelos de predição (FORTES, MORA-LOPEZ e TRIGUERO, 2006).

O método de remoção em pares ou casos disponíveis (*pairwise deletion*) é uma variação do método *listwise deletion*, no qual registros incompletos são mantidos somente quando o atributo analisado não possui valor ausente. Assim como o método *listwise deletion*, o método *pairwise deletion* pode causar resultados tendenciosos se o dado ausente não for MCAR (LITTLE e RUBIN, 2002). O método *pairwise deletion* tem a vantagem de usar todas as variáveis disponíveis e consequentemente aumentar o poder estatístico, porém exige procedimentos elaborados para lidar com as ausências das bases de dados.

Outro método convencional é descartar todas as variáveis (colunas) com valores ausentes, que por sua vez não é uma opção interessante, devido a razões evidentes.

Por fim, a ponderação é uma versão mais refinada do método de exclusão de registros. Após a exclusão dos registros incompletos os dados remanescentes são reponderados de modo que sua distribuição assemelha-se a amostra completa. Neste método, inferências aleatórias do dado ausente são baseadas em pesos projetados, que são inversamente proporcionais a

probabilidade de seleção. Os procedimentos de ponderação são muito utilizados nas aplicações de questionários de pesquisas no ajuste das unidades não respondidas (LITTLE e RUBIN, 2002).

2.3.2 Procedimentos baseados em Imputação

A imputação é o procedimento de substituição de valores ausentes. CARTWRIGHT, SHEPPERD e SONG (2003) recomendam a imputação para grandes bases de dados. Por outro lado, JUNNINEN et al. (2004) alerta que os métodos de imputação não podem ser usados como um tipo de "alquimia estatística" e sim para ajudar a remediar as situações com dados ausentes. Em situações específicas, a imputação não é aplicável, como por exemplo, coleções de dados assíncronos e distribuídos no qual a natureza do dado é temporal e/ou espacial (GORODETSKY, KARSAEV e SAMOILOV, 2005). Por exemplo, dados oriundos de um sistema de detecção de intrusos em uma rede de computadores possuem essas características. Nesse caso, a ausência de dados pode ser inclusive aparente, pois o dado em falta pode ser um dado vazio, caso sejam consideradas as diversas fontes de geração de intrusão. PYLE (1999) define dado vazio como o dado que não existe no mundo real, e para esta situação recomenda o uso de técnicas que possam lidar com os dados faltantes.

Já TWALA, CARTWRIGHT e SHEPPERD (2005) destacam que a imputação esconde a incerteza do dado, levando a intervalos de testes e confiança inválidos, uma vez que os valores são derivados dos dados existentes. Os métodos de imputação normalmente predizem valores mais comportados que os valores reais. Esse fato pode levar o risco de simplificar excessivamente o problema estudado. Uma vez que, o mais importante aspecto da substituição de valores ausentes é a não distorção das características originais da amostra (HRUSCHKA JR., 2003).

Segundo RUBIN (1996) a imputação tem dois objetivos. O objetivo básico é permitir que os usuários finais dos dados apliquem suas ferramentas de análise em qualquer base de dados como se não existisse nenhum dado ausente. O segundo objetivo é obter estatisticamente inferências válidas. Muitas técnicas de imputação têm sido desenvolvidas ao longo dos anos. DONDERS, HEIJDEN e STIGEN (2006) apresentam uma introdução aos métodos de imputação, enquanto HU, SALVUCCI e COHEN (1998) descrevem brevemente aproximadamente trinta métodos de imputação distintos.

O método de **imputação global baseada no atributo ausente** utiliza os valores existentes para preencher os valores ausentes. Esses métodos utilizam uma medida de tendência central, como a média, mediana ou moda. Neste método, o desvio padrão é distorcido, pois, os valores extremos são trazidos para o meio da distribuição, reduzindo a variância, mesmo sobre o padrão MCAR. É aconselhável que a imputação não altere a variância da amostra. Uma alternativa é a introdução de uma perturbação aleatória no valor da média. Esta perturbação tenta evitar a distorção dos resultados, mesmo assim não é um método satisfatório (MAGNANI, 2004).

A **imputação global baseada nos atributos não ausentes**, utiliza as correlações existentes entre variáveis com valores ausentes e variáveis disponíveis. Uma estratégia é a imputação por regressão. Os atributos ausentes são tratados como variáveis dependentes, e a regressão é executada para imputar os valores ausentes. Uma desvantagem é que a regressão é baseada em um modelo escolhido, que nem sempre representa satisfatoriamente os dados. Outra desvantagem é a distorção do erro padrão (LITTLE, 1992).

Outra técnica de imputação global é a regra de associação. As regras de associação buscam interconexões de registros na tentativa de expor características e tendências. O objetivo é encontrar regras na forma de antecedente e consequente, que sejam frequentes e válidas e que atendam a critérios mínimos de suporte e confiança. O suporte está relacionado à frequência que uma regra ocorre em uma tabela. A confiança reflete a validade de uma regra e expressa a qualidade de uma regra, indicando o quanto a ocorrência do seu antecedente pode assegurar a ocorrência de seu consequente.

O método MVC (*Missing Value Completion)* proposto por RAGEL e CRÉMILLEUX (1998) utiliza regras de associação robustas para lidar com dados ausentes através do algoritmo RAR-*Robust Association Rules*. Em RAGEL e CRÉMILLEUX (1999) demonstra-se a aplicação do método MVC em uma base de dados real. O algoritmo RAR descobre regras de associação em conjuntos de dados que não contêm dados ausentes em seus registros. Estas regras são obtidas a partir de um subconjunto dos dados que não possui nenhum valor ausente, chamado banco de dados válido. O algoritmo MVC decide qual dado imputar baseando-se nos consequentes das regras. Desse modo, duas situações são geradas:

1. Todas as regras chegam a mesma conclusão, assim assume-se o valor que aparece no consequente da regra como a melhor imputação;
2. Ocorre a intervenção do usuário no processo de complementação de dados, como critério heurístico. O usuário decide qual regra ou quais regras utilizar baseado em

parâmetros, como suporte e confiança de cada regra, e assume-se o consequente da regra como o valor a ser imputado.

A utilização de regras de associação para completar dados ausentes é encontrada em WU, WUN e CHOU (2004), a partir da idéia que regras de associação descrevem as dependências de relacionamentos entre atributos em uma base de dados, onde todo o *itemset* que possui atributo com valor ausente, também deve possuir relacionamentos similares.

A **imputação local** apresenta três métodos: *hot-deck, cold deck* e KNN. O método *hot-deck* é dividido em duas etapas: primeiro os registros são subdivididos em classes e depois, para cada registro incompleto, os valores imputados são escolhidos baseados em registros da mesma classe. MAGNANI (2004) destaca algumas vantagens do método *hot-deck*: redução do erro padrão sem a imposição de um modelo rígido, produção de um conjunto de dados sem dados ausentes e preservação da distribuição da população. Além disso, o método permite que técnicas distintas de imputação possam ser usadas para cada grupo gerado. SOARES (2007) salienta que a divisão dos dados em grupos pode trazer grandes benefícios ao processo de imputação.

Por sua vez, o método *cold-deck* substitui os valores ausentes por uma constante externa. Essa constante externa pode ser obtida a partir de dados históricos ou conhecimento do domínio do problema. LITTLE e RUBIN (2002) afirmam que o embasamento teórico desse método é insatisfatório.

O método KNN é uma variação do método *NN-Nearest Neighbor*. O KNN é muito utilizado na tarefa de classificação de registros. No algoritmo original a classificação é baseada apenas no vizinho mais próximo (COVER e HART, 1967). Já o KNN, para estimar a classe de um novo padrão X, calcula os k-vizinhos mais próximos a X e classifica-o como sendo da classe que aparece com maior frequência dentre os seus k-vizinhos. O parâmetro "k", indica o número de vizinhos utilizados pelo algoritmo durante a fase de teste e permite uma classificação mais refinada, porém o valor ótimo de k varia de um problema para outro. Segundo JONSSON e WOHLIN (2004) um valor apropriado para k é a raiz quadrada do número total de registros. O KNN também é utilizado como método de imputação, e neste caso os k registros que são próximos ao registro com valores ausentes são combinados na estimação. BATISTA e MONARD (2001) analisam o KNN como método de imputação e concluem que os resultados obtidos com o KNN são superiores aos obtidos pelas abordagens internas utilizadas pelos algoritmos de aprendizado CN2, C4.5 e pela imputação pela média ou moda. Além disso, apontam que o KNN, mesmo diante de uma grande quantidade de dados ausentes mantém a qualidade da imputação.

A **imputação múltipla** (*MI-Multiple Imputation*) proposta por RUBIN (1976) surgiu como uma alternativa flexível aos métodos de verossimilhança. A imputação múltipla foi utilizada pela primeira vez em arquivos de dados públicos oriundos de censo e pesquisas, devido à necessidade de disponibilizar dados completos para os usuários finais. A imputação múltipla objetiva reduzir a incerteza inerente ao dado imputado e é composta de três etapas: imputação, análise e combinação dos resultados. Na etapa de imputação, os conjuntos de valores plausíveis para as observações ausentes são criados utilizando algum método de imputação próprio. Um método próprio de imputação pode ser baseado em um modelo explícito (paramétrico, como regressão linear) ou implícito (como por exemplo, *hot-deck*) que incorporam a variabilidade apropriada entre as repetições (RUBIN, 1976). Cada conjunto de valores plausíveis é usado para preencher o dado ausente e criar uma base de dados completa. Na etapa de análise, cada base de dados é analisada utilizando métodos de dados completos. Finalmente, na etapa de combinação de resultado, os resultados são combinados considerando a incerteza da imputação. Formalmente, o processo de imputação consiste em:

Imputação: geração de um conjunto de $m > 1$ valores plausíveis para cada elemento ausente de $Z_{aus} = (Y_{aus}, X_{aus})$

Análise: análise das m bases de dados utilizando métodos de casos completos

Combinação: combina os resultados das m análises

A imputação múltipla a partir da perspectiva bayesiana é uma motivação natural devido ao tratamento dos parâmetros (SHAFER e GRAHAM, 2002). RUBIN (1988), recomenda que as imputações sejam criadas através de um processo bayesiano: especificar um modelo paramétrico para o dado completo, aplicar uma distribuição prévia para os parâmetros desconhecidos do modelo e simular m execuções independentes a partir da distribuição condicional do dado ausente, segundo o dado observado utilizando o Teorema de Bayes. Além da perspectiva bayesiana é possível utilizar outros métodos de imputação entre eles os métodos de propensão, análise de discriminante e regressão logística. Para aplicação do método de imputação múltipla é necessário que o mecanismo obedeça ao padrão MAR.

A imputação múltipla é uma aproximação do método máxima verossimilhança (*ML-Maximum Likelihood*). A imputação múltipla tenta alguns poucos valores plausíveis para o dado ausente, enquanto o método ML integra todos os valores possíveis, dando mais peso para os valores mais plausíveis. O resultado da estimação ML é o mesmo se infinitas imputações fossem executadas na imputação múltipla. A desvantagem de ML são as suposições sobre a distribuição do dado ausente. Sendo assim, quando é possível conviver

com essas suposições, ML torna-se a opção ideal, caso contrário MI (*MI-Multiple Imputation*) apresenta-se mais flexível.

A **imputação composta** foi proposta por SOARES (2007). Neste método, o processo de imputação de um atributo ausente é precedido da aplicação de outras tarefas, como por exemplo o agrupamento de dados e seleção de atributos, com o objetivo de melhorar a qualidade do dado imputado. O método propõe a aplicação de comitês de complementação de dados para o processo de imputação. Esta abordagem incorpora o conceito de meta-aprendizado encontrado em comitês de classificação. Os comitês de classificação utilizam várias sugestões para um processo decisório ou para uma combinação de resultados que podem produzir uma nova classificação para um registro de uma base de dados.

2.3.3 Procedimentos baseados em Modelos

A modelagem de dados engloba técnicas estatísticas, probabilísticas e de aprendizado para obtenção de um modelo que consiga representar de uma forma genérica as características dos dados. Nesta categoria destacam-se os algoritmos de verossimilhança, os métodos bayesianos, a seleção de modelos, o método de mistura de padrões e as redes neurais.

Os **métodos de verossimilhança** (*ML-Maximum Likelihood*) estimam os parâmetros da função de distribuição estatística para uma amostra incompleta. O objetivo é encontrar um modelo que represente o conjunto de dados e, com isso ter condições de regredir qualquer valor ausente existente. O algoritmo EM-*Expectation-Maximization*, proposto por DEMPSTER, LAIRD e RUBIN (1977), consiste em um procedimento iterativo de verossimilhança, que estima os parâmetros de uma função de densidade probabilística de uma amostra. Este método é baseado nos relacionamentos observados entre todas as variáveis e injeta um grau de erro aleatório para refletir a incerteza da imputação. Caso exista uma idéia dos valores ausentes, então a estimação dos parâmetros do modelo torna-se simples. Similarmente, conhecendo os parâmetros do modelo, então torna-se possível obter predições não distorcidas para os valores ausentes. Essa interdependência entre parâmetros do modelo e valores ausentes sugerem um método iterativo, onde primeiramente os valores ausentes são preditos com base nos valores assumidos para os parâmetros, e essas predições são utilizadas para atualizar os parâmetros estimados, e assim sucessivamente até a convergência do algoritmo. Uma maneira de monitorar a convergência do EM é a verificação da função

loglikelihood que deve aumentar a cada iteração até tornar-se estável (SCHAFER e OSLEN, 1998).

O algoritmo iterativo pode consumir muito tempo de processamento antes de convergir, principalmente diante de muitos valores ausentes e muitas variáveis. A taxa de convergência do EM é proporcional à quantidade de informação ausente da base de dados (LITTLE e RUBIN, 2002). Entretanto, como qualquer outro procedimento de otimização não linear, sofre do problema do mínimo local, sendo sensível aos valores iniciais dos parâmetros. Para usá-lo é necessário especificar a distribuição da amostra, o que nem sempre é possível no processo de KDD (MAGNANI 2004). Variações deste método visam melhorar a taxa de convergência, como por exemplo, os algoritmos ECM, ECME, AECM e PX-EM (RUBIN, 2006).

WILLIAMS et al. (2007) apresentam um algoritmo de regressão logística para classificação de dados incompletos. A função de densidade condicional é estimada utilizando um modelo de mistura gaussiano (GMM) e a estimação de parâmetro utiliza os algoritmos EM e VB-EM. Neste trabalho o problema do dado ausente é resolvido evitando-se as heurísticas de imputação.

Os **métodos bayesianos** são amplamente utilizados para o tratamento de dados ausentes, tanto na área estatística como na área de mineração de dados.

Segundo MAGALHÃES (2007), as redes bayesianas são estruturas que combinam a distribuição de probabilidades e grafos associados a um conjunto de variáveis de interesse. Um grafo G é composto por um conjunto de vértices V conectados por um conjunto de arcos A que representam as ligações entre os vértices. Uma rede bayesiana necessita de um grafo orientado (ou dígrafo) que estabelece uma relação de dependência direta (causa/efeito) entre os vértices. No grafo orientado as arestas são chamadas de arcos, e a relação definida pelo conjunto de arcos A não é simétrica, existindo uma orientação na relação entre os vértices. Um grafo direcionado acíclico (GDA) é um grafo composto por vértices e arcos no qual não existe ocorrência de ciclos. Uma rede bayesiana é um par $B = (G, \theta)$ definido sobre um conjunto de variáveis aleatórias $X = \{ X_1, X_2,...,X_n \}$ onde cada X_i corresponde a um vértice, G é um grafo direcionado acíclico, chamado de estrutura, e θ é um conjunto de parâmetros que especificam as distribuições de probabilidades condicionais que satisfaçam a condição de Markov:

$$P_\theta[X_i \mid X_j, pa(X_i)] = P_\theta[X_i \mid pa(X_i)] \quad (4)$$

onde $pa(X_i)$ é o conjunto de vértices que são pais de X_j. Ou seja, a condição de Markov para uma rede bayesiana diz que qualquer vértice da rede é condicionalmente independente de seus não descendentes condicionados a seus pais. Uma vez definida a estrutura da rede e seu conjunto de parâmetros θ, pode-se calcular uma função que atribui um valor para cada GDA baseada nos dados, chamada de função *score,* cujo cálculo depende da distribuição de probabilidades associadas às variáveis aleatórias do problema em questão. Um modelo bayesiano pode ser utilizado na predição de um parâmetro θ ou qualquer outro valor não observado em qualquer uma das outras variáveis do modelo. Esta característica de trabalhar com valores desconhecidos é uma motivação para utilizar esta teoria em tratamento de valores ausentes.

GARCÍA e HRUSCHKA (2005) utilizam o classificador bayesiano como ferramenta de imputação para problemas de classificação considerando a taxonomia RUBIN (1976) do mecanismo de ausência. O trabalho ilustra como o processo de imputação influencia a tarefa de classificação. Um trabalho similar é proposto por LIU e LEI (2006), que propõe um método de imputação bayesiano, utilizando o atributo de imputação como atributo classe para construir um classificador bayesiano.

FRIEDMAN (1997) apresenta um método para aprender a estrutura da rede bayesiana e a estimação dos parâmetros a partir de dados incompletos, baseado em uma extensão do algoritmo EM.

Um novo método de imputação baseado na representação do conhecimento através de redes bayesianas é apresentado por HRUSCHKA JR (2003). Nesse trabalho as redes bayesianas são utilizadas como mecanismo de inferência na predição de valores adequados para a substituição de valores ausentes.

RIGGELSEN e FEELDERS (2005), propõem um método bayesiano para aprendizado de redes bayesianas com dados incompletos. O objetivo é obter a distribuição posterior dos modelos a partir dos dados observados.

Cabe ressaltar que o processo de inferência bayesiana envolve o cálculo das distribuições posteriores a partir de uma distribuição prévia. Este cálculo é condicionado aos dados observados, portanto quanto maior a amostra de dados utilizada na obtenção da distribuição posterior, menor é a influência da distribuição prévia na predição (HRUSCHKA JR, 2003). Entretanto, frente a muitos dados ausentes, os cálculos das distribuições posteriores tornam-se comprometidos.

Para o mecanismo NMAR deve-se explicitamente especificar uma distribuição para a ausência a ser adicionada ao modelo do dado completo. Há duas formas de implementação: seleção de modelos e modelo de mistura de padrões.

O método **seleção de modelos** foi usado primeiramente na área de Econometria para descrever como o problema de resposta para um item de questionário depende do próprio item (AMEMIYA, 1984 e HECKMAN, 1976). Na seleção de modelos, primeiro especifica-se uma distribuição para o dado completo, depois é requerido que assuma-se uma distribuição para o dado ausente. Ou seja, é necessário que suposições da distribuição do dado ausente sejam feitas para a identificação do modelo. Entretanto, essas suposições nem sempre são suficientes para a identificação do modelo. Assim, seleção de modelos apresenta problemas de convergência devido à estimação de muitos parâmetros. Estes modelos são considerados instáveis para aplicações científicas e geralmente geram mais questões que respostas (LAIRD, 1994). Matematicamente, a seleção de modelos constrói uma junção de distribuição para o dado completo Y e a ausência R pela especificação de uma distribuição marginal para $Y = (Y_{obs}, Y_{aus})$ e uma distribuição condicional para R dado Y:

$$P(Y, R; \theta; \varepsilon) = P(Y; \theta) P(R \mid Y; \varepsilon) \tag{5}$$

Onde θ representa os parâmetros desconhecidos da população do dado completo e ε representa os parâmetros desconhecidos da distribuição condicional da ausência.

Como uma alternativa para a seleção de modelos, LITTLE (1993) descreve uma nova classe de métodos baseada na formulação de **mistura de padrões.** Essa classe de modelos não requer especificação precisa do mecanismo de ausência. O modelo de mistura de padrões categoriza os diferentes padrões de ausência em uma variável preditora, incorporada ao modelo estatístico. Assim, é possível determinar se o padrão de ausência tem um poder preditivo no modelo, por ele mesmo (efeito principal) ou em conjunto com outros preditores (efeito interação). A desvantagem primária do método refere-se à convergência. Caso o número de padrões de ausência e o número de variáveis com ausência sejam altos em relação ao número de casos ou registros da amostra, o modelo pode não convergir devido à insuficiência de dados para suportar o uso de muitos efeitos principais e efeitos de interação.

Um modelo genérico de mistura de padrões pode ser escrita como:

$$P(Y, R; \theta; \varepsilon) = P(R; \eta) P(Y \mid R; \upsilon) \tag{6}$$

Onde η representa a proporção do dado completo em cada grupo de ausência e υ representa os parâmetros das distribuições condicionais do dado completo dentro dos grupos de ausência. A estimação de υ requer algumas suposições inverificáveis, pois uma parte do dado completo Y está escondida em cada grupo de ausência. LITTLE (1993) chama essas suposições de identificação de restrições.

A mistura de padrões é uma imputação múltipla que é executada sobre uma variedade de suposições acerca do mecanismo de ausência. Como por exemplo, considere uma base de dados de pessoas com o atributo peso. Na imputação múltipla normalmente assume-se que os indivíduos que informaram seus pesos são similares aos indivíduos que não informaram. O modelo de mistura de padrões, por exemplo, assume que os indivíduos que não informaram seus pesos são em média 20% mais pesados do que os indivíduos que informaram o peso. Claramente é uma suposição arbitrária: a idéia da mistura de padrões é tentar uma variedade de suposições plausíveis e verificá-las em relação ao resultado.

THIJS et al (2002) descrevem estratégias de ajuste para o modelo de mistura de padrões, como identificação de restrições e ajuste por padrão.

As **redes neurais** artificiais são citadas em vários trabalhos como uma estratégia para lidar, estimar ou imputar dados ausentes. As redes neurais artificiais são funções matemáticas não-lineares que detectam e representam padrões sofisticados e são utilizadas para tarefas preditivas (classificação) e tarefas descritivas (agrupamento). Há também alguns trabalhos que apostam na capacidade das redes de lidar com dados ausentes devido à robustez do algoritmo. Dentre os trabalhos que utilizam as redes neurais como estimadores, destacam-se: WEI e TANG (2003), ABDELLA e MARWALA (2005), ZÁRATE, NOGUEIRA e SANTOS (2005, 2007), GARCIA-LAENCINA, SANCHO-GÓMEZ e FIGUEIRA-VIDAL (2006) e PENG e ZHU (2007). ZÁRATE, NOGUEIRA e SANTOS (2005, 2007) demonstram a utilização de uma RNA que estima atributos ausentes em bancos de dados com massivos dados ausentes a partir de um conjunto mínimo de registros consistentes. Além disso, apontam que a eficiência da representação neural requer uma base de dados consistente que represente o problema. Por outro lado, os trabalhos VIHAROS, MONOSTORI e VINCZE (2002), ORRE et al (2003), JIANG, CHEN e YUAN (2005) e NELWAMONDO e MARWALA (2007) evidenciam a capacidade dessas redes em lidarem com situações de dados incompletos.

WEI e TANG (2003) apostam na etapa de preparação de dados e propõem um *framework* para imputação de dados ausentes utilizando a rede neural *Self Organization Map*

(SOM), através das quais os dados são agrupados em subconjuntos. Dessa forma, pode aumentar a acurácia da imputação. YU, WANG e LAI (2006) propõem um esquema integrado de preparação de dados para análises de dados que utilizam redes neurais. O esquema proposto auxilia o processo de aprendizado, reduz a complexidade do modelo neural e aumenta o desempenho das tarefas de análise de dados.

Em ABDELLA e MARWALA (2005), o método apresentado combina algoritmos genéticos e redes neurais para ajudar na estimativa dos dados ausentes. O algoritmo genético é utilizado para minimizar a função erro derivada de uma rede neural auto-associativa.

Já PENG e ZHU (2007) apresentam o modelo ISOM-DH baseado na análise de componentes principais e redes *Self Organization Map* (SOM) para lidar com dados ausentes. Ainda utilizando RNA, GARCIA-LAENCINA, SANCHO-GÓMEZ e FIGUEIRA-VIDAL (2006) propõem um método de aprendizado multitarefa que paralelamente realiza a tarefa de imputação e classificação para estimar dados ausentes. A imputação é guiada pela tarefa de classificação e os valores imputados são aqueles que contribuem para melhorar o aprendizado.

Na outra linha de investigação, VIHAROS, MONOSTORI e VINCZE (2002) apostam em modelos de redes neurais artificiais capazes de lidar com a situação de dados ausentes. A idéia principal é encontrar uma configuração apropriada para as entradas e saídas da representação neural. O algoritmo apresentado adapta a estrutura da rede para dados individuais com componentes ausentes, então seleciona o estado de proteção dos neurônios da rede de acordo com as partes ausentes dos vetores de dados. A proteção do neurônio é feita através de um *flag* que indica se o neurônio está protegido ou não. Se o neurônio estiver protegido significa que o mesmo não será considerado no cálculo dos pesos.

ORRE et al. (2003) descrevem uma rede neural recorrente modificada para lidar com dados ausentes na fase de treinamento. A rede possui dois modos de operação: treinamento e *recall*. O trabalho apresenta métodos não supervisionados para reconhecimento de padrão que incluem mudanças na fase de *recall* do algoritmo permitindo lidar com dados de treinamento com valores ausentes.

JIANG, CHEN e YUAN (2005) propõem um modelo de redes neurais artificiais para classificação de dados incompletos. A base de dados incompleta é dividida em grupos de base de dados completos, que são utilizados como conjuntos de treinamento. O método proposto utiliza toda a informação fornecida pelo dado com valores ausentes, mantém máxima consistência do dado incompleto e evita a dependência de suposições da distribuição

Outra proposta apresentada em NELWAMONDO e MARWALA (2007) utiliza conjuntos de Fuzzy-ARTMAPs para a tarefa de classificação e *multi-layer perceptrons* para a

tarefa de regressão como estimador. O método proposto é adequado para operações de estimações on-line usando as redes neurais previamente treinadas.

INGRASSIA e DOMMA (2000) relacionam os modelos estatísticos e neurais. O treinamento de uma rede RNA é comparado a um problema de estimação ML.

Um trabalho de comparação das técnicas de redes neurais auto-associativas combinadas com algoritmos genéticos e as técnicas de EM-*Expectation Maximization* é apresentado por NELWAMONDO, MOHAMED e MARWALA (2007). A conclusão da comparação enfatiza que as técnicas EM apresentam desempenho superior quando há pouca ou nenhuma interdependência entre as variáveis de entrada, enquanto as técnicas de RNA e algoritmos genéticos apresentam um desempenho melhor quando há relacionamentos não lineares inerentes entre algumas variáveis de entrada, e que é uma realidade em problemas reais de mineração de dados.

2.3.4 Procedimentos baseados na manipulação direta dos dados ausentes

Os métodos de manipulação direta do dado ausente conseguem tratar os valores ausentes, sem a necessidade de imputação. Nesta categoria destacam-se os algoritmos baseados em **árvores de decisão**. Os métodos baseados em árvores não fazem nenhuma suposição sobre a forma de distribuição dos dados e não exigem uma especificação estruturada do modelo, ou seja, são métodos não paramétricos. A baixa correlação dos atributos pode influenciar negativamente o desempenho desses algoritmos. A preocupação de tornar os algoritmos robustos e eficientes diante da ausência de parte dos dados é evidenciada em (QUILAN, 1993).

FORTES, MORA-LÓPEZ e TRIGUERO (2006) apresentam um método para lidar com dados ausentes baseado no algoritmo de Árvore de Decisão de Indução *Top-Down*. Três aspectos são observados nesse trabalho: critérios de divisão, alocação de valores para os atributos com valores ausentes e predição de novas observações.

Em HEWETT (2004) é investigado um método para lidar com dados ausentes que utiliza heurísticas do aprendizado de máquina, que não requerem conjuntos completos para as inferências. É apresentado um sistema de aprendizado (*SORCER-Second-Order Relation Compression for Extraction Rules*) que induz um conjunto de regras da base de dados representado como uma tabela de decisão de segunda ordem, ou seja, relações de banco de

dados nas quais as linhas possuem conjuntos de valores atômicos como componentes. Esses conjuntos de valores são interpretados como disjunções que possuem representações compactas, facilitam a administração e aumentam a compreensibilidade. O SORCER é baseado no *framework* teórico de relações de segunda ordem, no qual lidam com dados ausentes de forma única, utilizando o conjunto vazio como uma representação natural.

WEISS e INDURKHYA (1999) discutem métodos para aprendizado e aplicação de regras de decisão para classificação de dados com muitos valores ausentes. Os autores apresentam um método para induzir regras de indução a partir de registros com dados ausentes onde o formato da regra não difere das regras do dado completo e nenhuma característica especial é especificada para preparar o dado original ou aplicar as regras induzidas. O método gera regras na forma normal disjuntiva. As regras de decisão são relacionadas às árvores de decisão. O nodo terminal de uma árvore pode ser agrupado em regras na forma normal disjuntiva (DNF), em que somente uma regra é satisfeita para um novo caso. As regras de decisão são também regras DNF que permitem sobreposição de regras, permitindo ter um conjunto de regras mais compactas.

GORODETSKY, KARSAEV e SAMOILOV (2005) apresentam um método que gera dois conjuntos de regras utilizados como limite inferior e superior para outros conjuntos de regras correspondentes às associações arbitrárias de dados ausentes. Assim, baseado nesses conjuntos, o subconjunto de regras é selecionado para ser utilizado na classificação. Este método é indicado para aplicações onde a imputação não é teoricamente justificada, como por exemplo no aprendizado de detecção de intrusos em uma rede baseada em *data streams* assíncronos oriundos de múltiplas fontes.

Já SCHONER (2004), motivado pelo fato da imputação introduzir dependências ou regularidades não presentes no dado real, explora novas direções através da utilização de **SVM (*Support Vector Machine*) e máxima entropia**, e conclui que esses métodos podem melhorar a predição para uma base de dados incompleta. O SVM é um método de aprendizado supervisionado usado para classificação e regressão. Uma propriedade desse método é que ele simultaneamente minimiza o erro de classificação empírico e maximiza a margem geométrica. O treinamento SVM sempre encontra um mínimo global, e sua interpretação geométrica possibilita terreno fértil para outras investigações. Por outro lado, a entropia é uma medida da falta de informação de uma distribuição de probabilidades (SHANNON, 1948). O princípio da máxima entropia permite selecionar a partir de uma família de distribuições consistentes a distribuição que maximiza a entropia (JAYNES, 1957).

Essa teoria é capaz de determinar distribuições de probabilidades com pequenas amostras de dados, o que a torna eficiente em estudos onde os dados são escassos.

Outras contribuições exploram a teoria de ***rough sets***, introduzida por PAWLAK (1982). Esta teoria possui propriedades que permitem eliminar variáveis ou atributos irrelevantes através do processo de redução do sistema de informação, baseando-se na definição de redutos, os quais são subconjuntos de atributos capazes de manter as mesmas propriedades de representação do conhecimento quando esta é feita utilizando todos os atributos. Os objetos contidos em um sistema, de acordo com suas características, são agrupados em classes. Os objetos de uma classe são indiscerníveis entre si. Esta teoria é capaz de administrar imprecisões, informações ruidosas e incompletas. O objetivo da teoria de *rough sets* é a redução do sistema de informação e consequentemente a geração de regras. Entretanto, a formalização clássica da teoria de r*ough sets* não trata o problema do valor ausente. Os primeiros métodos r*ough sets* para valores ausentes de atributos foram os algoritmos LEM1 e LEM2 (GRZYMALA-BUSSE, 2003). Esses algoritmos de regras de indução utilizam aprendizado por exemplo (*LERS-Learning from Examples*) baseado na idéia de blocos de pares de valores de atributos. Esses blocos são usados para construir conjuntos de características, relações de características e aproximações *upper* e *lower* para tabelas de decisão com valores ausentes nos atributos.

LI e CERCONE (2006) apresentam o método RSFit que prediz valores ausentes baseado na teoria de *rought sets* e *itemsets* frequentes.

Outra alternativa para lidar com dados ausentes dentro dessa teoria é o método de decomposição, no qual nem o processo e nem a relação de indiscernibilidade são modificados. Na decomposição a base de dados é decomposta em subconjuntos sem valores ausentes. Então, métodos de classificação por indução são aplicados nesses conjuntos. Finalmente, métodos de resolução de conflitos são utilizados para obter a classificação final a partir das classificações parciais. O método de decomposição introduz dificuldades na interpretação do último passo que é a resolução de conflitos. Normalmente utiliza-se uma estratégia de votação para soluciná-los. O melhoramento da decomposição em comparação com outros trabalhos é evitar a necessidade de combinar vários classificadores (LATKOWSKI e MIKOLAJEZYK, 2004).

A **reconstrução parcial** proposta por AGGARWAL e PARTHASARATHY (2001) cria representações conceituais com dimensões reduzidas. O método utiliza matriz de covariâncias e componentes principais para obter os conceitos que retêm o mínimo possível de informações contidas nas variáveis originais. A matriz de covariâncias não pode ser

computada sobre uma base de dados com valores ausentes, então utiliza-se a estimação de valores ausentes baseada no algoritmo EM.

As **regras de associação** também são utilizadas como técnica para lidar com dados ausentes. Os dados ausentes trazem problemas para os algoritmos baseados em regras de associação, pois distorcem as medidas de suporte e confiança e implicam na perda de boas regras. Uma inovação para lidar com os dados ausentes foi a criação das medidas de representatividade e extensibilidade que complementam as medidas de suporte e confiança. A representatividade indica a quantidade de *itemsets* que não possuem valores ausentes nos atributos. Um *itemset* é extensível se possuir um *superset* frequente e representativo. Essas novas medidas foram propostas devido às distorções causadas nas medidas de suporte e confiança, ou seja, ocasionadas pelo dado ausente. Uma implementação dessas novas medidas é encontrada em CALDERS, GOEHALS e MAMPAEY (2007), através do algoritmo *Xminer*.

Por sua vez, NAYAK e COOK (2001) apresentam o algoritmo de regras aproximadas que considera a existência de dado ausente e ruidoso.

Por último, MAMPAEY (2006) ressalta que as regras geradas podem conter itens com ausências e que essa característica pode ser explorada para minerar regras interessantes sobre o mecanismo de ausência que afeta a base de dados. Um item é um valor de atributo escrito como $(A=a_i)$. A ausência no item é escrita como $(A=?)$, indicando um valor ausente para o atributo A. Dessa forma, regras no formato $(A=a_i) => (B=?)$ podem ser encontradas.

Resumidamente, a Tabela 5 apresenta os principais métodos revisados e suas respectivas suposições de mecanismo de ausências.

Tabela 5

Métodos e mecanismos

Método	**Mecanismo**
Imputação pela média	MCAR
Listwise deletion	MCAR
Imputação múltipla	MAR
Imputação por regressão	MAR
Modelos bayesianos	MAR
EM	MAR
Maximum likelihood	MAR
Seleção de modelos	NMAR
Mistura de padrões	NMAR

2.4 Considerações sobre os procedimentos

Além da abordagem estatística, nota-se na área de mineração de dados uma preocupação com dados ausentes. PEARSON (2005a, 2005b) alerta que, além do problema do dado ausente, existe também o problema do dado ausente disfarçado. Segundo o autor, dado ausente disfarçado é o dado codificado como válido. O dado está explicitamente representado, porém o valor compromete a interpretação. O principal efeito é a introdução de distorções significativas nos resultados das análises. HUA e PEI (2007) desenvolveram um método capaz de identificar valores frequentes de dados ausentes disfarçados. O método é baseado no modelo de distribuição do dado ausente disfarçado e na heurística que utiliza amostras não distorcidas.

Na literatura de mineração de dados há uma preocupação no tratamento dos dados ausentes na fase de preparação dos dados e no desempenho da aplicação dos algoritmos.

Em BROWN e KROS (2003) foi avaliado o impacto dos dados ausentes nos principais algoritmos de mineração de dados.

Alguns trabalhos apresentam *frameworks* para a escolha automática do método para o tratamento do dado ausente. ZOU (2005) apresenta um *framework* baseado em número de registros, número de atributos, número de atributos simbólicos, entropia da base de dados, número de classes e taxa de dados ausentes. SHALABI, NAJJAR e KAYED (2006) propõem um *framework* que implementa quatro técnicas de tratamento de dados ausentes, e a escolha da técnica a ser utilizada baseia-se na teoria dos *rough sets*.

De uma forma geral, na literatura de dados ausentes no contexto de mineração de dados são poucos os trabalhos que consideram o mecanismo de ausência, dentre eles destacam-se HULSE, KHOSHGOFTAAR e SEIFFERT, (2006), SONG e SHEPPERD, (2007), GARCÍA e HRUSCHKA (2005), BATISTA e MONARD (2003). Nesse contexto os trabalhos podem ser divididos em duas categorias: comparação de técnicas e proposta de métodos.

Vários trabalhos efetuam a comparação de métodos para lidar com dados ausentes, entre eles GRZYMALA-BUSSE e HU (2000), BATISTA e MONARD (2003), LIU, LEI e WU (2005), HULSE, KHOSHGOFTAAR, SEIFFERT (2006), TWALA, CARTWRIGHT, SHEPPERD (2006), SONG, SHEPPERD (2007), TSECHANSKY e PROVOST (2007).

ACUÑA e RODRIGUEZ (2003) avaliam o efeito da taxa de erro de classificação dos classificadores LDA-Análise de Discriminante Linear e KNN frente aos seguintes métodos de

tratamento de dados ausentes: *case deletion*, imputação pela média, imputação pela mediana e KNN.

FARHANGLAR, KURGAN e PEDRYCZ (2004) comparam os métodos de imputação pela média, *hot-deck*, algoritmos probabilísticos, árvores de decisão e regras de decisão considerando as seguintes características do banco de dados: número de registros, número de atributos, proporção de atributos booleanos e número de classes.

DELAVALLADE e DANG (2007) propõem uma nova técnica baseada na medida de entropia, que encontra para cada valor ausente um valor de substituição com maior poder de discriminação. Além disso, propõem uma nova taxonomia para os métodos dividindo-os em: espaço de observação ou espaço de variável, iterativo ou não iterativo, informação local ou global, estocástico ou determinístico, modelo de predição e informação de classe.

Uma modificação do algoritmo *K-means* proposta por WAGSTAFF (2004) incorpora informação do dado ausente. Esse algoritmo chamado de KSC (*K-means with soft constraint*) utiliza os dados completos e não descarta os dados com valores ausentes. O dado completo é utilizado para o agrupamento e o dado ausente para gerar um conjunto de restrições para o algoritmo de agrupamento. Uma demonstração prática do algoritmo KSC é encontrada em WAGSTAFF e LAIDLER (2005).

2.5 Considerações sobre os mecanismos

A identificação dos mecanismos de ausência é uma tarefa importante, pois as propriedades dos métodos estatísticos que lidam com dados ausentes dependem fortemente da natureza desses mecanismos (LITTLE e RUBIN, 2002). Nenhum método pode ser considerado ótimo para todas as situações de dado ausente, e quase sempre os métodos estatísticos fazem determinadas suposições sobre o mecanismo de ausência.

O tratamento de dados com padrão de ausência MAR ou NMAR aumenta consideravelmente a complexidade do problema, pois nestes casos a ausência acontece por uma conjunção de fatores em princípio desconhecidos. Assim, a menos que se conheçam as condições nas quais os dados se tornaram MAR ou NMAR, tratá-los como MCAR apresenta-se como uma boa primeira opção (SOARES, 2007).

COLLINS, SHAFER e KAM (2001) investigam formas de descobrir o mecanismo de ausência através da incorporação de variáveis auxiliares. As estratégias foram denominadas

restritivas e inclusivas. As variáveis são incluídas para melhorar o desempenho dos métodos de imputação. A estratégia restritiva incorpora poucas variáveis auxiliares, enquanto a estratégia inclusiva utiliza todas ou quase todas as variáveis auxiliares disponíveis. Assim é possível identificar subcategorias do mecanismo MAR, as quais são chamadas de: MAR-linear, MAR-convex e MAR-sinister. Por exemplo, sejam *X* e *Y* variáveis observadas, *Y* uma variável com valores ausentes e *Z* uma variável auxiliar que contém a possível causa da ausência em *Y*. Para o mecanismo MCAR, os valores ausentes são impostos a *Y* independentemente de *X*, *Y* e *Z*. Para o MAR-linear, a probabilidade de ausência é linearmente relacionada a *Z*. Para o MAR-convex, a probabilidade de ausência é maior nos extremos de *Z* e menor no meio. Já o MAR-sinister, a probabilidade de ausência é uma função correlação entre *Z* e *X*. Os autores concluem que a adoção de estratégias inclusivas ajudam a esclarecer as causas da ausência. Além disso, apontam outras variedades plausíveis de MAR.

Outro trabalho que investiga variedades do mecanismo MAR é apresentado em PREISSER, LOHMAN e RATHOUZ (2002), o qual aponta os mecanismos MAR-Strong e MAR-Weak relacionados, respectivamente, com alta e baixa correlação de variáveis.

No mundo real é difícil existir uma ausência puramente MAR ou puramente NMAR, ou seja, MAR e NMAR podem coexistir em uma base de dados com valores ausentes (GRAHAM, 2007). Essa observação leva a uma nova investigação em trabalhos futuros, de um quarto mecanismo de ausência denominado HMAR (*hybrid missing at random*). Suponha uma base de dados com os atributos sexo e idade, onde os valores ausentes estão associados ao atributo idade. Quando a ausência da idade está relacionada ao atributo sexo, tem-se o mecanismo unicamente MAR. Quando a ausência da idade está relacionada ao próprio valor da idade, então têm-se o mecanismo unicamente NMAR. Entretanto em situações reais é aceitável que a ausência esteja relacionada a um conjunto de fatores, como por exemplo, indivíduos do sexo feminino e com idade superior a 30 anos não informam a idade, neste caso tem-se o mecanismo híbrido HMAR.

O mecanismo NMAR é o padrão de ausência mais complexo. O mecanismo NMAR também é chamado de mecanismo não ignorável, porque os modelos de análise de dados convencionais não podem lidar eficientemente com esse tipo de ausência. É muito difícil de identificar a partir do dado observado a existência do mecanismo NMAR, assim a literatura recomenda a exploração dos resultados utilizando a análise de sensibilidade (JASEN et al., 2006). A análise de sensibilidade é usada para avaliar o grau de confiança dos resultados em situações de decisões incertas ou suposições sobre os dados e resultados utilizados.

De um modo geral, é uma boa prática aplicar a análise de sensibilidade ao empregar diferentes técnicas de tratamento de dados ausentes, assim assegura-se a robustez das conclusões de cada método. MOLENBERGHS et al. (1999) salienta que a análise de sensibilidade para a especificação de um modelo é uma necessidade, e conhecimento subjetivo pode desempenhar um papel crítico.

Devido à incerteza oriunda do dado ausente não ignorável (NMAR), deve-se tomar grande precaução ao se concluir qualquer inferência sobre um modelo para esse mecanismo.

CHANG (2005) demonstra a distorção do resultado ao tratar dado NMAR como se fosse MAR através de pontuações de propensão. A pontuação de propensão proposta por ROSENBAUM e RUBIN (1983) é a probabilidade condicional de observar as variáveis alvos Y dado as covariáveis X. Em estudos observacionais, as pontuações de propensão são estimadas a partir do dado observado utilizando regressão logística .

Por sua vez, MOLENBERGHS et al. (2002) demonstram que não é possível distinções empíricas entre MAR e NMAR. Através de ajustes, um modelo NMAR pode ser reproduzido exatamente por uma contrapartida MAR.

Outra linha de investigação aponta a visualização da ausência como meio de auxiliar a identificação do mecanismo de ausência, como evidenciado em WANG e WANG (2007) e TEMPL e FILZMOSER (2008).

Por fim, KONING, FINKER e DAIMER (2005) apresentam como solução para o mecanismo NMAR os modelos *Heckman* e *Tobite*, salientando que os modelos são aplicáveis à modelos lineares e para modelos não lineares não há nenhuma ferramenta disponível.

2.6 Considerações finais

Em suma, não há uma conclusão sobre os métodos de tratamento de dados ausentes. Não é possível obter uma orientação de qual método é apropriado para uma determinada configuração de dados que apresentam valores ausentes. A literatura estatística recomenda os métodos de imputação múltipla e ML. O grande problema dessas técnicas é que necessitam fortes suposições de modelo, que é difícil de justificar no contexto do KDD. Já a literatura de *machine learning* apresenta uma variedade de métodos, dentre eles, KNN, RNA, regras de associação, *rough sets*, árvores de decisão e classificadores bayesianos. No contexto da

mineração de dados não há um guia coerente de métodos a serem utilizados em bases de dados com valores ausentes (LATKOWSKI e MIKOLAJCZYK, 2004).

Uma vez que a identificação do mecanismo de ausência não é uma tarefa trivial, e que os métodos de tratamento de dados ausentes são na maioria das vezes relacionados ao mecanismo, e que o mecanismo NMAR é o mais difícil de lidar e também o mais provável de ocorrer, este trabalho apresenta um método para a identificação do mecanismo de ausência baseado na técnica de regras de associação.

3 PROPOSTA DE UM MÉTODO PARA IDENTIFICAÇÃO DO MECANISMO DE AUSÊNCIA BASEADO EM REGRAS DE ASSOCIAÇÃO

Este capítulo tem como objetivo principal apresentar o método RAIMA para identificação do mecanismo de ausência. O método RAIMA é baseado na utilização de regras de associação, assim os principais conceitos referentes as regras de associação são apresentados, como o algoritmo Apriori e as medidas de interesse.

3.1 Considerações iniciais

Como dito anteriormente uma base de dados pode conter atributos com valores ausentes. Essa ausência possui um mecanismo associado, o qual pode ser MCAR, MAR ou NMAR. Em outras palavras a ausência pode ser totalmente aleatória, dependente de algum atributo ou dependente do próprio atributo que possui o valor ausente. Para identificar tais mecanismos é necessário analisar as relações entre os atributos com valores ausentes e os demais atributos. Com esse objetivo, neste trabalho é proposto o uso das regras de associação. Estas regras permitem extrair conhecimento a partir de uma base de dados cuja finalidade é descobrir relações, padrões ou correlações entre itens.

Muitos algoritmos já foram propostos para obter regras de associação, sendo que o mais referenciado é o algoritmo Apriori (AGRAWAL, IMIELINSKI e SWANI, 1993). Basicamente, o algoritmo Apriori encontra os *itemsets* frequentes de uma base de dados e define as regras de associação entre eles, considerando tipicamente medidas de confiança e suporte. Enquanto o suporte restringe a quantidade de regras geradas, a confiança reflete a coesão entre os itens relacionados.

A interpretação e a análise das regras geradas pelos algoritmos de regras de associação é um problema ainda em aberto. Muitos estudos ainda estão sendo propostos em torno deste tema (NEVES, 2002). Algumas contribuições já foram propostas, entre elas a definição de novas medidas de interesse (TAN, KUMAR e SRIVASTAVA, 2002 e GONÇALVES, 2005) e também métodos de visualização das regras geradas (KLEMETTINEN et al., 1994).

Ainda no contexto de regras de associação, uma questão que requer atenção é o tratamento dos atributos numéricos das bases de dados. Os atributos numéricos podem possuir

valores diversos que implicam em grande geração de itens. Sendo assim, alguns métodos de tratamento de valores numéricos, como técnicas de discretização, serão abordados neste capítulo.

Além disso, as regras obtidas através das regras de associação podem ser irrelevantes ou redundantes. Uma das formas de evitar esse problema é efetuar a redução da dimensionalidade do banco de dados. A aplicação das regras de associação em um conjunto reduzido de atributos reduz o número de atributos que aparecem nos padrões descobertos e facilita o entendimento das mesmas.

3.2 Regras de Associação

As regras de associação foram introduzidas por AGRAWAL, IMIELINSKI e SWANI (1993). Primeiramente foram utilizadas para análises de "cestas de supermercados", com os objetivos de ajudar os gerentes a direcionar estratégia de marketing, auxiliar nas estratégias de promoções e reorganizar a disposição dos produtos vendidos nas lojas.

Supondo que cada item de um supermercado seja representado por uma variável booleana, onde 1 indica a presença de um produto na compra e 0 a sua ausência, e cada compra seja representada como uma transação, em uma tabela de booleanos dos produtos adquiridos, as transações podem ser analisadas para se obter padrões de compra de produtos. Estes padrões de compra podem ser representados como regras de associação. As regras de associação são representadas da seguinte forma: A⇒B (lê-se A implica em B), onde *A* é o antecedente e *B* o consequente e *A* e *B* são dois conjuntos de itens distintos na base de dados. Como dito anteriormente, cada regra A⇒B possui duas medidas que determinam sua validade no conjunto de dados e também limitam a quantidade de regras geradas. A medida de suporte refere-se à frequência relativa das tuplas para cada padrão, ou seja, o grau de importância do padrão, e a confiança está relacionada a certeza da regra.

Formalmente, a regra de associação pode ser definida como a seguir. Considere $I = \{i_1, i_2,...,i_m\}$ um conjunto de itens e $D = \{T_1,T_2,...,T_n\}$ um conjunto de transações, onde cada transação *T* é composta por um conjunto de itens, tal que $T \subseteq I$. Cada transação é associada a um identificador, chamado TID. Seja *A* um conjunto de itens. Uma transação *T* contém *A* se e somente se $A \subseteq T$. Uma regra de associação é uma implicação na forma A⇒B, onde $A \subset I$, *B*

$\subset I$, *e* $A \cap B = \varnothing$. A regra A⇒B possui um suporte (Sup), onde Sup é o percentual de transações em *D* que contém $A \cup B$ e a confiança (Conf), a qual corresponde ao percentual de transações em *D* contendo *A* que também contém *B* (AGRAWAL, IMIELINSKI e SWANI, 1993). Ou seja,

$$Sup(A \Rightarrow B) = P(A \cup B) \quad (7)$$

$$Conf(A \Rightarrow B) = P(A \mid B) \quad (8)$$

A obtenção de regras de associação é um processo de dois passos:

1. Encontrar todos os *itemsets* frequentes na base de dados que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário.
2. Utilizar todos os *itemsets* frequentes para gerar as regras de associação. Estas regras devem satisfazer as restrições de suporte mínimo e confiança mínima estabelecidas pelo usuário

Toda a complexidade da regra de associação é determinada pelo primeiro passo, encontrar os *itemsets* mais frequentes.

3.3 Algoritmo Apriori

Vários algoritmos já foram propostos para encontrar os *itemsets* frequentes (HAN, PEI e YIN, 2000), um dos mais referenciados é o algoritmo Apriori.

O algoritmo Apriori é utilizado apenas para regras de associação booleana, se existe ou não na transação, para procurar por *itemsets* frequentes, o que resulta em conjuntos de *itemsets* candidatos e então realiza cálculos para determinar se os mesmos são frequentes ou não.

O Apriori utiliza a propriedade de antimonotonia, que diz que para um *itemset* ser frequente, todos os seus subconjuntos também devem ser frequentes. Isto significa que se um *itemset* A não satisfaz o suporte mínimo, então A não é frequente, então se um item B é adicionado ao *itemset* A, então o resultado desse novo *itemset* ($A \cup B$) não pode ser mais frequente que A. Então, esse novo *itemset* ($A \cup B$) não pode ser frequente.

Outra propriedade empregada pelo Apriori é o método iterativo conhecido como busca *level-wise*, onde k-*itemsets* são utilizados para explorar (k+1)-*itemsets*.

O algoritmo Apriori processa os *itemsets* frequentes através de várias iterações. Cada iteração tem dois passos: geração dos conjuntos candidatos e a poda dos conjuntos candidatos.

Em cada iteração o Apriori constrói o maior conjunto de candidatos, conta o número de ocorrências de cada candidato e determina quais *itemsets* satisfazem o suporte mínimo definido pelo usuário.

Na primeira fase da primeira iteração, o conjunto obtido contém todos os 1-*itemsets* (todos os itens do banco de dados). O algoritmo conta o suporte para cada item em todo o banco de dados. No final desta fase, o algoritmo encontrará todos os 1-*itemsets* frequentes que satisfarão o suporte mínimo definido pelo usuário, estes *itemsets* são denominados C_1.

No primeiro passo da primeira iteração, todos os itens são candidatos. O algoritmo simplesmente percorre todo o banco de dados e gera uma lista de candidatos, denominada C_1. No próximo passo o algoritmo conta quantas ocorrências cada candidato teve no banco de dados e, baseado no suporte mínimo, seleciona os 1-*itemsets* mais frequentes, gerando L_1.

Na segunda iteração todos os pares de itens são candidatos. Baseado no conhecimento adquirido da primeira fase dos *itemsets* infrequentes, o algoritmo Apriori reduz o conjunto dos *itemsets* candidatos (propriedade antimonotonia) descartando todos os 2-*itemsets* que não podem ser frequentes. O resultado da segunda iteração será o conjunto de todos os 2-*itemsets* frequentes que satisfaçam o suporte mínimo, denominado C_2.

Para descobrir os 2-*itemsets* mais frequentes o algoritmo utiliza a união do resultado L_1 com ele mesmo ($L_1 \times L_1$) para gerar os próximos candidatos a 2-*itemsets* mais frequentes. O operador × é junção de dois conjuntos. Então C_2 consiste do 2-*itemset* gerado a partir de $L_1 \times L_1$. Gerado a lista C_2 o algoritmo conta a ocorrência de cada candidato na base de dados e seleciona os 2-*itemsets* mais frequentes que satisfaçam o suporte mínimo, gerando assim o L_2. O algoritmo repete a mesma iteração para descobrir C_3 a partir de L_2.

Uma vez definidos todos os *itemsets* frequentes, o próximo passo é definir as regras de associação entre os mesmos. Esta fase é relativamente simples e direta. Nesta fase é necessário apenas analisar as possíveis associações entre os *itemsets* descobertos pelo algoritmo Apriori e selecionar regras que satisfaçam a confiança mínima exigida pelo usuário.

Passos simplificados do Apriori:

1. Dados de entrada: coleção de dados, suporte mínimo, confiança mínima
2. Considerar k=1 para criação de k-*itemsets*

3. Analisar os dados e criar uma tabela de k-*itemsets* com suporte acima do suporte mínimo
4. Criar com os *itemsets* filtrados um conjunto de candidatos a (K+1) *itemsets*. Usar a propriedade do Apriori para eliminar *itemsets* infrequentes
5. Repetir a partir do passo 3 até que o conjunto gerado seja vazio
6. Listar as regras de associação, aplicar limites de confiança e filtrar regras

Muitos melhoramentos foram sugeridos no algoritmo Apriori, entretanto mesmo assim, a complexidade computacional do algoritmo Apriori é exponencial. Entre eles, destaca-se o algoritmo *Frequent pattern growth* (*FP-growth*) que procura pelos *itemsets* frequentes sem gastar tempo com a geração dos candidatos a *itemsets* frequentes (HAN, PEI e YIN, 2000). O algoritmo *FP-growth* adota a estratégia de "dividir para conquistar". Ele detecta, na base de dados, todos os itens que são frequentes e os armazena na memória principal, em uma estrutura compacta chamada de *frequent-pattern tree* (*FP-tree*). O algoritmo *FP-growth* é mais rápido que o Apriori. Várias técnicas de otimização foram adicionadas a ele, o que torna ainda mais rápido.

3.4 Medidas para avaliação de regras

Um dos grandes problemas na utilização das regras de associação é a interpretação e a análise do grande volume de regras que são potencialmente produzidas. Segundo TOIVONEN et al. (1995) a interpretação das regras descobertas pode ser um novo problema de *data mining,* devido à dificuldade de extrair conhecimento da grande quantidade de regras que podem ser produzidas. TAN, KUMAR e SRIVASTAVA (2002) apresentam uma revisão de 21 medidas de interesse utilizadas na avaliação das regras de associação. Segundo os autores, não há uma medida que seja consistentemente melhor em todas as situações. A medida deve ser selecionada de acordo com suas características e com as expectativas do especialista no domínio do problema. No anexo I encontra-se um quadro resumo contendo as principais medidas de interesse.

As medidas de interesse são classificadas em *objetivas* e *subjetivas*. As medidas de interesse objetivas são índices estatísticos utilizados para selecionar regras interessantes descobertas pelo algoritmo de regras de associação (GONÇALVES, 2005). O suporte e

confiança são exemplos de medidas de interesse objetivas. As medidas que medem a dependência entre os itens (*lift*, *Rule Interest* e convicção) também são consideradas objetivas.

As medidas de interesse subjetivas dependem da avaliação do usuário. Alguns exemplos: a medida de utilidade (*actionability*) considera que uma regra é interessante se o usuário pode fazer algo a partir da regra descoberta. A medida de inesperabilidade (*unexpectedness*) diz que uma regra é interessante quando contradiz a expectativa do usuário. Finalmente, a medida de novidade (*novelty*), quando revela uma regra completamente nova para o usuário (NEVES, 2002).

A seguir uma revisão das principais medidas objetivas será apresentada.

O **suporte** é uma medida de significância do *itemset*. A desvantagem do suporte é o problema do item que ocorre raramente em uma base de dados. Os itens infrequentes são podados deixando de produzir regras interessantes e potencialmente valiosas.

$$Sup(A \Rightarrow B) = Sup(A \cup B) \tag{9}$$

A **confiança** é definida como a proporção de exemplos cobertos pelo antecedente que são também cobertos pelo consequente. A medida de confiança deve ser alta para que exista coesão entre os itens. Uma confiança baixa não reflete nenhum padrão de comportamento. O problema da confiança é que ela é sensível à frequência do consequente no banco de dados. Consequentes com alto suporte automaticamente produzem altos valores de confiança mesmo que não existam associações entre os itens.

$$Conf(A \Rightarrow B) = \frac{Sup(A \cup B)}{Sup(A)} \tag{10}$$

A **cobertura** também chamada de suporte do antecedente, define a proporção de exemplos cobertos pelos itens que compõem o antecedente da regra.

$$Cobertura(A \Rightarrow B) = Sup(A) \tag{11}$$

A **convicção** compara a probabilidade que A apareça sem B se eles são dependentes da atual frequência da presença de A sem B. Esta medida tenta capturar o grau de implicação

entre A e B, é unidirecional, ou seja, convicção (A⇒B) é diferente de convicção (B⇒A). A convicção permite definir a independência do antecedente A, face ao consequente B (BRIN et al, 1997).

$$Convicção(A \Rightarrow B) = \frac{Sup(A)(1 - Sup(B))}{Sup(A) - Sup(A \cup B)} \tag{12}$$

A medida de interesse, também chamada de ***lift***, mede quantas vezes A e B ocorrem juntos se são estatisticamente independentes, e não mede a implicação. (BRIN et al ,1997). Quando o valor é igual a 1 indica que A e B são independentes. Quanto maior o valor do lift, mais a existência de associação entre A e B não é uma transação aleatória, e sim devida a algum relacionamento entre os itens.

$$Lift(A \Rightarrow B) = \frac{Conf(A \Rightarrow B)}{Sup(B)} = \frac{Sup(A \cup B)}{Sup(A)Sup(B)} \tag{13}$$

A medida ***leverage*** é a diferença entre a proporção de exemplos cobertos, simultaneamente, pelo antecedente e pelo consequente da regra e a proporção de exemplos que seriam cobertos se o antecedente e o consequente fossem independentes (PIATETSKY-SHAPIRO, 1991).

$$Leverage(A \Rightarrow B) = Sup(A \cup B) - (Sup(A)Sup(B)) \tag{14}$$

O teste do χ^2 (qui-quadrado) pode ser uma forma de medir a correlação entre o antecedente e o consequente de uma regra (LIU, HSU e MA, 1999). Este teste permite identificar a direção da correlação, positiva, negativa ou independência. Baseia-se na comparação das frequências observadas com as frequências esperadas. Quanto mais próximas estas freqüências, maior será a probabilidade de se tratarem de casos independentes. Considerando f_o uma frequência observada e f uma frequência esperada, o valor χ^2 é definido pela seguinte fórmula:

$$\chi^2 = \sum \frac{(f_o - f)^2}{f} \tag{15}$$

O índice ***Jaccard*** foi inicialmente proposto para medir a similaridade de espécies de flora (JACCARD, 1912). O índice ***Jaccard***, também chamado de coeficiente de similaridade ***Jaccard***, é usado para comparar a similaridade e diversidade dos elementos dos conjuntos. O índice ***Jaccard*** é definido como o número de elementos da interseção de dois conjuntos dividido pela união dos conjuntos, ou seja:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{16}$$

Sejam *A* e *B* dois conjuntos, *A = {2, 3, 4, 6,7}* e *B = {1, 4, 5, 7, 8}*, então $A \cup B$ *= {1, 2, 3, 4, 5, 6, 7, 8}* e $A \cap B$ *= {4,7}*, o coeficiente ***Jaccard***, é assim calculado:

$$J(A,B) = \frac{\text{elementos da interseção}}{\text{elementos da união}} = \frac{2}{8} = 0{,}25$$

Outra medida complementar ao índice ***Jaccard***, é a medida distância ***Jaccard***, que mede a dissimilaridade entre os elementos dos conjuntos, assim definido,

$$J\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{17}$$

No contexto das regras de associação, o índice ***Jaccard*** é adaptado para medir o grau de sobreposição entre os casos A e B, através da distância entre A e B pela fração entre os casos cobertos pelos dois e os casos cobertos por um só (A ou B). O índice ***Jaccard*** varia entre 0 e 1, sendo que valores altos indicam sobreposição de casos cobertos.

$$Jacc(A \Rightarrow B) = \frac{Sup(A \cup B)}{Sup(A) + Sup(B) - Sup(A \cup B)} \tag{18}$$

A maior parte das medidas tem por objetivo medir a independência entre o antecedente e o consequente das regras. Estas medidas podem ser utilizadas como forma de filtrar as regras de associação que obedeçam a determinados valores de independência, permitindo reduzir o número de regras produzidas.

Neste trabalho, a medida escolhida para identificar as regras responsáveis pela ausência foi o índice ***Jaccard***. Esta escolha baseou-se no objetivo da medida. O índice ***Jaccard*** mede a sobreposição entre o antecedente e o consequente, e quanto maior o índice, maior é a tendência do antecedente e consequente cobrirem os mesmos itens.

3.5 Técnicas de discretização

Os atributos numéricos possuem uma diversidade de valores, por isso quando as regras de associação são utilizadas em bancos de dados com atributos numéricos, podem originar inúmeros itens e tendem a produzir muitas regras sem valor de interesse e demasiadamente específicas.

As técnicas de discretização são usadas para reduzir e categorizar o número de valores para um dado atributo, cria-se intervalos de valores ou categorias que substituem os valores atuais do atributo (LIU et al.,2002).

As técnicas de discretização podem ser divididas em abordagens supervisionadas e não supervisionadas. Na abordagem supervisionada um atributo comando o processo de discretização, enquanto na abordagem não supervisionada o processo é independente de outros atributos. Algumas técnicas da discretização supervisionada são: Fayyad e Irani (orientada por Entropia), Intervalo de Classes e Chi-Merge. A discretização não supervisionada apresenta as técnicas: Agrawal e Srikant, *equi-depth*, *equi-width* e baseada em distância. A técnica proposta por AGRAWAL e SRIKANT (1994) cria os intervalos a partir de uma medida de perda de informação (*partial completeness*). Essa medida permite definir o número ideal de intervalos em relação à perda de informação admitida. A discretização *equi-width* particiona os valores em partições de tamanhos iguais. Este tipo de discretização não leva em consideração as classes do conjunto de dados. Quando o tamanho do intervalo é muito pequeno ou muito grande para o atributo, várias instâncias de classes diferentes podem ficar no mesmo intervalo. Na discretização *equi-depth*, cada partição possui o mesmo número de elementos. Por sua vez, quando a discretização é baseada na distância, as partições são mais significantes, uma vez que se considera o número de pontos em um intervalo e a coesão dos pontos no intervalo.

Outro tratamento interessante para os dados numéricos no contexto das regras de associação é a analise da distribuição dos valores dos atributos numéricos, com o objetivo de

identificar subpopulações que se distinguem em relação à população geral, esta análise pode ser feita através de análises de histogramas, ou ainda utilizando o teste g*oodness of fit*, que compara as distribuições pela assimetria (s*kewness*) e afunilamento (k*urtosis*).

Os atributos categóricos podem também serem agrupados quando há uma variedade de valores, como por exemplo, o atributo CEP (código de endereçamento postal) pode ser agrupado em regiões. Desta forma, grupos de atributos categóricos podem ser criados quando o usuário e/ou especialista detém o domínio dos valores do atributo.

3.6 Técnicas de redução de dimensionalidade

As técnicas de redução de dimensionalidade são usadas para reduzir o número de atributos de uma base de dados, através da detecção de atributos irrelevantes ou redundantes. O objetivo é encontrar um conjunto mínimo de atributos que representem a distribuição das classes como na base de dados original. De uma maneira geral, a redução de atributos quase sempre melhora a precisão de modelos em problemas reais, pois torna os modelos mais inteligíveis e ajuda a explicar melhor um problema real. A redução de atributos, no contexto de regras de associação, reduz o número de regras geradas na execução do algoritmo de regras de associação facilitando a análise do resultado. Uma implementação encontrada em PLASSE et al. (2007) é o agrupamento de variáveis correlacionadas com o objetivo de diminuir o número de regras de associação produzidas. Dentre as técnicas de redução de dimensionalidade destacam-se: analise de correlação, ganho de informação, teste do qui-quadrado, abordagem *wrapper*, análise de componentes principais, SVD dentre outras.

A análise de correlação utiliza o **coeficiente de correlação**, também chamado **coeficiente Pearson**, que mede a força e a direção de um relacionamento linear entre 2 variáveis quantitativas.

$$R = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \text{ ou ainda} \tag{19}$$

$$R = \sum_{i=1}^{n} Z_{xi} Z_{yi} \tag{20}$$

Onde Z_x e Z_y representam as observações padronizadas, ou seja, após a subtração da média e divisão pelo desvio padrão. O coeficiente de correlação pode variar entre -1 e 1; quando as amostras são independentes, o seu valor será próximo de zero.

O **coeficiente Spearman** é similar ao coeficiente Pearson, sendo que as amostras são substituídas pelos seus respectivos *ranks*.

$$S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n} \tag{21}$$

Onde D_i representa a diferença entre os *ranks* correspondentes a cada par de observação x_i, y_i.

As associações entre variáveis categóricas dependem do número de categorias das mesmas. Uma forma de cálculo é através da tabela de contingência. O número de linhas e colunas da tabela de contingência depende do número de categorias de cada variável. A partir da tabela de contingência pode-se calcular o qui-quadrado para averiguar se as variáveis são independentes. O qui-quadrado é uma medida de associação entre duas variáveis categóricas que assume valores próximos de zero quando as variáveis são independentes e valores elevados quando existe dependência entre as variáveis. Esta medida não esta limitada ao intervalo [0,1] e o seu valor depende do numero total de observações. Devido a estas limitações, surgiram novas propostas de coeficientes de associações para dados categóricos, entre o coeficiente Cramer e o coeficiente Phi.

O **coeficiente Cramer** é uma medida aferida em uma escala categórica, portanto pode ser aplicado em situação onde a informação encontra-se distribuída por categorias nominais não ordenáveis.

$$C = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{22}$$

Onde n é o número total de observações e k mínimo entre o número de linhas e colunas da tabela de contingência.

O coeficiente Cramer apresenta algumas vantagens, seu valor esta limitado ao intervalo [0,1]. Quando as variáveis são totalmente independentes o seu valor é igual a zero. Quanto maior a associação, maior o valor do coeficiente. Ao contrário do qui-quadrado, o

coeficiente pode ser aplicado para comparar tabelas de contingência de dimensão diferente ou baseadas em amostras de dimensões diferentes. Uma desvantagem é que quando o valor do coeficiente for igual a 1, pode não haver associação perfeita entre as duas variáveis, uma vez que a associação só é perfeita se o numero de linhas for igual ao número de colunas. Por fim, este coeficiente não deve ser comparado diretamente a outros. Por exemplo, o coeficiente Cramer pode ser usado em dados ordinais, mas o seu valor não pode ser comparado ao coeficiente Pearson.

Informações complementares sobre o coeficiente Pearson, coeficiente Cramer e coeficiente Spearman são encontradas em KINNEAR e GRAY (2000).

O **coeficiente Phi**, é similar ao coeficiente Cramer, foi proposto inicialmente para tabelas de contingência 2 x 2. Seja uma tabela de contingência

A	B
C	D

Phi é dado por:

$$Phi = \frac{|\, AD - BC \,|}{\sqrt{(A+B)(C+D)+(A+C)+(B+D)}} \tag{23}$$

O **ganho de informação** mede a qualidade de classificação do atributo, ou seja, quanto um atributo é capaz de separar um conjunto de exemplos em categorias, é dado pela equação:

$$G(X,a) = Entropia(X) - \sum_{v} \frac{|\, X_v \,|\, Entropia(X_v)}{|\, X \,|} \tag{24}$$

Onde v = valores possíveis para a, X_v = subconjunto de X em que o valor de a é igual a v e $|X|$ é o número de elementos de X.

A entropia mede a quantidade de informação de um atributo, caracterizando a impureza de um conjunto de exemplos, é dada por:

$$Entropia(X) = \sum_{i=1}^{c} - p_i \log_2(p_i) \tag{25}$$

Onde X = conjunto de exemplos de um conceito, c = número possíveis para o conceito a e p_i = percentual de exemplos que a é igual a v_i.

Neste trabalho, para ilustrar a redução de dimensionalidade será utilizada a análise de correlação Pearson para atributos numéricos e o coeficiente Cramer para atributos categóricos.

3.7 Proposta do método RAIMA para identificação do mecanismo de ausência

O método RAIMA – Regras de Associação para Identificação do Mecanismo de Ausência, tem como objetivo auxiliar a identificação do mecanismo de ausência em bases de dados que possuam ausências univariadas. A seguir a formulação do método RAIMA será descrita.

Considere uma base de dados original , definida como:

$$X = \begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{21} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{Bmatrix}$$

Onde N corresponde ao número de exemplos (casos) da base de dados e M ao número de características do exemplo.

Seja E o conjunto de exemplos (linhas) da matriz X, definido como: $E = \{E_1, E_2,...,E_N\}$, onde cada exemplo $E_i \in E$ possui um conjunto de características (atributos) definido por: $\{A_1, A_2,...,A_M\}$. Cada atributo A_k é representado por seu conjunto de valores a_{jk} definido por $A_k = \{a_{jk}\}$ para $j=1..N$ e $k=1..M$.

Um atributo A_k pode ser categórico ou numérico. Atributos com valores categóricos podem ser reagrupados segundo uma especificação explícita de uma ordem definida pelo usuário. Por sua vez, os atributos com valores numéricos são discretizados em intervalos.

Supondo que A_M corresponda ao atributo que contém valores ausentes na base de dados original X e que o valor ausente foi recuperado por algum processo de estimação ou imputação, cada exemplo $E_i \in E$ possui agora também um atributo rótulo A_{M+1}= $\{presente, ausente\}$ que indica respectivamente se o a valor x_{jM} para $j=1..N$ estava presente ou ausente na base de dados original.

A Tabela 6 representa uma base de dados com N exemplos e M+1 atributos. Colunas ($A_1,...A_{m+1}$) representam atributos e linhas ($E_1,...,E_N$) representam exemplos. O atributo classe neste contexto é o atributo A_M e o atributo A_{M+1} representa a situação do atributo classe na base de dados original.

Tabela 6

Base de dados na forma de valores de atributos

	A_1	A_2	...	A_M	A_{M+1}
E_1	a_{11}	a_{12}	...	a_{1M}	ausente
E_2	a_{21}	a_{22}	...	a_{2M}	presente
:	:			:	:
E_N	a_{N1}	a_{N2}	...	a_{NM}	presente

O método RAIMA é composto de quatro passos:

i) Aplicar o algoritmo Apriori sobre o conjunto de dados *E* selecionando as regras definidas como:

R_t: $A_K = \{a_{jk}\} \Rightarrow A_{M+1}$= *ausente*, para $\forall$ $k=1..M$ e onde *t* representa o número total de regras extraídas.

ii) Para cada regra extraída, calcular o suporte (*Sup*), a confiança (*Conf*), a cobertura (*Cobertura*) e o índice Jaccard (*Jacc*).

Para melhor entendimento do procedimento RAIMA as regras R_t serão generalizadas como: B $\Rightarrow$ C, então:

$$Sup(B \Rightarrow C) = \frac{frequencia(B \cup C)}{N} = \frac{|B \cup C|}{N}$$

$$Conf(B \Rightarrow C) = \frac{frequencia(B \cup C)}{frequencia(B)} = \frac{|B \cup C|}{|B|}$$

$$Cobertura(B \Rightarrow C) = Sup(B) = frequencia(B)$$

$$Jacc(B \Rightarrow C) = \frac{Sup(B \cup C)}{Sup(B) + Sup(C) - Sup(B \cup C)}$$

iii) Verificar se o mecanismo é MCAR

Se o conjunto $C=\{Conf(R_1), Conf(R_2), ... Conf(R_t)\}$ apresenta valores com variância próxima de zero, onde a variância é definida como:

$$s^2 = \frac{\sum_{i=1}^{t} (Conf(R_i) - Conf(\overline{R}))^2}{t-1} \quad (26)$$

Então o mecanismo de ausência é MCAR

iv) Verificar se o mecanismo é MAR ou NMAR

Verificar o quanto B é coberto por C e vice-versa, usando o índice Jaccard. Selecionar a regra que apresentar o maior índice Jaccard.

$Jacc(R_r) = max\ \{Jacc(R_1), Jacc(R_2),...,Jacc(R_l)\}$, para $1 <= r <= l$.

Observando a regra R_r, identificar o antecedente e respectivo atributo $A_k = \{a_{rk}\}$. Se $A_k = A_M$ então o mecanismo de ausência é NMAR, senão o mecanismo é MAR.

Exceções:

i. Quando os atributos de uma base de dados são altamente correlacionados várias regras com valores próximos ao máximo índice *Jaccard* podem ser encontradas. Isto impede a identificação do caso MAR e pode-se concluir erroneamente que o mecanismo é NMAR. Uma possível forma de contornar esta exceção é efetuar a redução da dimensionalidade, excluindo da análise os atributos altamente correlacionados com o atributo classe.

ii. Quando a base de dados apresenta muitos itens raros (valores infrequentes) e o suporte mínimo é muito baixo, são geradas regras com alto valor de confiança que podem dificultar a identificação do mecanismo MCAR.

3.8 Considerações finais

Neste capítulo foram exploradas as características da técnica de Regras de Associação para a identificação das relações na base de dados. Devido suas características, as regras de associação foram utilizadas no método RAIMA, que tem como objetivo auxiliar na identificação do mecanismo de ausência. A partir do princípio que em uma base de dados os itens podem estar ausentes ou presentes, espera-se que alguma relação possa ser encontrada entre os itens ausentes e os itens presentes. Além disso, foram revisadas as principais medidas

de interesse de uma regra, algumas técnicas de discretização e de redução de dimensionalidade.

No próximo capítulo o método RAIMA será avaliado em três bases de dados que possuem características distintas de correlação entre os atributos.

4 AVALIAÇÃO EXPERIMENTAL

4.1 Considerações iniciais

Este capítulo apresenta a avaliação experimental efetuada para verificar o desempenho do método RAIMA. Os experimentos foram realizados em três bases de dados com características específicas.

As regras de associação foram geradas através do Apriori, disponibilizado no software WEKA 3.4 (WEKA, 2007). O cálculo do ganho de informação de cada atributo também foi obtido no WEKA e os demais cálculos estatísticos foram efetuados no software R 2.5.1 (R Development Core Team, 2007).

4.2 Descrição das Bases de Dados a avaliar

As bases de dados utilizadas neste trabalho foram obtidas no repositório da UCI *Machine Learning* (BLAKE, KEOGH e MERZ, 1999), exceto a base de dados do processo de laminação a frio (ZÁRATE, 1998). No total, 3 bases de dados foram utilizadas nesta avaliação experimental: Banco de dados *Laminação a Frio*, *Mushroom Database, Wisconsin Breast Cancer Database*.

As bases de dados selecionadas foram classificadas pelo grau de correlação entre seus atributos (ver Tabela 7). O critério de seleção de banco de dados pela correlação deve-se ao fato que atributos altamente correlacionados poderiam prejudicar a identificação do mecanismo. Na prática atributos altamente correlacionados podem ser eliminados, reduzindo a dimensionalidade da base de dados.

Tabela 7

Classificação da base de dados em relação à correlação entre os atributos

Base de dados	Número de atributos correlação > 0,7	Número de atributos correlação < 0,7	% de atributos correlação > 0,7	Classificação quanto à correlação
Mushroom	2	19	9,5%	Baixa
Laminação a Frio	2	4	33%	Média
Wisconsin	8	1	88%	Alta

A base de dados "laminação a frio" descreve os principais parâmetros do processo siderúrgico da laminação a frio. A Tabela 8 ilustra as características desta base de dados, que possui 117649 registros e 7 atributos numéricos. O valor da carga de laminação é a variável dependente, calculada a partir dos demais parâmetros. Cabe ressaltar que esta base foi gerada artificialmente através de um modelo teórico (ZÁRATE, 1998).

Tabela 8

Base de dados Laminação a Frio

Atributo	Tipo	Mínimo	Máximo	Média	id
espessura de entrada	numérico	4,6	5,4	5	a1
espessura de saída	numérico	3,492	3,708	3,6	a2
coeficiente de atrito	numérico	0,096	0,144	0,12	a3
tensão a ré	numérico	0,3087	0,5733	0,441	a4
tensão frente	numérico	6,3686	11,83	9,098	a5
tensão de escoamento	numérico	39,14	54,55	46,81	a6
Carga de laminação	numérico	976	2930	1755,5	a7

Como esta base apresenta dados numéricos, um processo de discretização para aplicação do algoritmo de regras de associação é necessário. Os atributos a1, a2, a3 e a4 não foram discretizados, pois, possuem apenas 7 valores distintos, então, cada valor numérico foi tratado como um valor categórico, recebendo os respectivos rótulos, MB (muito baixo), B (baixo), M1 (média 1). M2 (média 2), M3 (média 3), A (alto) e MA (muito alto).

A discretização utilizada para os atributos a6 e a7, foi a técnica *equi-width*, ou seja os intervalos foram subdivididos em 6 tamanhos iguais. Os intervalos receberam os seguintes rótulos: MB (muito baixo), B (baixo), M1 (média 1). M2 (média 2), A (alto), MA (muito alto). A distribuição das classes após a discretização é visualizada na Tabela 9.

Tabela 9

Distribuição das classes – Laminação a Frio

Classe	Número de registros	% Classe
MB	4858	4,129
B	24549	20,866
M1	32381	27,523
M2	26495	22,52
A	26730	22,72
MA	2636	2,24

A base de dados *Mushroom* possui 8124 registros e 23 atributos categóricos. Esta base de dados descreve características de cogumelos e o atributo classe identifica se o cogumelo é comestível ou venenoso. A Tabela 10 lista as principais características da base de dados *Mushroom*. O dicionário dos valores de domínio encontra-se no Anexo II.

Tabela 10

Base de dados *Mushroom*

Atributo	Tipo	Domínio	Id
Class	categórico	e, p	a1
cap-shape	categórico	b, c, x, f, k, s	a2
cap-surface	categórico	f, g, y, s	a3
cap-color	categórico	n, b, c, g, r, p, u, e, w, y	a4
Bruises	categórico	t, f	a5
Odor	categórico	a, l, c, y, f, m, n, p, s	a6
gill-attachment	categórico	a, d, f, n	a7
gill-spacing	categórico	c, w, d	a8
gill-size	categórico	b, n	a9
gill-color	categórico	k, n, b, h, g, r, v, p, u, e, w, y	a10
stalk-shape	categórico	e, t	a11
stalk-root	categórico	b, c, u, e, z, r	a12
stalk-surface-above-ring	categórico	z, f, y, k, s	a13
stalk-surface-below-ring	categórico	f, y, k, s	a14
stalk-color-above-ring	categórico	n, b, c, g, v, p, e, w, y	a15
stalk-color-below-ring	categórico	n, b, c, g, v, p, e, w, y	a16
veil-type	categórico	p, u	a17
veil-color	categórico	n, o, w, y	a18
ring-number	categórico	n, o, t	a19
ring-type	categórico	c, e, f, l, n, p, s, z	a20
spore-print-color	categórico	k, n, b, h, r,o, u, w, y	a21
population	categórico	a, c, n, s, v, y	a22
Habitat	categórico	g, l, m, p, u, w, d	a23

A distribuição das classes da base de dados *Mushroom* é apresentada na Tabela 11.

Tabela 11

Distribuição das classes – *Mushroom*

Classe	Número de registros	% Classe
e	4208	51,8
p	3916	48,2

A base de dados *Wisconsin Breast Cancer* possui 699 registros e 10 atributos categóricos (ver Tabela 12). Esta base de dados descreve características de células mamárias e o atributo classe identifica se a célula é benigna ou maligna.

Tabela 12

Base de dados *Wisconsin*

Atributo	Tipo	Domínio	Id
Clump Thickness	categórico	1,2,3,4,5,6,7,8,9,10	a1
Uniformity of Cell Size	categórico	1,2,3,4,5,6,7,8,9,10	a2
Uniformity of Cell Shape	categórico	1,2,3,4,5,6,7,8,9,10	a3
Marginal Adhesion	categórico	1,2,3,4,5,6,7,8,9,10	a4
Single Epithelial Cell Size	categórico	1,2,3,4,5,6,7,8,9,10	a5
Bare Nuclei	categórico	1,2,3,4,5,6,7,8,9,10	a6
Bland Chromatin	categórico	1,2,3,4,5,6,7,8,9,10	a7
Normal Nucleoli	categórico	1,2,3,4,5,6,7,8,9,10	a8
Mitoses	categórico	1,2,3,4,5,6,7,8,9,10	a9
Class	categórico	2,4	a10

Nesta base 16 registros possuem valores ausentes para o atributo a6, sendo que 2 destes registros pertencem a classe 2 (maligna) e 14 registros à classe 4 (benigna). Neste experimento, os valores ausentes foram substituídos pelo valor mais comum do atributo em relação à classe, para a classe maligna o valor do atributo ausente foi substituído pelo valor 10 e para a classe benigna o valor ausente foi substituído por 1.

A distribuição das classes na base de dados *Wisconsin* é apresentada na Tabela 13.

Tabela 13

Distribuição das classes – *Wisconsin*

Classe	Número de registros	% Classe
2	458	65,5
4	241	34,5

4.3 Procedimentos experimentais para geração artificial da ausência

Para cada base de dados foi criado um novo atributo para indicar a situação de presente ou ausente do atributo classe. As ausências foram geradas artificialmente para os três tipos de mecanismos de ausência: MCAR, MAR, NMAR. As simulações foram executadas com os seguintes percentuais de ausência: 25%, 50%, 75% e 100%. Para o mecanismo MCAR foi considerado 95% de ausência ao invés de 100%. O tipo de ausência utilizado está restrito ao tipo univariado, a ausência ocorre em um único atributo. Para o mecanismo MCAR as ausências foram geradas aleatoriamente de acordo com os percentuais de ausência utilizado em cada experimento. Para a geração de ausência com o mecanismo MAR foi escolhido um atributo que representasse a maior quantidade de casos de ausência após aplicado o critério de ausência (ver Tabela 14). Para o mecanismo NMAR, o atributo classe foi utilizado nos critérios de geração de ausência. A Tabela 14 ilustra a geração de ausência para cada mecanismo e base de dados considerada.

Tabela 14

Geração artificial da ausência nas bases de dados

Base de Dados	Mecanismo	Critério	% Ausência	Registros com atributo classe presente	Registros com atributo classe ausente
Laminação	MCAR	aleatório	25	88237	29412
		aleatório	50	58825	58824
		aleatório	75	29412	88237
		aleatório	95	5882	111767
	MAR	a3 > 0,12	25	105039	12610
		a3 > 0,12	50	92435	25214
		a3 > 0,12	75	79829	37820
		a3 > 0,12	100	67227	50422
	NMAR	a7 > 1755	25	103677	13972
		a7 > 1755	50	89712	27937
		a7 > 1755	75	75748	41901
		a7 > 1755	100	61788	55861
Mushroom	MCAR	aleatório	25	6093	2031
		aleatório	50	4062	4062
		aleatório	75	2031	6093
		aleatório	95	406	7718
	MAR	a22 = v	25	7114	1010
		a22 = v	50	6104	2020
		a22 = v	75	5094	3030
		a22 = v	100	4084	4040
	NMAR	a1= e	25	7072	1052
		a1= e	50	6020	2104
		a1= e	75	4968	3156
		a1= e	100	3916	4208
Wisconsin	MCAR	aleatório	25	524	175
		aleatório	50	349	350
		aleatório	75	174	525
		aleatório	95	35	664
	MAR	a2 < 3	25	579	120
		a2 < 3	50	459	240
		a2 < 3	75	338	361
		a2 < 3	100	218	481
	NMAR	a10 = 2	25	584	115
		a10 = 2	50	470	229
		a10 = 2	75	354	345
		a10 = 2	100	241	458

4.4 Aplicação do método RAIMA nas Bases de Dados

4.4.1 Base de dados Laminação a Frio

A análise exploratória da base de dados Laminação a Frio, resultou na Figura 1, onde é apresentado um histograma para cada atributo da base de dados.

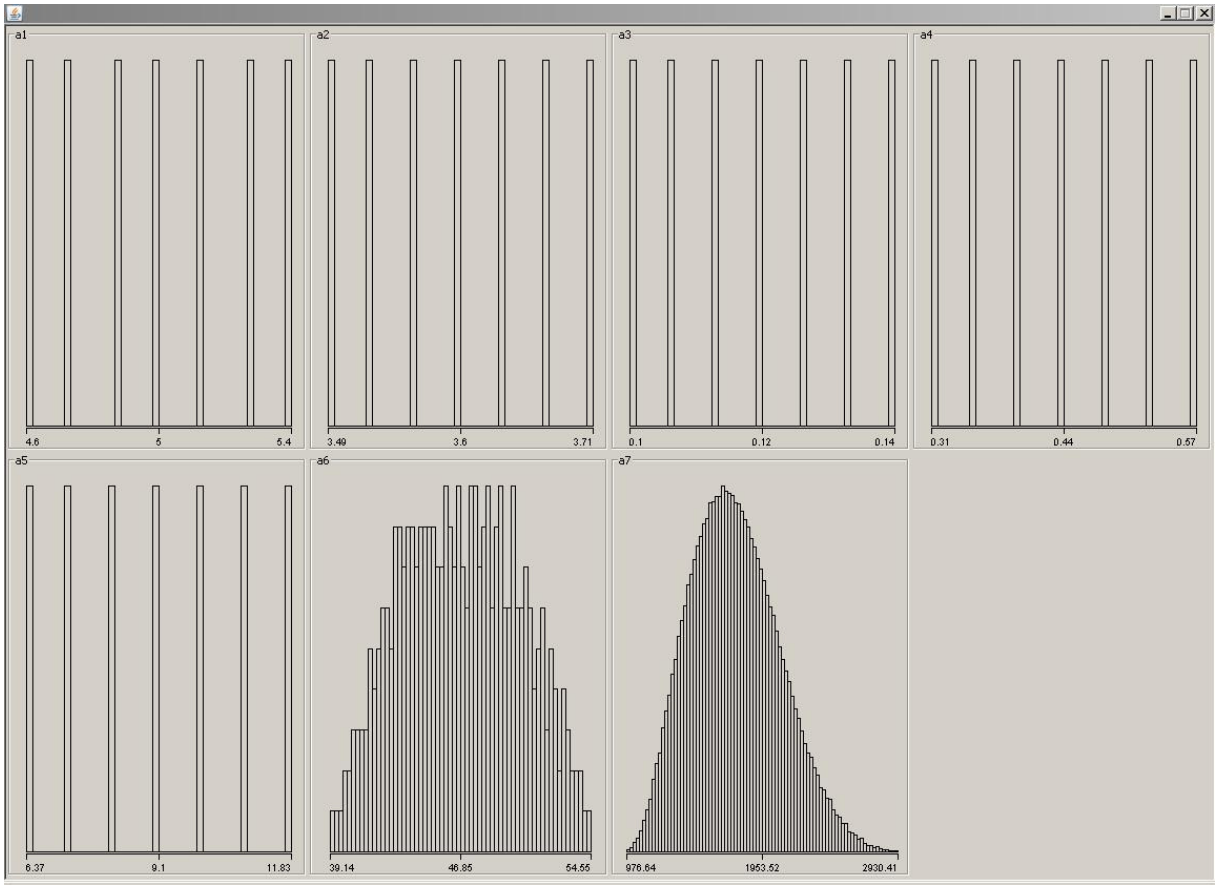


Figura 1 – Histogramas dos atributos da base de dados Laminação a Frio

Pela análise visual dos histogramas nota-se que os atributos a1, a2, a3, a4 e a5 possuem apenas 7 valores distintos, enquanto que os atributos a6 e a7 possuem valores diversos. Devido a isso, aplicou-se a técnica de discretização descrita na Seção 4.2.

Como todos os atributos são numéricos efetuou-se uma análise de correlação (coeficiente Pearson) para cada atributo em relação ao atributo carga de laminação (atributo classe), com os resultados apresentados na Tabela 15.

Tabela 15

Análise Correlação base de dados Laminação a Frio

Atributo	Coeficiente Pearson
A6	0,78
A1	0,71
A3	0,38
A2	-0,25
A5	-0,096
A4	-0,008

O algoritmo Apriori foi executado para cada situação prevista na tabela de geração artificial de ausência. Os parâmetros de suporte mínimo e confiança mínima foram de 1%. Valores muito baixos de suporte e confiança a permitem produção de muitas regras de associação. Uma vez que o RAIMA analisa todas as regras com consequente igual a ausente, a execução do Apriori considerou apenas o atributo situação da classe (ausente e presente) e cada atributo individualmente. Os gráficos a seguir representam os respectivos valores de suporte e confiança de cada regra produzida para cada experimento.

Para o mecanismo MCAR nota-se que as regras que implicam em ausência estão distribuídas uniformemente entre todos os atributos. Além disso, o percentual de ausência é proporcional ao valor da confiança. O valor de suporte também é constante, salvo para os atributos a6 e a7. Conforme já averiguado pela análise de correlação, o atributo a6 é fortemente correlacionado ao atributo a7, indicando que o mesmo poderia ser retirado da base de dados para diminuir a dimensionalidade e consequentemente diminuir também a quantidade de regras geradas pela algoritmo Apriori.

.

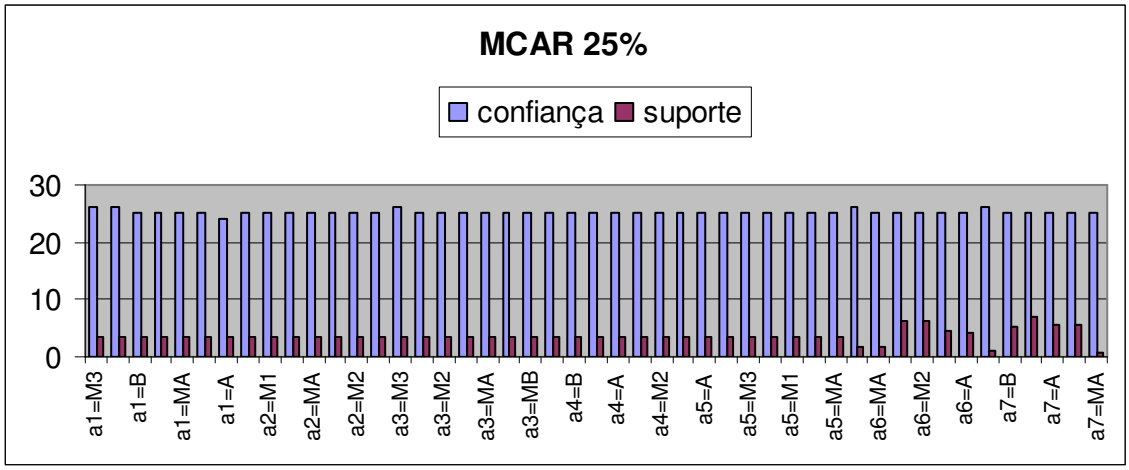


Gráfico 17 – Laminação a Frio – MCAR 25% ausência

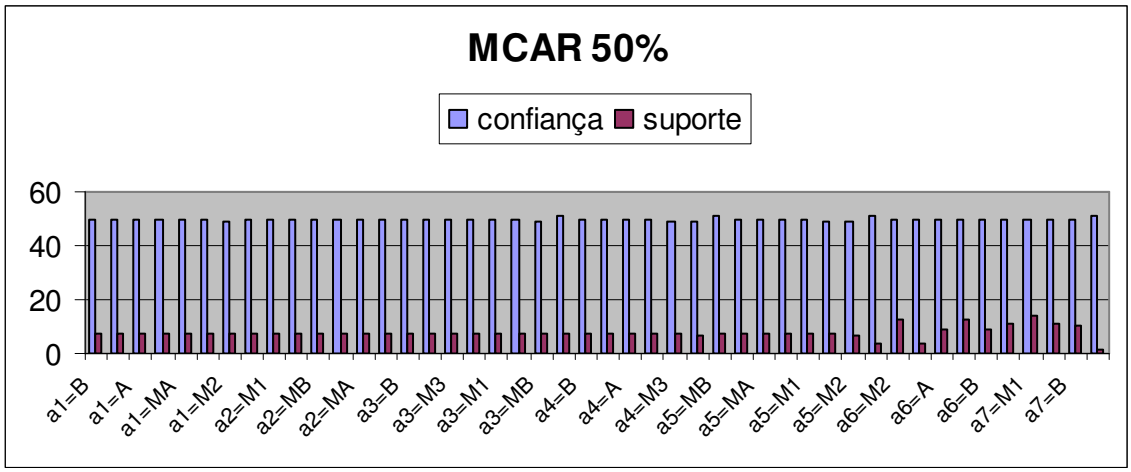


Gráfico 18 – Laminação a Frio – MCAR 50% ausência

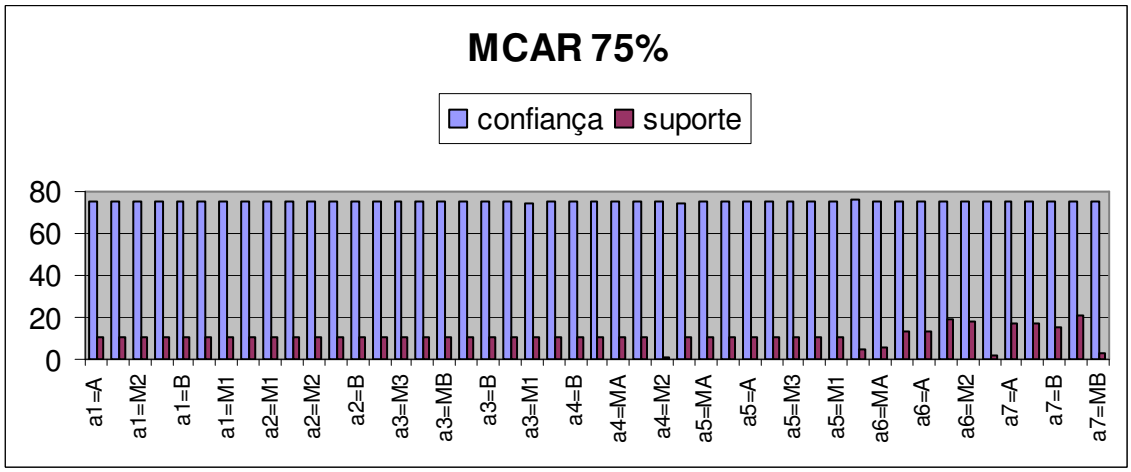


Gráfico 19 – Laminação a Frio – MCAR 75% ausência

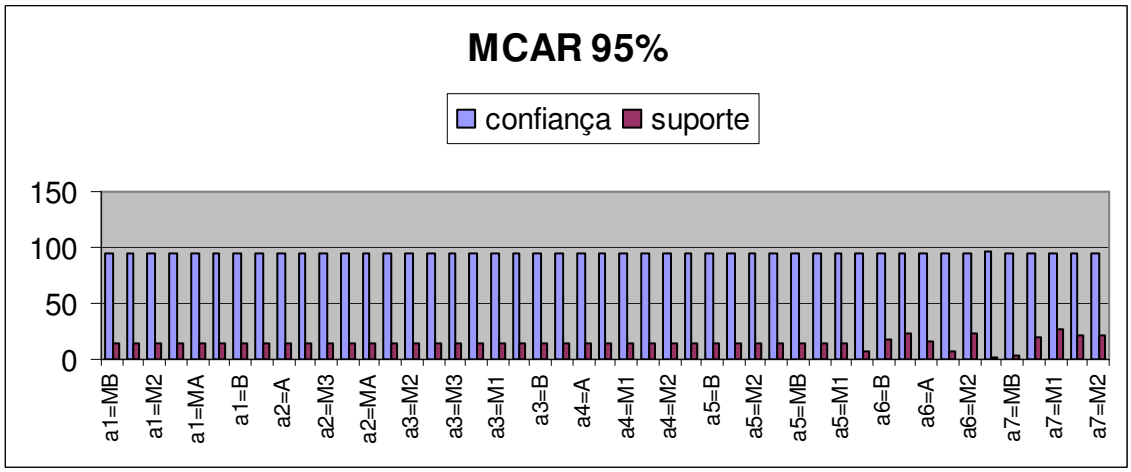


Gráfico 20 – Laminação a Frio – MCAR 95% ausência

Na geração da ausência do mecanismo MAR utilizou-se o critério a3 > 0,12. Essa condição é facilmente identificada nos gráficos abaixo com percentual de confiança proporcional ao percentual de ausência utilizado. Entretanto, destacam-se também algumas regras relacionadas ao atributo a7 (classe) à medida que o percentual de ausência atinge 50%. Esta observação confirma que o mecanismo MAR e NMAR confundem-se em determinadas situações. Além disso, como sugerido por GRAHAM (2007), no mundo real as ausências são parcialmente MAR e NMAR. Essa hipótese poderá ser investigada em trabalhos futuros e até mesmo chegar à definição de um quarto mecanismo de ausência, o caso híbrido (MAR e NMAR).

As regras apontadas nos gráficos com alto valor de confiança são analisadas com o objetivo de obter-se o melhor conjunto de regras que represente a ausência. Para esta análise é utilizado o índice *Jaccard*, que mede o grau de sobreposição entre os casos cobertos por cada parte da regra. Altos valores indicam que o antecedente e o consequente da regra tendem a cobrir os mesmos casos. Esse índice varia entre o intervalo 0 e 1. A Tabela 16 apresenta os resultados do índice *Jaccard* para cada experimento.

Tabela 16

Medidas – Laminação a Frio

Experimento	Regra	Confiança	Suporte	Índice *Jaccard*
MAR 50%	a3=MA ⇒ Ausente	51	7	0,253
	a3=M3 ⇒ Ausente	50	7	0,250
	a3=A ⇒ Ausente	49	7	0,246
	a7=MA ⇒ Ausente	49	1	0,048
MAR 75%	a3=MA ⇒ Ausente	75	11	0,299
	a3=M3 ⇒ Ausente	75	11	0,300
	a3=A ⇒ Ausente	75	11	0,300
	a7=MA ⇒ Ausente	73	2	0,050
MAR 100%	a3=MA ⇒ Ausente	100	14	0,333
	a3=M3 ⇒ Ausente	100	14	0,333
	a3=A ⇒ Ausente	100	14	0,333
	a7=MA ⇒ Ausente	97	2	0,050
NMAR 25%	a6=MA ⇒ Ausente	25	2	0,106
	a7=MA ⇒ Ausente	25	1	0,042
	a7=M2 ⇒ Ausente	25	6	0,196
	a7=A ⇒ Ausente	25	6	0,195
	a1=MA ⇒ Ausente	23	3	0,145
	a6=A ⇒ Ausente	22	4	0,151
	a1=A ⇒ Ausente	20	3	0,124
NMAR 50%	a6=MA ⇒ Ausente	50	4	0,136
	a7=MA ⇒ Ausente	50	1	0,045
	a7=M2 ⇒ Ausente	50	11	0,322
	a7=A ⇒ Ausente	50	11	0,322
	a1=MA ⇒ Ausente	45	6	0,205
	a6=A ⇒ Ausente	44	8	0,227
	a1=A ⇒ Ausente	40	6	0,175
NMAR 75%	a6=MA ⇒ Ausente	75	4	0,095
	a7=MA ⇒ Ausente	75	2	0,046
	a7=A ⇒ Ausente	75	17	0,413
	a7=M2 ⇒ Ausente	75	17	0,407
	a1=MA ⇒ Ausente	67	10	0,238
	a6=A ⇒ Ausente	67	6	0,120
	a1=A ⇒ Ausente	61	9	0,251
NMAR 100%	a6=MA ⇒ Ausente	100	8	0,159
	a7=A ⇒ Ausente	100	23	0,478
	a7=M2 ⇒ Ausente	100	23	0,474
	a7=MA ⇒ Ausente	100	2	0,047
	a1=MA ⇒ Ausente	90	13	0,263
	a6=A ⇒ Ausente	89	15	0,308
	a1=A ⇒ Ausente	81	12	0,229

Nota-se que as regras que apresentam o maior índice *Jaccard* para cada experimento são as regras que identificam o atributo causador da ausência, atributo a3 no caso MAR e atributo a7 no caso NMAR.

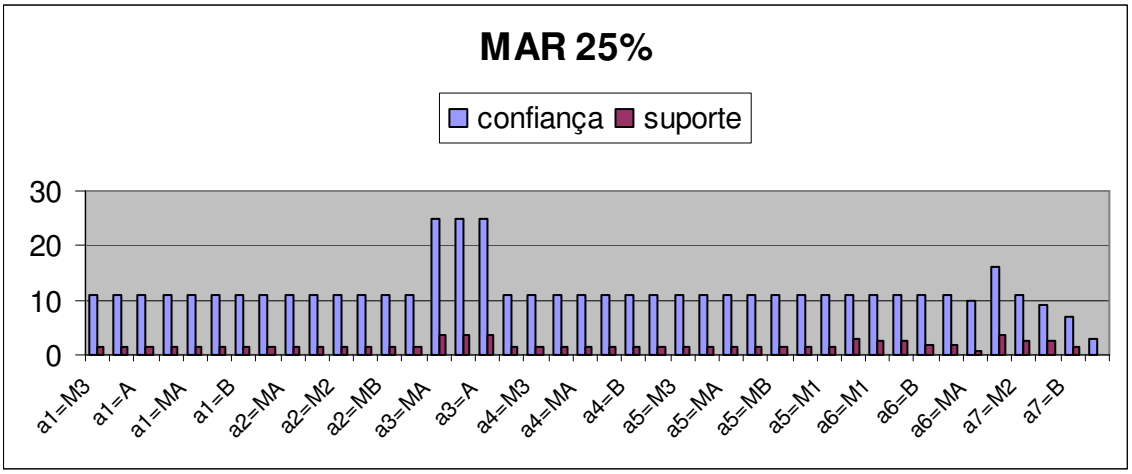


Gráfico 21 – Laminação a Frio – MAR 25% ausência

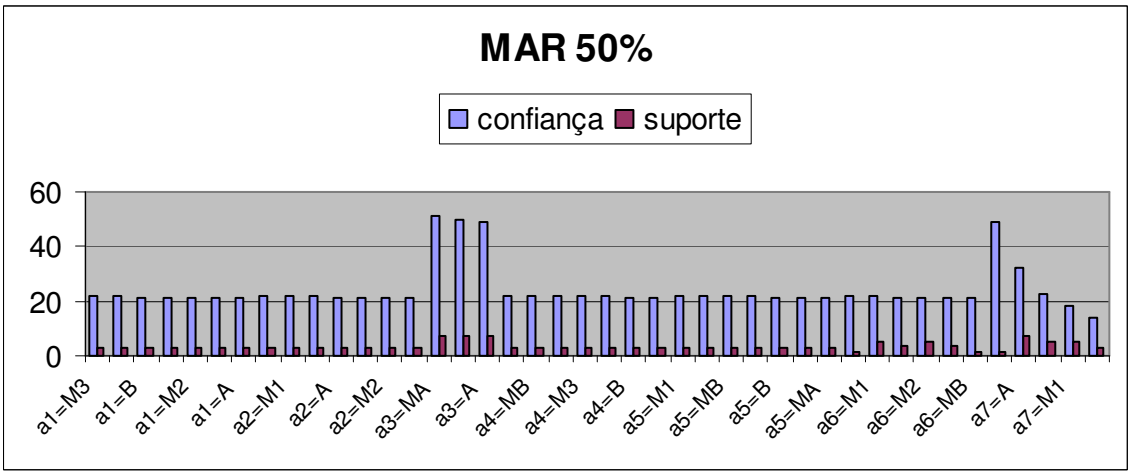


Gráfico 22 – Laminação a Frio – MAR 50% ausência

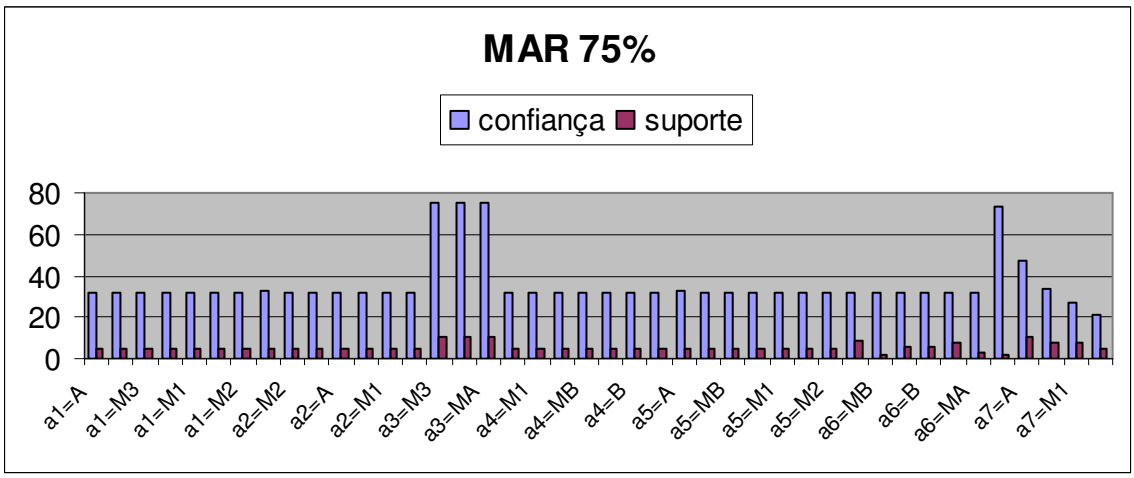


Gráfico 23 – Laminação a Frio – MAR 75% ausência

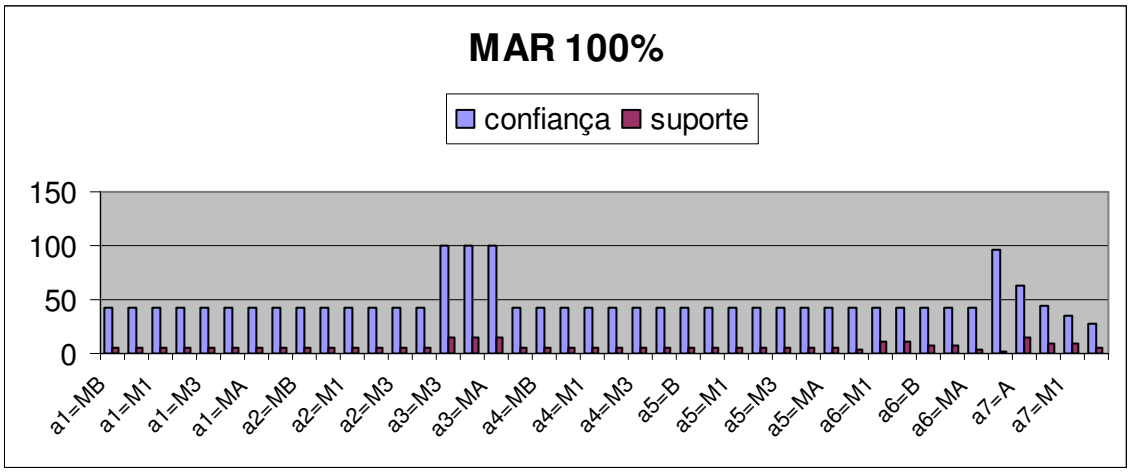


Gráfico 24 – Laminação a Frio – MAR 100% ausência

Na geração da ausência do mecanismo NMAR utilizou-se o critério a7 > 1755, que representa as categorias M2, A e MA. As regras relacionadas ao atributo a7 são identificadas nos gráficos como responsáveis pela ausência dos registros. As regras relacionadas ao atributo a6 poderiam também sinalizar a causa da ausência, porém essa suspeita pode ser eliminada, uma vez que o atributo a6 é altamente correlacionado ao atributo a7, podendo até mesmo ser descartado do domínio do problema.

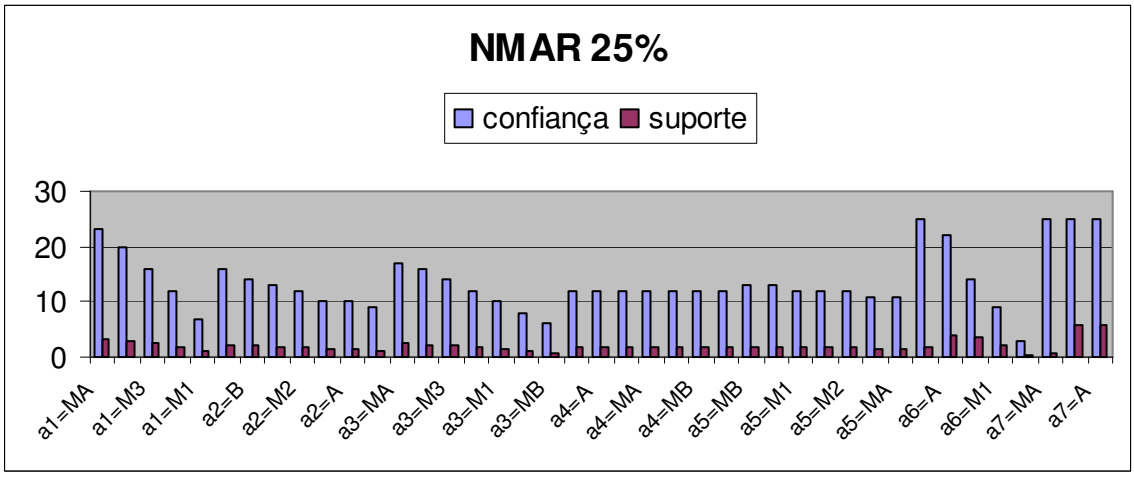


Gráfico 25 – Laminação a Frio –NMAR 25% ausência

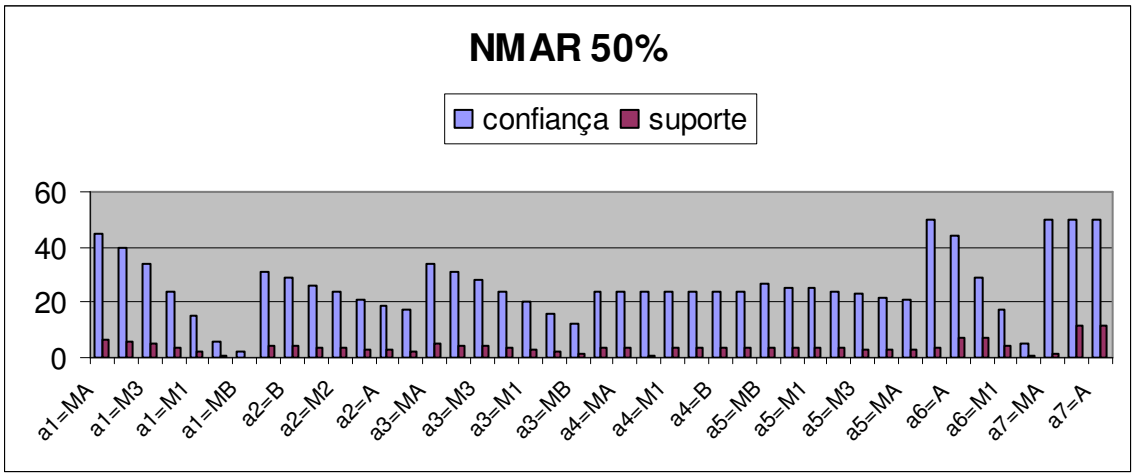


Gráfico 26 – Laminação a Frio –NMAR 50% ausência

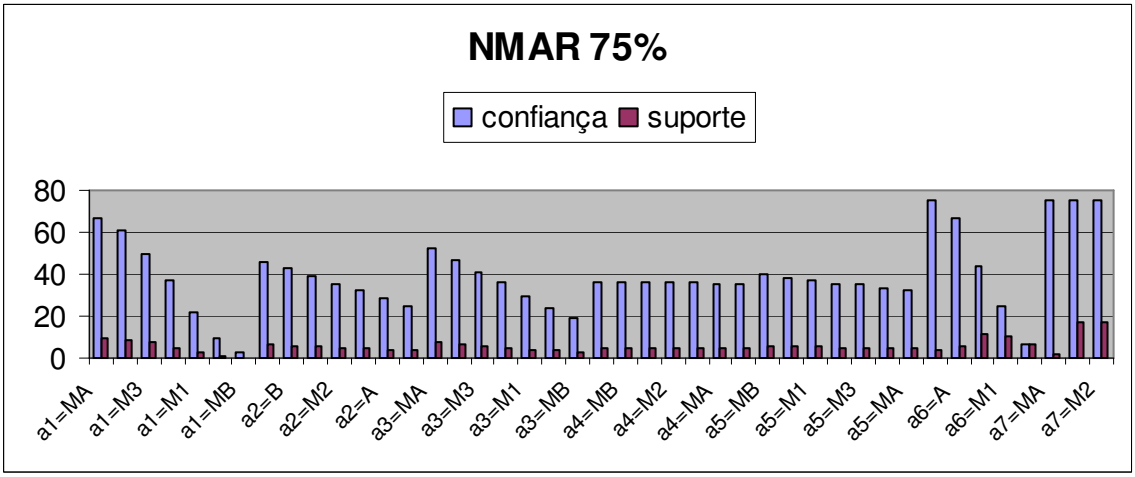


Gráfico 27 – Laminação a Frio –NMAR 75% ausência

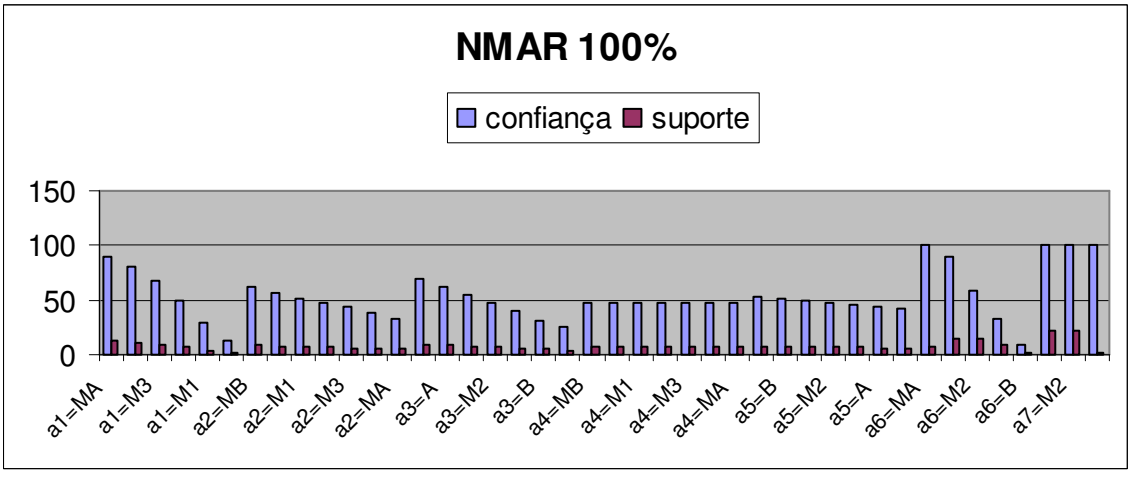


Gráfico 28 – Laminação a Frio –NMAR 100% ausência

No caso da base de dados Laminação a Frio, pode-se concluir que o método RAIMA conseguiu identificar o mecanismo responsável pela ausência em todos os experimentos pela satisfação das três hipóteses propostas pelo método.

4.4.2 Base de dados Mushroom

A análise exploratória da base de dados *Mushroom* através de histogramas é apresentada na Figura 2. Através da observação dos histogramas, nota-se que o atributo a17 possui um único valor na base de dados, portanto pode ser desconsiderado, uma vez que para qualquer tipo de mecanismo e percentual de ausência implicará em regras de alta confiança e alto suporte. Além disso, alguns atributos como os atributos a7, a8, a9, a18 e a19 possuem uma distribuição de valores desbalanceada em relação às classes, o que implicará em alto valor de suporte para os valores mais frequentes..

A base de dados *Mushroom* possui apenas atributos categóricos, por isso não foi necessário utilizar as técnicas de discretização.

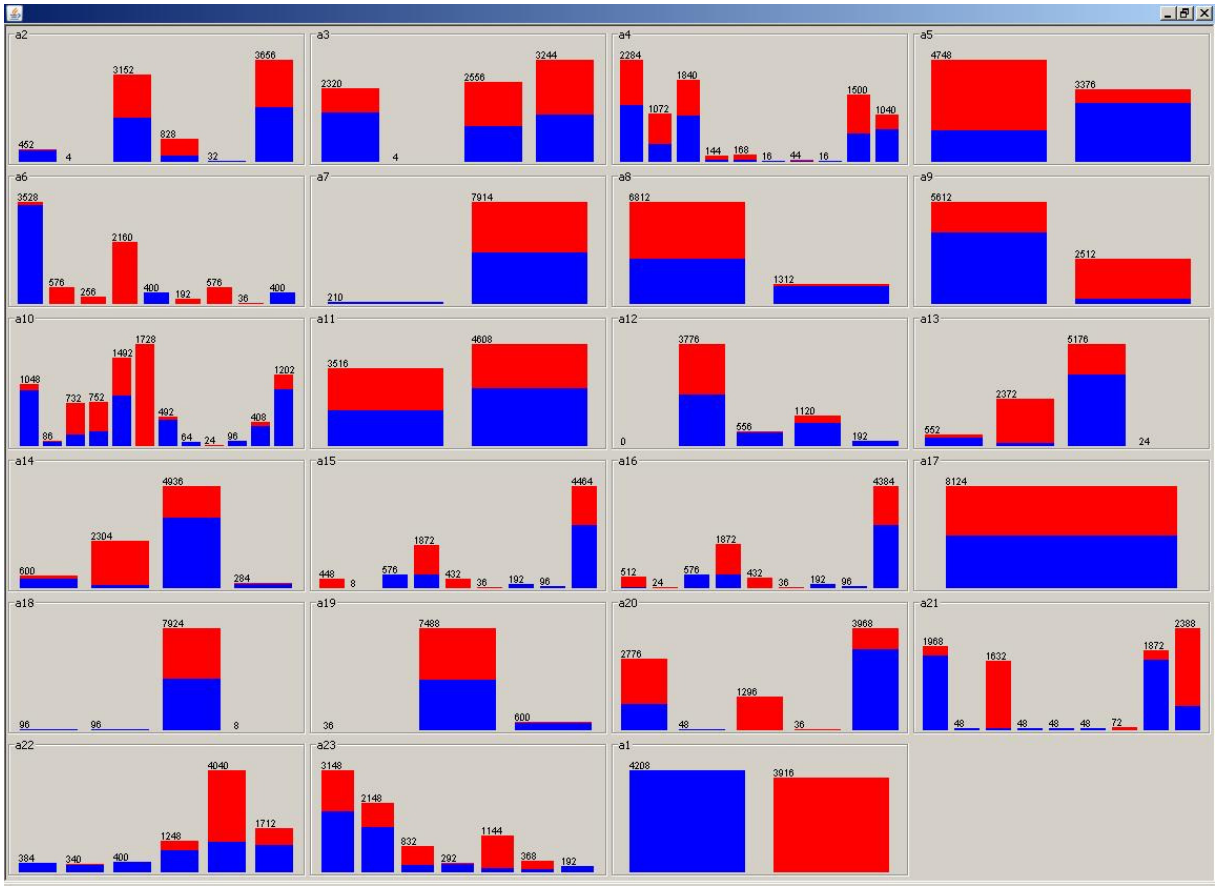


Figura 2 – Histogramas da base de dados *Mushroom*

A correlação de atributos foi verificada utilizando índice Cramer, Tabela de Contingência, coeficiente Phi e Qui-quadrado. Além disso, foi efetuada também análise de ganho de informação de cada atributo. Os resultados são apresentados na Tabela 17.

Tabela 17

Resultado dos índices Cramer, Contingência, Phi, Qui-quadrado e ganho de informação para a base de dados *Mushroom*

atributo	Cramer	Contingência	Phi	Qui-quadrado	Ganho de informação
a2	0,246	0,238	0,246	489,92	0,0488
a3	0,197	0,193	0,197	315,04	0,02859
a4	0,218	0,213	0,218	387,59	0,03605
a5	0,506	0,448	0,502	2043,45	0,19238
a6	**0,971**	**0,697**	**0,971**	**7659,72**	**0,90607**
a7	0,129	0,128	0,129	135,61	0,01417
a8	0,348	0,329	0,348	986,03	0,10088
a9	0,54	0,475	0,54	2369,17	0,23015
a10	0,681	0,563	0,681	3765,71	0,41698
a11	0,102	0,101	0,102	84,55	0,00752
a12	0,407	0,377	0,407	419,26	0,03834
a13	0,588	0,507	0,588	2808,28	0,28473

a14	0,575	0,498	0,575	2684,47	0,27189
a15	0,525	0,465	0,525	2237,89	0,25385
a16	0,515	0,458	0,515	2152,39	0,24142
a17	-	-	-	0	0
a18	0,153	0,152	0,153	191,22	0,02382
a19	0,215	0,21	0,215	374,73	0,03845
a20	0,603	0,517	0,603	2956,61	0,3108
a21	0,753	0,601	0,753	4602,03	0,4807
a22	0,487	0,438	0,487	1929,74	0,20196
a23	0,44	0,403	0,44	1573,77	0,15683

Os atributos com os maiores índices são respectivamente a6 e a21. Atributos altamente correlacionados podem confundir a identificação dos mecanismos MAR e NMAR, situação que será evidenciada nos respectivos gráficos dos mecanismos.

Para o mecanismo MCAR evidencia-se uma distribuição uniforme dos valores de confiança em todas as observações de ausência. Além disso, observa-se que o valor da confiança obedece ao percentual de ausência aplicado à base de dados. Outra observação, não evidenciada na base de dados de Laminação, é uma variação nas medidas de suporte. Ao investigar a razão dessa variação, observa-se "picos" nas medidas de suporte quando a distribuição dos valores para o atributo é muito desbalanceada em relação às categorias. Esta observação é exemplificada com a regra a7=f $\Rightarrow$ ausente, Conf 50% e Sup 49% gerada para o mecanismo MCAR com 50% de ausência. O atributo a7 possui apenas 2 valores que são a e f, sendo que a7=f possui 7914 instâncias e a7=a apenas 210 instâncias. Outro exemplo é a regra a17=p, o atributo a17 possui somente o valor p, o que faz com que essa regra apareça com alto valor de suporte.

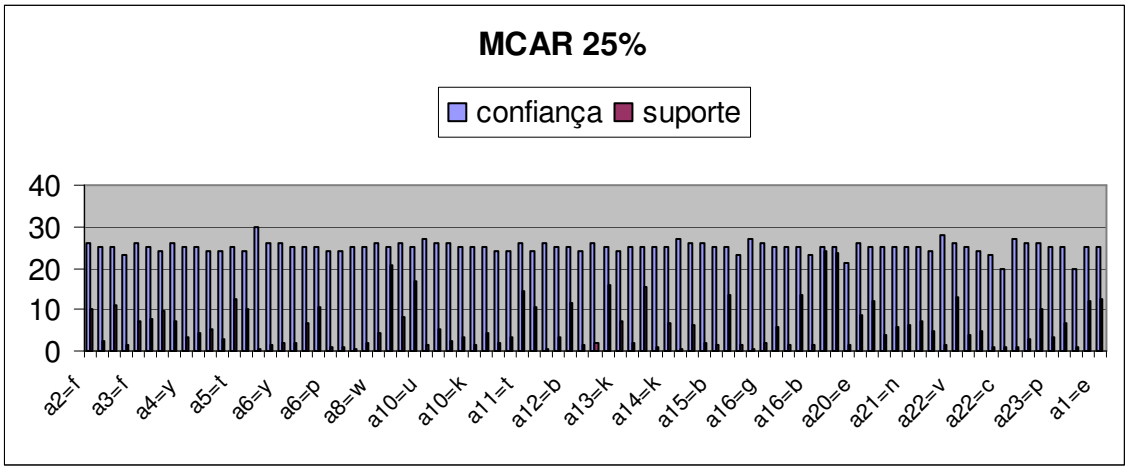


Gráfico 29 – *Mushroom* – MCAR 25%

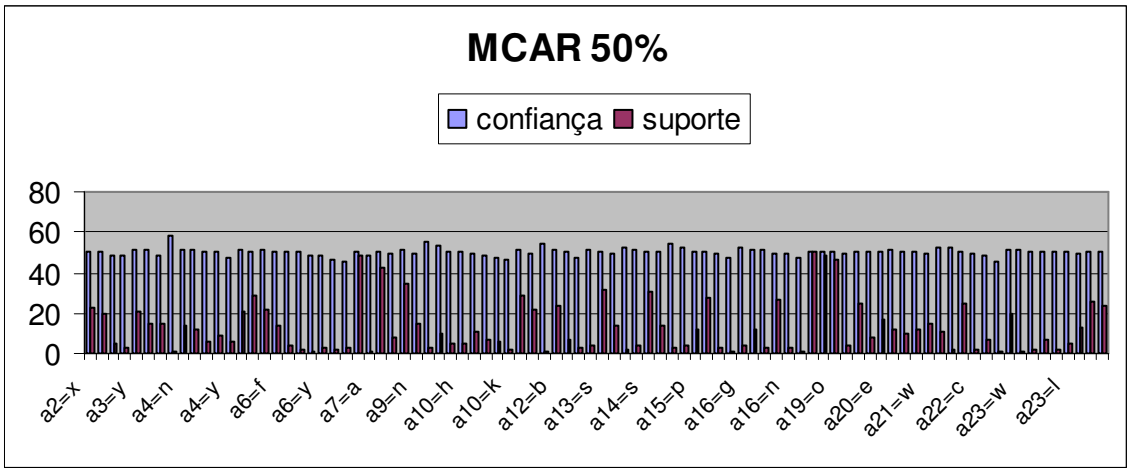


Gráfico 30 – *Mushroom* – MCAR 50%

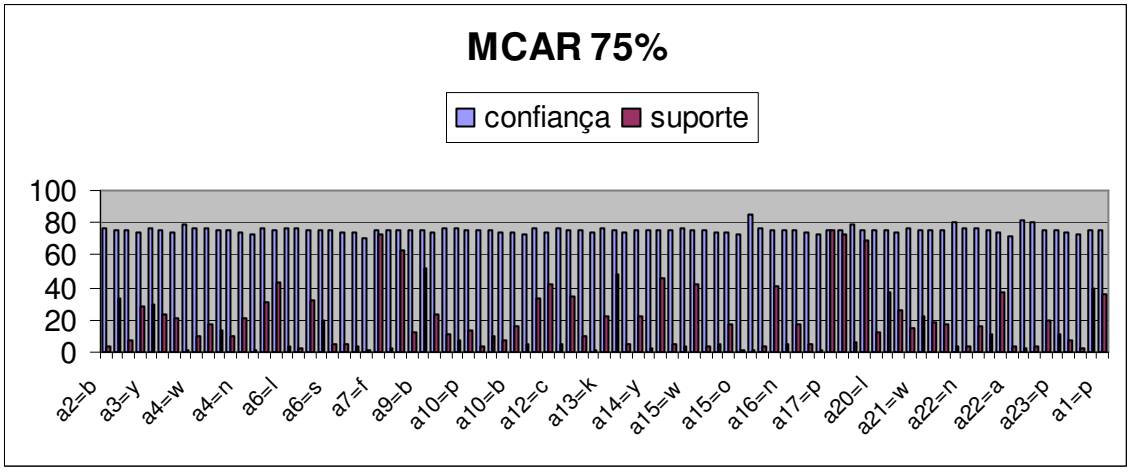


Gráfico 31 – *Mushroom* – MCAR 75%

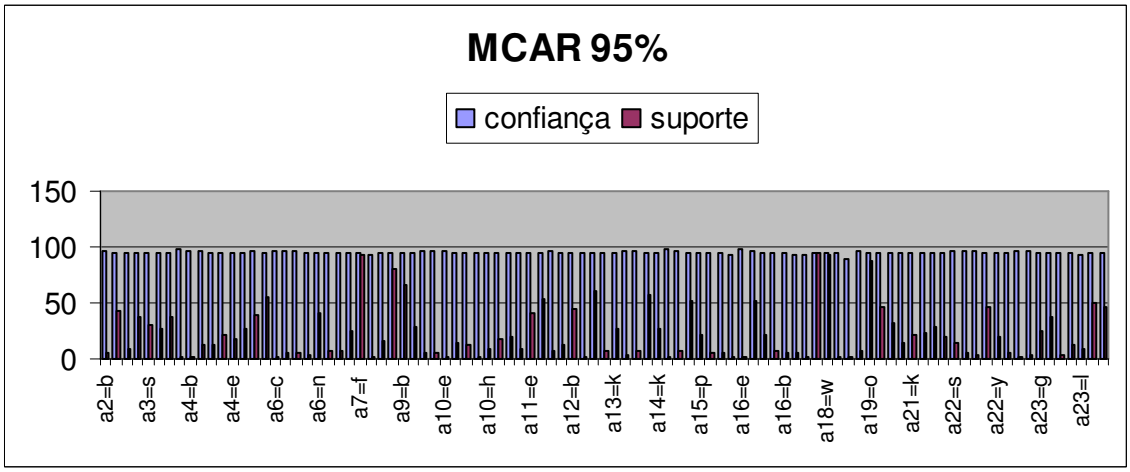


Gráfico 32 – *Mushroom* – MCAR 95%

A geração da ausência para o mecanismo MAR utilizou o critério a22 = v, que pode ser observado nos gráficos abaixo com alto valor de confiança e suporte. Os atributos

altamente correlacionados (a6 e a21), também apresentam altos valores de confiança, porém suas medidas de suporte são inferiores à medida de suporte da regra a22. Esses dois atributos podem ser desconsiderados na análise das regras geradas. A Tabela 18 apresenta o índice *Jaccard* para as regras que apresentaram alto valor de confiança.

Tabela 18

Medidas – *Mushroom*

Experimento	Regra	Confiança	Suporte	Índice *Jaccard*
MAR 25%	a6=s ⇒ Ausente	26	2	0,103
	a10=b ⇒ Ausente	25	5	0,189
	a22=v ⇒ Ausente	25	12	0,250
	a6=y ⇒ Ausente	24	2	0,096
	a9=n ⇒ Ausente	22	7	0,187
	a23=l ⇒ Ausente	22	2	0,112
MAR 50%	a6=s ⇒ Ausente	50	4	0,087
	a10=b ⇒ Ausente	50	11	0,220
	a22=v ⇒ Ausente	50	25	0,400
	a6=y ⇒ Ausente	48	3	0,082
	a9=n ⇒ Ausente	43	13	0,244
	a23=l ⇒ Ausente	41	4	0,096
MAR 75%	a6=s ⇒ Ausente	75	5	0,135
	a10=b ⇒ Ausente	75	16	0,374
	a22=v ⇒ Ausente	75	37	0,750
	a6=y ⇒ Ausente	74	5	0,134
	a23=l ⇒ Ausente	66	7	0,164
MAR 100%	a6=y ⇒ Ausente	100	7	0,142
	a6=s ⇒ Ausente	100	7	0,142
	a10=b ⇒ Ausente	100	21	0,142
	a22=v ⇒ Ausente	100	50	1
	a9=n ⇒ Ausente	88	27	0,500
	a23=l ⇒ Ausente	87	9	0,173
NMAR 25%	a6=l ⇒ Ausente	26	1	0,077
	a15=g ⇒ Ausente	26	2	0,101
	a2=b ⇒ Ausente	25	1	0,080
	a22=a ⇒ Ausente	25	1	0,070
	a1=e ⇒ Ausente	25	13	0,250
	a6=n ⇒ Ausente	24	11	0,229
	a16=g ⇒ Ausente	24	12	0,091
	a22=n ⇒ Ausente	24	1	0,069
	a6=a ⇒ Ausente	23	1	0,067
	a10=n ⇒ Ausente	23	3	0,131
	a10=w ⇒ Ausente	23	1	0,077
	a12=c ⇒ Ausente	23	2	0,087
	a21=n ⇒ Ausente	23	6	0,180

	a8=w ⇒ Ausente	22	4	0,137
	a19=t ⇒ Ausente	22	2	0,085
	a5=t ⇒ Ausente	21	9	0,185
	a21=k ⇒ Ausente	21	5	0,154
NMAR 50%	a16=g ⇒ Ausente	53	4	0,127
	a6=l ⇒ Ausente	51	3	0,088
	a23=w ⇒ Ausente	51	1	0,044
	a15=g ⇒ Ausente	50	4	0,120
	a1=e ⇒ Ausente	50	26	0,500
	a6=n ⇒ Ausente	49	21	0,438
	a10=u ⇒ Ausente	49	3	0,102
	a12=r ⇒ Ausente	48	1	0,041
	a15=o ⇒ Ausente	48	1	0,041
	s16=o ⇒ Ausente	48	1	0,041
	a22=a ⇒ Ausente	48	2	0,080
	a22=n ⇒ Ausente	47	2	0,080
	a6=a ⇒ Ausente	46	2	0,078
	a23=m ⇒ Ausente	46	2	0,058
	a2=b ⇒ Ausente	45	2	0,086
	a8=w ⇒ Ausente	45	7	0,209
	a10=n ⇒ Ausente	45	6	0,175
	a12=c ⇒ Ausente	45	3	0,104
	a21=k ⇒ Ausente	45	10	0,267
	a21=n ⇒ Ausente	45	11	0,274
	a7=a ⇒ Ausente	44	1	0,041
	a10=k ⇒ Ausente	43	2	0,074
	a19=t ⇒ Ausente	43	3	0,104
	a22=c ⇒ Ausente	43	2	0,063
NMAR 75%	a15=g ⇒ Ausente	76	5	0,100
	a1=e ⇒ Ausente	75	39	0,600
	a6=n ⇒ Ausente	73	32	0,494
	a16=g ⇒ Ausente	73	5	0,096
	a8=w ⇒ Ausente	68	5	0,192
	a10=n ⇒ Ausente	67	9	0,155
	a21=n ⇒ Ausente	67	16	0,271
NMAR 100%	a6=a ⇒ Ausente	100	5	0,095
	a6=l ⇒ Ausente	100	5	0,095
	a10=e ⇒ Ausente	100	1	0,027
	a12=r ⇒ Ausente	100	2	0,045
	a15=g ⇒ Ausente	100	7	0,136
	a15=o ⇒ Ausente	100	2	0,045
	a15=e ⇒ Ausente	100	1	0,022
	a16=g ⇒ Ausente	100	7	0,136
	a16=o ⇒ Ausente	100	2	0,045
	a16=e ⇒ Ausente	100	1	0,022
	a18=n ⇒ Ausente	100	1	0,022

	a18=o ⇒ Ausente	100	1	0,022
	a22=n ⇒ Ausente	100	5	0,095
	a22=a ⇒ Ausente	100	5	0,091
	a23=w ⇒ Ausente	100	2	0,045
	a1=e ⇒ Ausente	100	52	1
	a6=n ⇒ Ausente	97	42	0,787
	a12=c ⇒ Ausente	92	6	0,120
	a7=a ⇒ Ausente	91	2	0,045
	a8=w ⇒ Ausente	91	15	0,277
	a10=u ⇒ Ausente	90	5	0,104

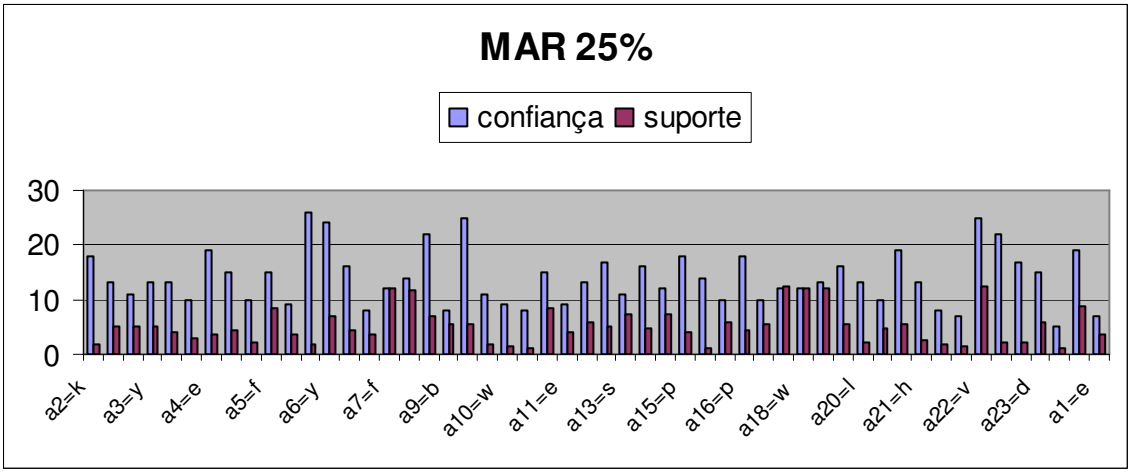


Gráfico 33 – *Mushroom* – MAR 25%

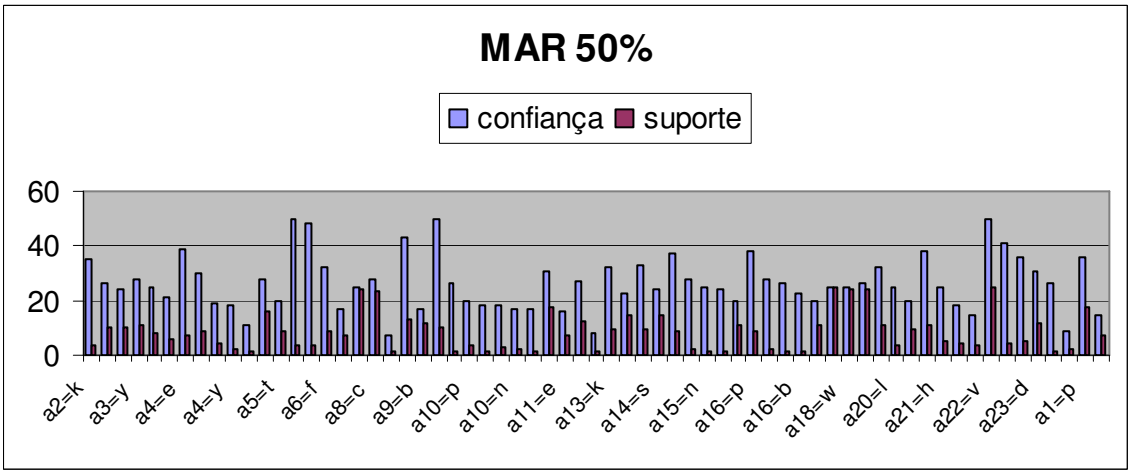


Gráfico 34 – *Mushroom* – MAR 50%

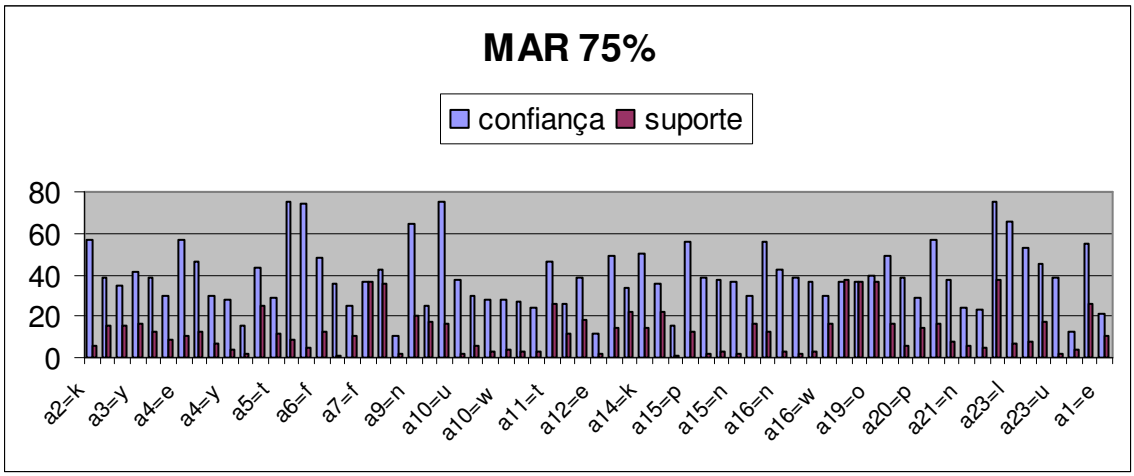


Gráfico 35 – *Mushroom* – MAR 75%

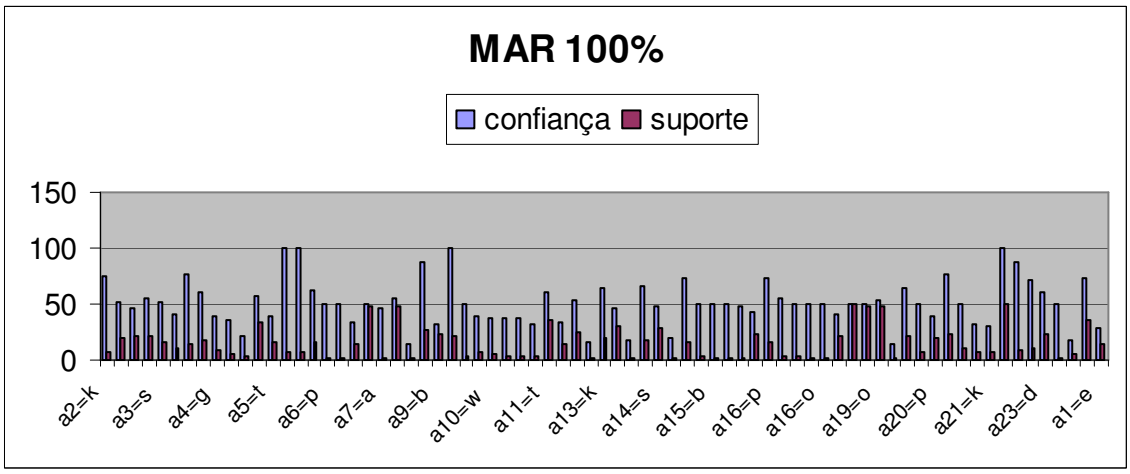


Gráfico 36 – *Mushroom* – MAR 100%

O mecanismo NMAR foi gerado com o critério a1=e, que representa a classe e (*edible*, cogumelos comestíveis). O mecanismo NMAR confunde-se com o mecanismo MAR: várias regras com alto valor de confiança são geradas, porém nota-se que dentre essas regras o suporte é maior para as regras referentes ao atributo classe. Como já citado anteriormente, os atributos altamente correlacionados ao atributo classe podem ser desconsiderados na análise, eliminando assim algumas regras que confundem a identificação do mecanismo. Outra observação, também ocorrida nos experimentos MCAR e MAR, é a variação da medida de suporte.

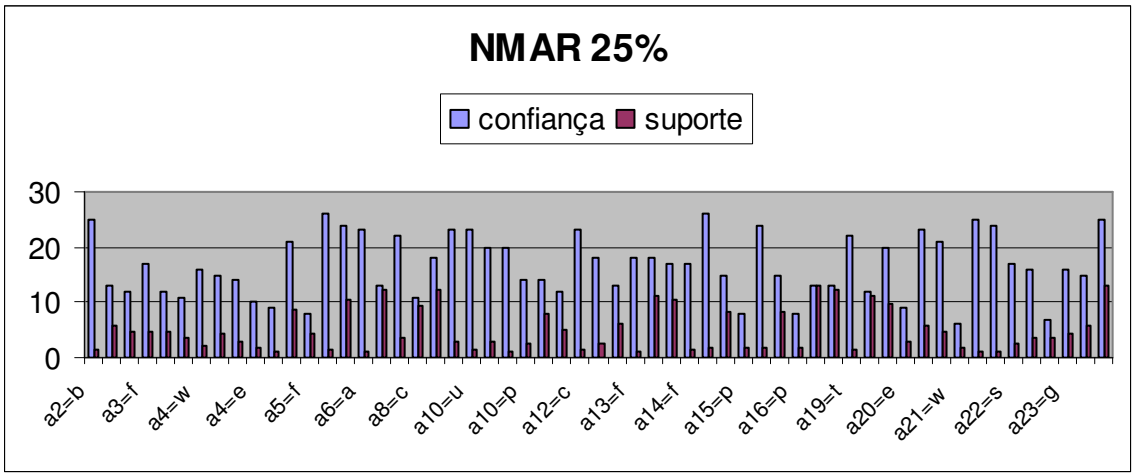


Gráfico 37 – *Mushroom* – NMAR 25%

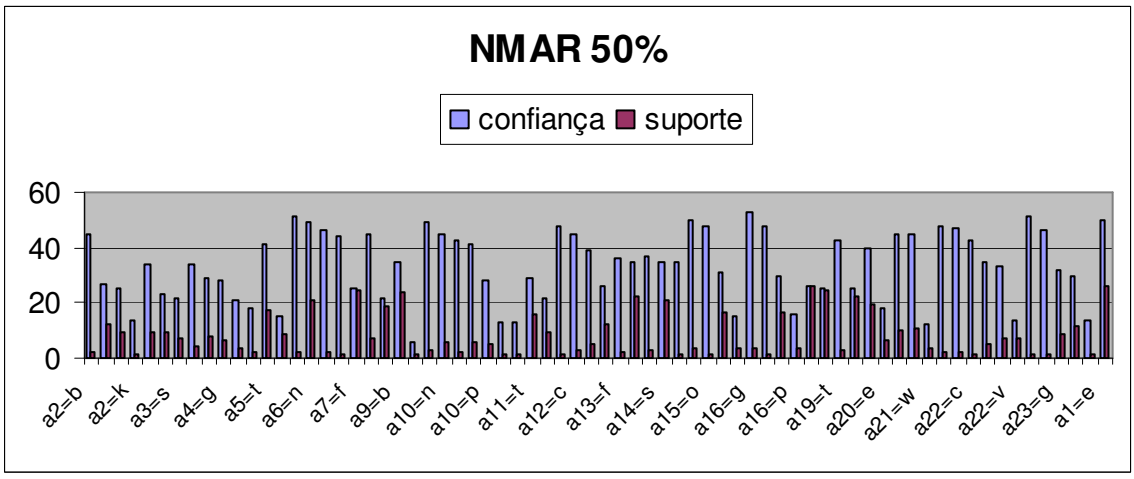


Gráfico 38 – *Mushroom* – NMAR 50%

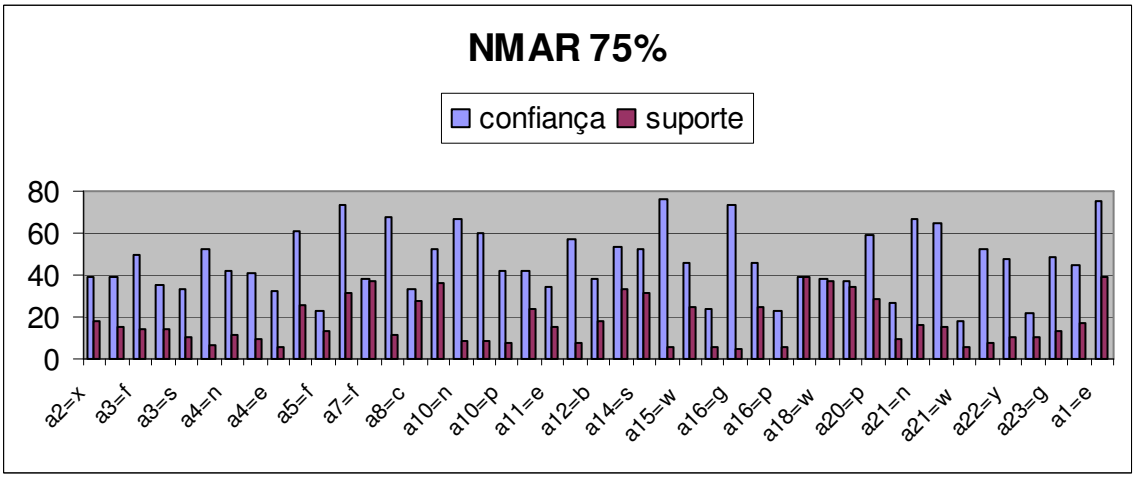


Gráfico 39 – *Mushroom* – NMAR 75%

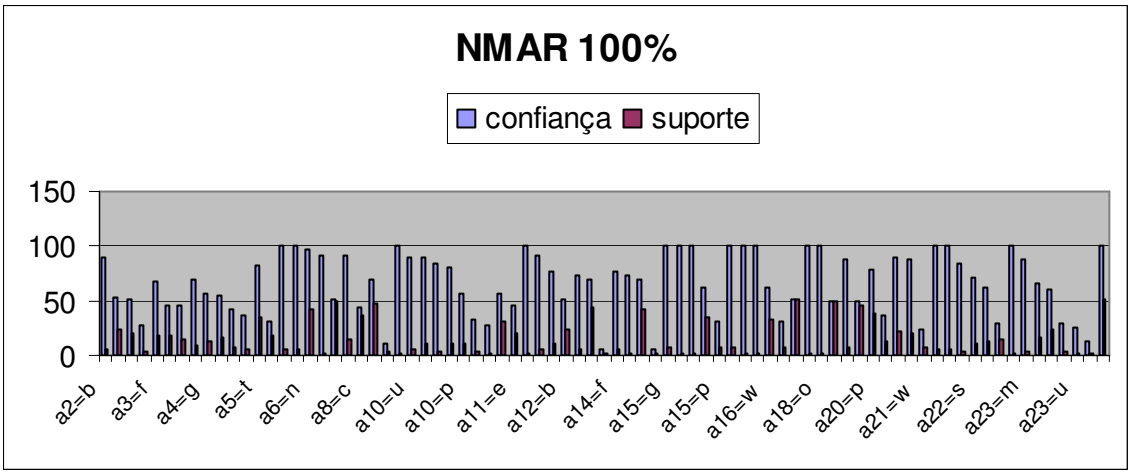


Gráfico 40 – *Mushroom* – NMAR 100%

A aplicação do método RAIMA na base de dados *Mushroom* obteve com sucesso a identificação do mecanismo de ausência, pois as três hipóteses dos métodos foram satisfeitas. Nesta base de dados foi encontrada uma grande variação da medida de suporte devido às características dos valores de determinados atributos. O atributo a17 possui um único valor e os atributos a7, a8, a9, a18 e a19 possuem valores desbalanceados entre as classes.

4.4.3 Base de dados Wisconsin

A análise exploratória da base dados *Wisconsin* é sumarizada nos histogramas da Figura 3.

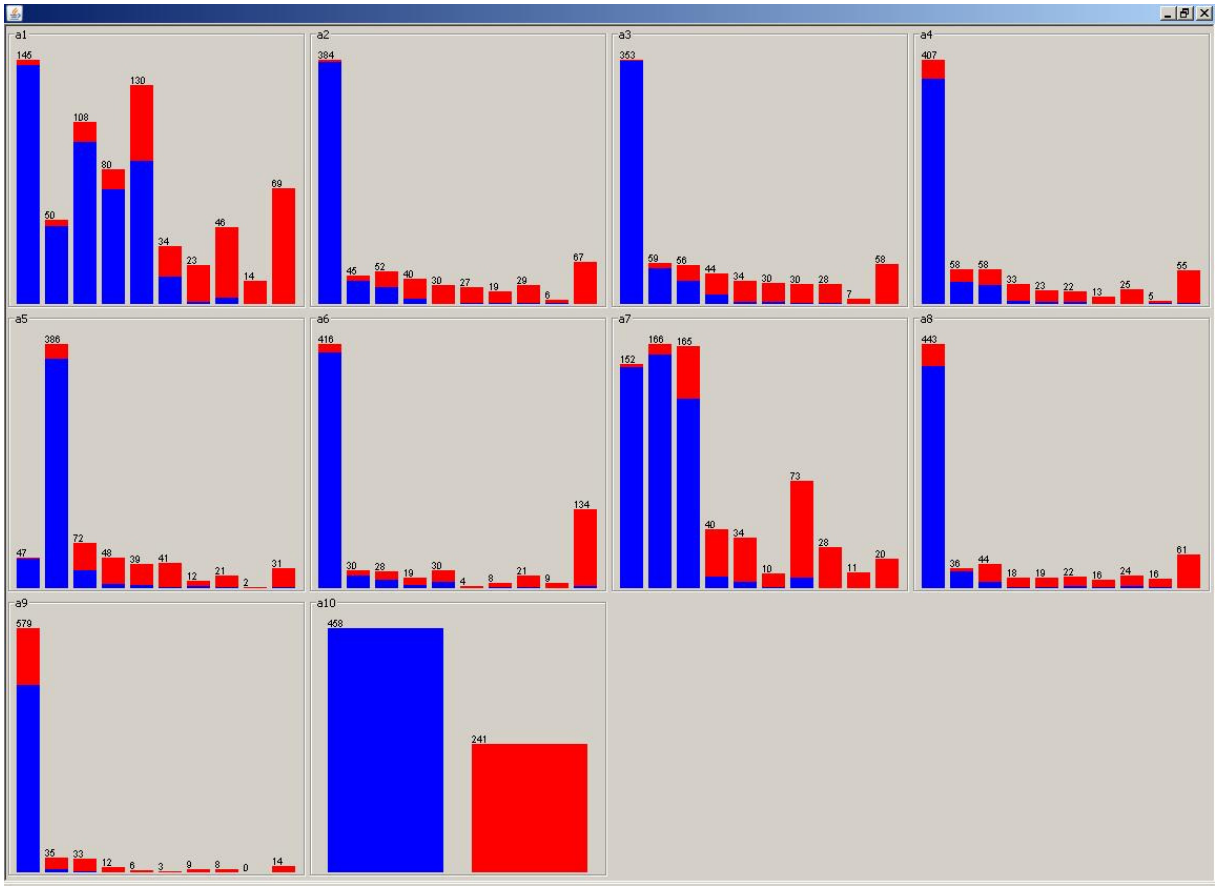


Figura 3 – Histogramas dos atributos da base de dados *Wisconcin*

Apesar dos atributos da base de dados serem numéricos, os mesmos foram considerados como categóricos. Para tanto foi efetuada a análise de correlação (ver Tabela 19) e também a análise dos índices Cramer, Tabela de Contingência, Coeficiente Phi, Qui-Quadrado e Ganho de Informação de cada atributo (ver Tabela 20).

Tabela 19

Coeficiente Pearson para os atributos da base de dados *Wisconsin*

Atributo	Coeficiente Pearson
a1	0,716
a2	0,817
a3	**0,818**
a4	0,696
a5	0682
a6	0,817
a7	0,756
a8	0,712
a9	0,423

Tabela 20

Resultado dos índices Cramer, Contingência, Phi, Qui-quadrado e ganho de informação para a base de dados *Wisconsin* discretizado

atributo	Cramer	Contingência	Phi	Qui-quadrado	Ganho de informação
a1	0,74	0,59	0,74	389,19	0,465
a2	**0,88**	**0,66**	**0,88**	**543,67**	**0,684**
a3	0,86	0,66	0,86	525,75	0,661
a4	0,74	0,59	0,74	389,52	0,449
a5	0,79	0,62	0,79	445,81	0,514
a6	0,84	0,64	0,84	495,91	0,594
a7	0,81	0,63	0,81	460,04	0,548
a8	0,75	0,61	0,77	420,29	0,475
a9	0,53	0,46	0,53	195,98	0,21

Todos os atributos da base de dados *Wisconsin* são altamente correlacionados, exceto o atribuo a9. Esta característica pode representar uma dificuldade na identificação do mecanismo, uma vez que muitas regras com alta confiança serão produzidas.

Na tentativa de evitar esse problema, foi aplicada a técnica de discretização à base de dados. A técnica *equi-width* não é recomendada para esta base de dados, pois os intervalos gerados serão similares aos valores existentes. Por outro lado, a técnica *equi-depth* determina um tamanho fixo de pontos para cada intervalo. Sendo assim, foi empregado o particionamento baseado na distância, no qual as tuplas de cada intervalo são distribuídas de forma uniforme. A Tabela 21 apresenta as novas categorias após a discretização dos atributos.

Tabela 21

Categorias dos atributos da base de dados *Wisconsin* após discretização

Atributo	Nova categoria	Domínio
a1	1	1,2
	2	3,4
	3	5,6
	4	7,8,9,10
a2	1	1
	2	2,3,4,5,6,7,8,9,10
a3	1	1
	2	2,3,4,5,6,7,8,9,10
a4	1	1
	2	2,3,4,5,6,7,8,9,10
a5	1	1,2
	2	3,4,5,6,7,8,9,10
a6	1	1,2
	2	3,4,5,6,7,8,9,10
a7	1	1
	2	2
	3	3
	4	4,5,6,7,8,9,10
a8	1	1
	2	2,3,4,5,6,7,8,9,10
a9	1	a
	2	2,3,4,5,6,7,8,9,10

A Figura 4 apresenta a base de dados *Wisconsin* após a discretização dos atributos.

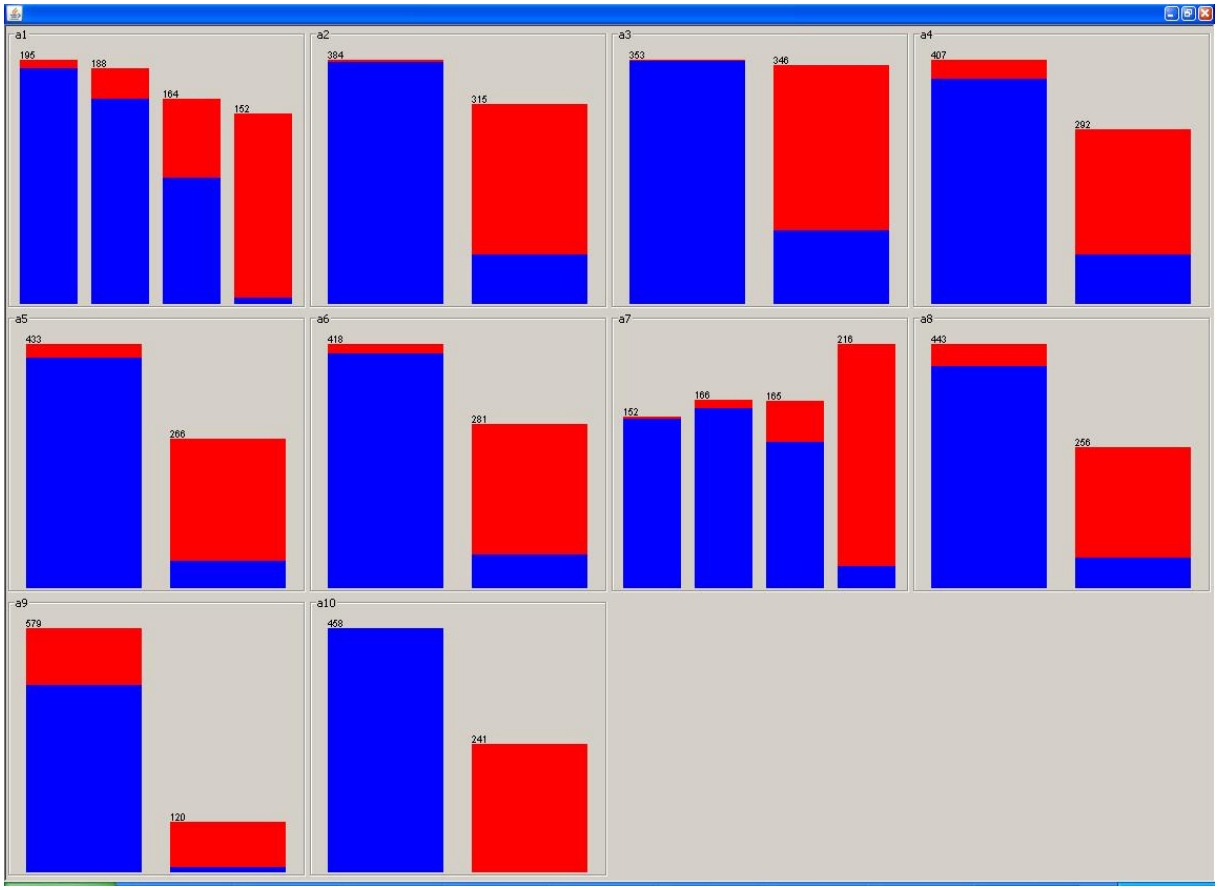
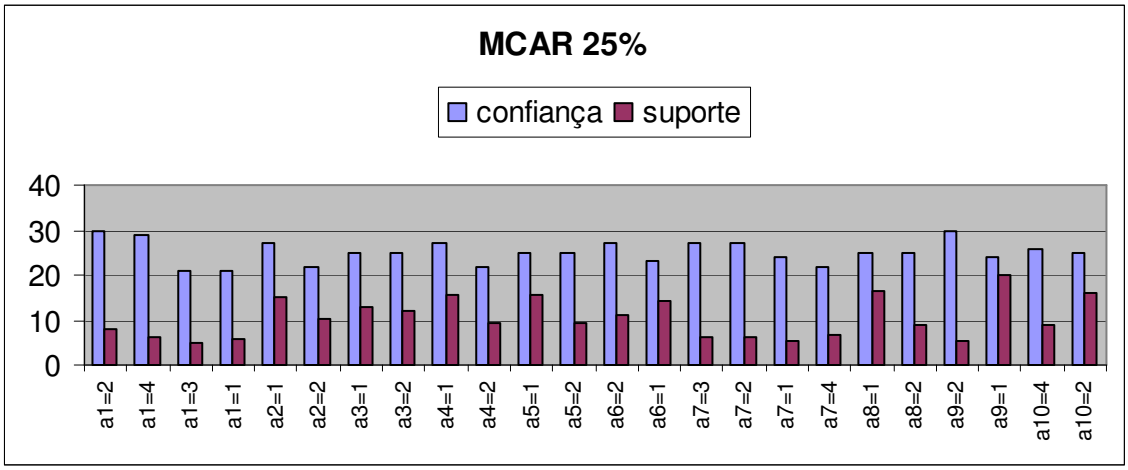


Figura 4 – Histogramas *Wisconsin* após discretização

Outro fator considerado na discretização da base de dados *Wisconsin* foi a coesão dos valores de cada intervalo. Após observação dos histogramas originais, nota-se que os valores de cada atributo relacionam-se com a classe, ou seja, valores inferiores de atributos (1, 2 e 3) geralmente representam a classe 2 (benigna), enquanto valores superiores (8, 9 e 10) referem-se a classe 4 (maligna).

Para o mecanismo MCAR nota-se que o valor da confiança é praticamente constante em todos os experimentos e também proporcional ao percentual de ausência aplicado. Além disso, nota-se variação nos valores de suporte. Esta variação ocorre quando a distribuição dos valores é muito desbalanceada entre as categorias, fato observado também no experimento com a base de dados *Mushroom*.



.

Gráfico 41 – *Wisconsin* – MCAR 25%

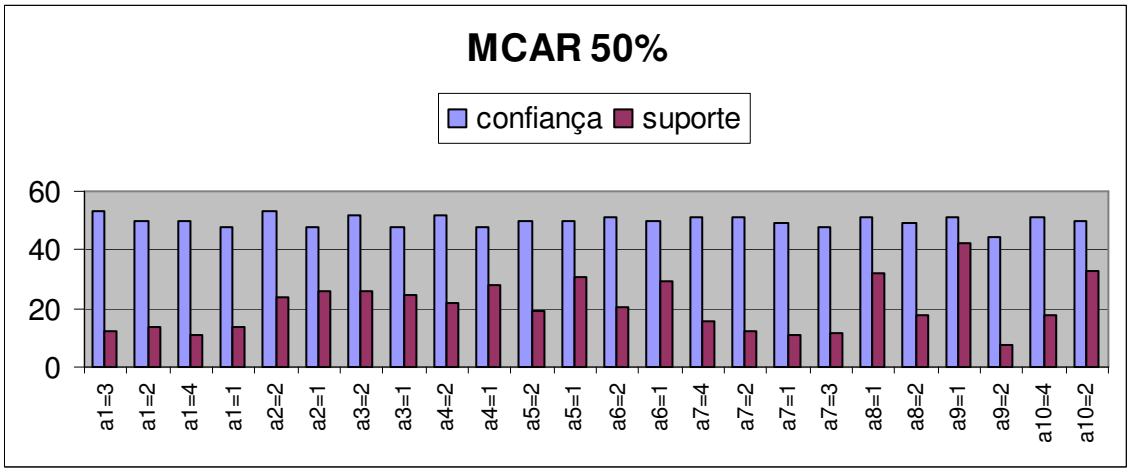


Gráfico 42 – *Wisconsin* – MCAR 50%

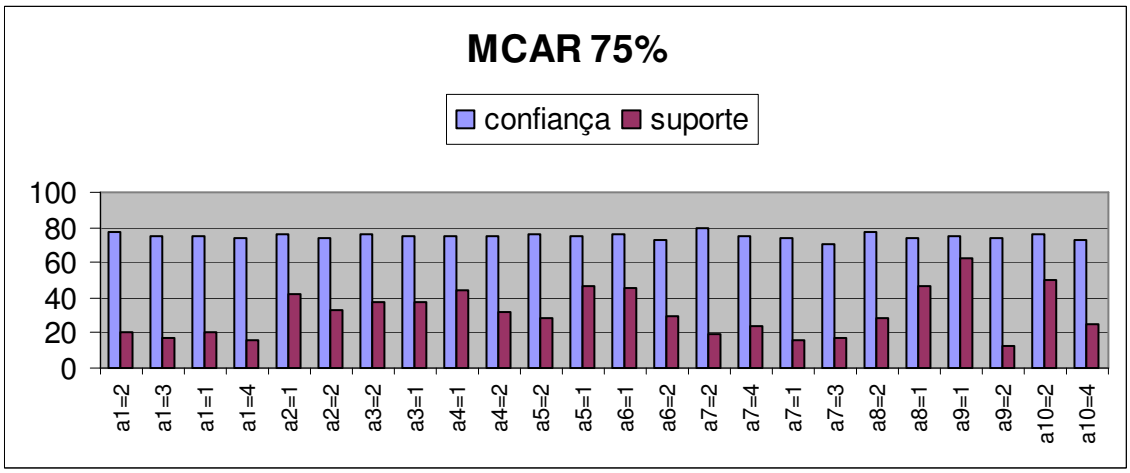


Gráfico 43 – *Wisconsin* – MCAR 75%

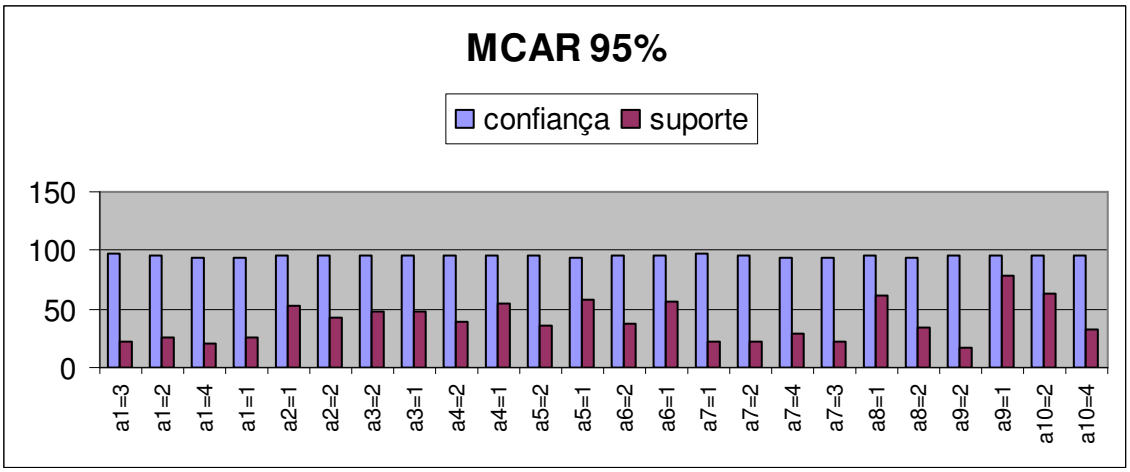


Gráfico 44 – *Wisconsin* – MCAR 95%

Na geração de ausências para os experimentos MAR utilizou-se o critério a2 < 3 no valor original do atributo, esse critério reflete na base de dados discretizada na categoria a2=1. Na análise dos gráficos MAR, várias regras apresentam alto valor de confiança, mesmo se fossem retirados os atributos mais correlacionados, ainda assim não é possível a identificação do mecanismo.

No experimento MAR 25%, a regra a3=1 $\Rightarrow$ Ausente, apresenta o maior índice *Jaccard*, entretanto a regra responsável pela ausência é a regra a2=1 $\Rightarrow$ Ausente, que apresenta um índice com valor inferior embora próximo do maior valor apurado. Nos demais experimentos MAR (50%,75% e 100%) as regras que apresentam o maior índice de *Jaccard* referem-se ao atributo a10 (atributo classe), caracterizando uma confusão entre MAR e NMAR.

Na base de dados *Wisconsin* não foi possível identificar o mecanismo MAR. Pode-se afirmar que qualquer atributo pode ser o causador da ausência, uma vez que praticamente todos os atributos possuem uma forte correlação (ver Tabela 19).

A Tabela 22 apresenta as medidas de confiança, suporte e índice *Jaccard* das principais regras produzidas.

Tabela 22

Medidas – *Wisconsin*

Experimento	Regra	Confiança	Suporte	Índice *Jaccard*
MAR 25%	a3=1 ⇒ Ausente	26	13	0,238
	a2=1 ⇒ Ausente	24	13	0,220
	a4=1 ⇒ Ausente	24	14	0,228
	a5=1 ⇒ Ausente	24	15	0,237
	a6=1 ⇒ Ausente	24	14	0,228
	a10=2 ⇒ Ausente	24	16	0,232
	a1=2 ⇒ Ausente	23	6	0,162
	a7=1 ⇒ Ausente	23	5	0,147
	a7=2 ⇒ Ausente	23	5	0,153
	a8=1 ⇒ Ausente	23	14	0,215
	a7=3 ⇒ Ausente	22	5	0,149
	a1=l ⇒ Ausente	21	6	0,149
MAR 50%	a3=1 ⇒ Ausente	50	25	0,418
	a2=1 ⇒ Ausente	49	27	0,427
	a5=1 ⇒ Ausente	48	29	0,441
	a7=2 ⇒ Ausente	48	11	0,241
	a10=2 ⇒ Ausente	48	31	0,460
	a6=1 ⇒ Ausente	47	28	0,424
	a8=1 ⇒ Ausente	47	29	0,431
	a4=1 ⇒ Ausente	46	27	0,412
	a7=3 ⇒ Ausente	46	11	0,231
MAR 75%	a8=2 ⇒ Ausente	80	29	0,497
	a2=1 ⇒ Ausente	75	41	0,626
	a3=1 ⇒ Ausente	74	37	0,579
	a7=1 ⇒ Ausente	74	16	0,279
	a4=1 ⇒ Ausente	73	42	0,630
	a5=1 ⇒ Ausente	73	45	0,661
	a6=1 ⇒ Ausente	72	43	0,622
	a10=2 ⇒ Ausente	72	47	0,678
MAR 100%	a2=1 ⇒ Ausente	100	55	0,838
	a3=1 ⇒ Ausente	100	51	0,770
	a7=1 ⇒ Ausente	100	22	0,331
	a1=1 ⇒ Ausente	97	27	0,407
	a10=2 ⇒ Ausente	97	64	0,940
	a5=1 ⇒ Ausente	96	59	0,871
NMAR 25%	a1=2 ⇒ Ausente	27	7	0,197
	a2=1 ⇒ Ausente	26	14	0,244
	a7=2 ⇒ Ausente	26	6	0,180
	a3=1 ⇒ Ausente	25	13	0,238
	a6=1 ⇒ Ausente	25	15	0,245
	a10=2 ⇒ Ausente	25	16	0,251
	a5=1 ⇒ Ausente	24	15	0,228

	a7=1 ⇒ Ausente	24	5	0,155
	a4=1 ⇒ Ausente	22	13	0,211
	a8=1 ⇒ Ausente	22	14	0,213
NMAR 50%	a2=1 ⇒ Ausente	50	27	0,456
	a3=1 ⇒ Ausente	50	25	0,440
	a10=2 ⇒ Ausente	50	33	0,500
	a6=1 ⇒ Ausente	49	29	0,457
	a7=1 ⇒ Ausente	49	11	0,245
	a1=1 ⇒ Ausente	48	13	0,280
	a4=1 ⇒ Ausente	47	27	0,426
	a5=1 ⇒ Ausente	47	29	0,445
	a7=2 ⇒ Ausente	46	11	0,238
NMAR 75%	a1=1 ⇒ Ausente	76	21	0,381
	a10=2 ⇒ Ausente	75	49	0,749
	a2=1 ⇒ Ausente	74	40	0,634
	a3=1 ⇒ Ausente	73	37	0,589
	a6=1 ⇒ Ausente	72	43	0,658
	a7=1 ⇒ Ausente	72	16	0,280
	a5=1 ⇒ Ausente	71	44	0,651
NMAR 100%	a10=2 ⇒ Ausente	100	6	1
	a2=1 ⇒ Ausente	99	54	0,822
	a3=1 ⇒ Ausente	99	50	0,763
	a7=1 ⇒ Ausente	99	21	0,326
	a1=1 ⇒ Ausente	96	27	0,404
	a6=1 ⇒ Ausente	96	57	0,844
	a7=2 ⇒ Ausente	96	23	0,341

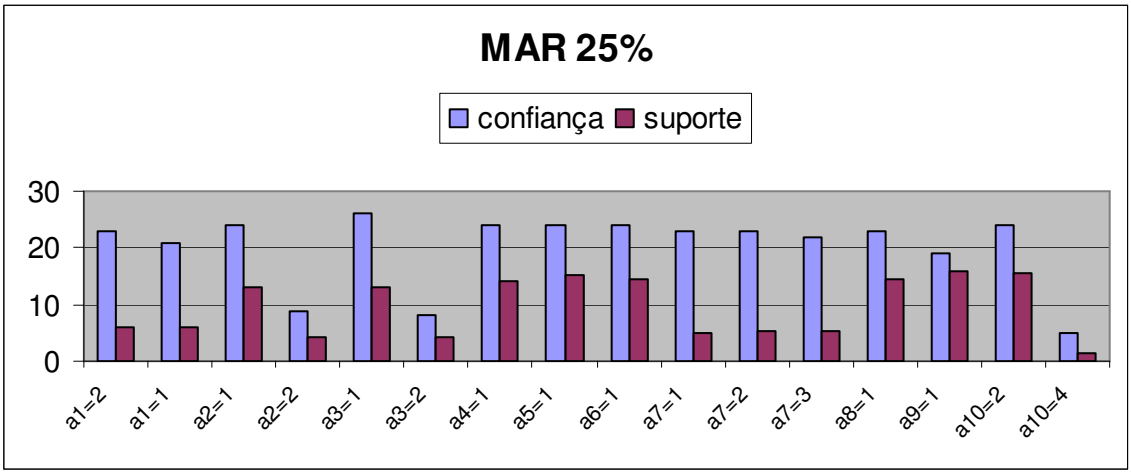


Gráfico 45 – *Wisconsin* – MAR 25%

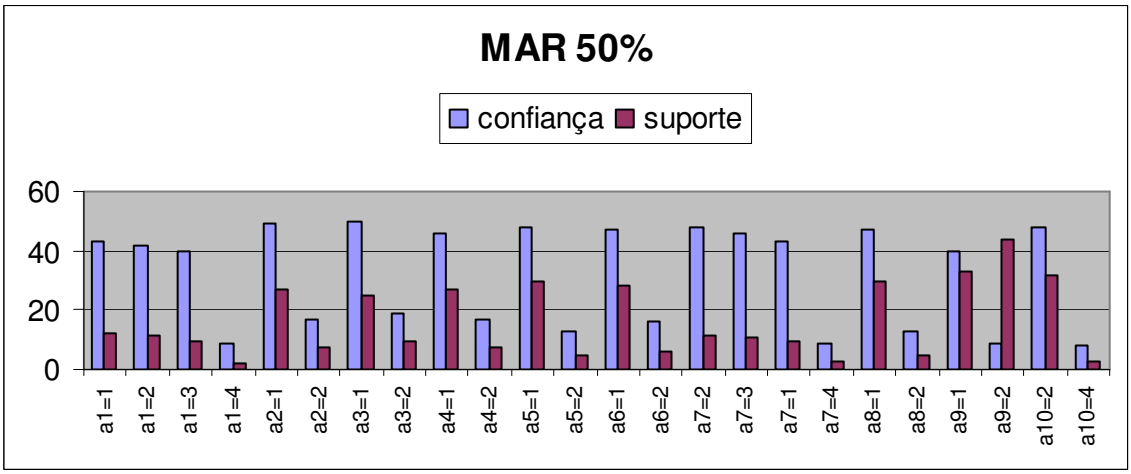


Gráfico 46 – *Wisconsin* – MAR 50%

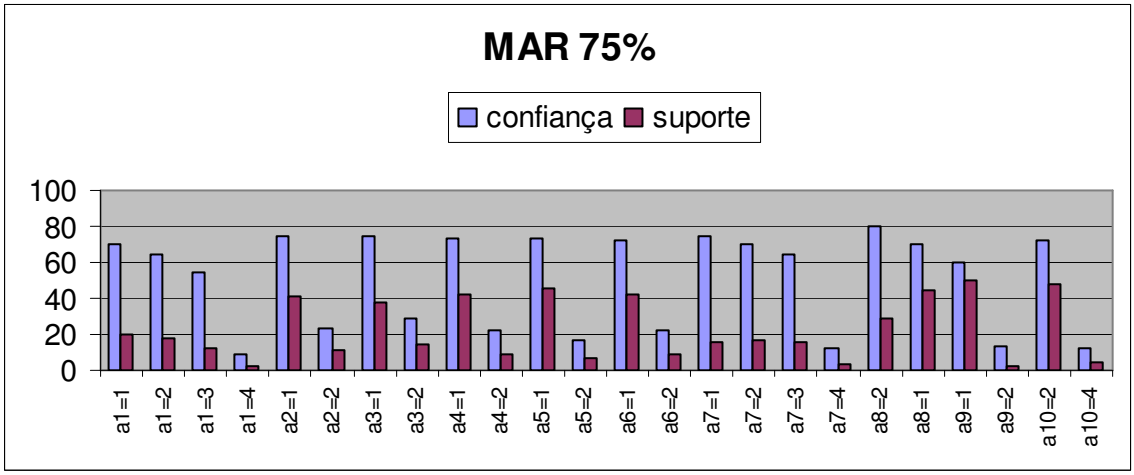


Gráfico 47 – *Wisconsin* – MAR 75%

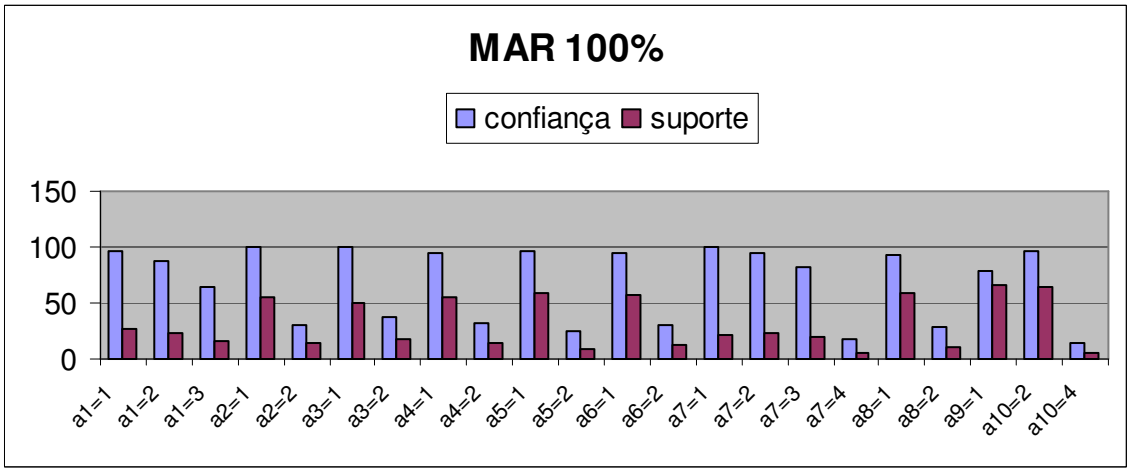


Gráfico 48 – *Wisconsin* – MAR 100%

O mecanismo NMAR foi gerado com o critério a10=2, que representa a classe benigna. O mecanismo NMAR confunde-se com o mecanismo MAR, várias regras com alto

valor de confiança são produzidas. No entanto, é possível a identificação da melhor regra que identifica a causa da ausência em cada experimento através da análise do índice *Jaccard*, como evidenciado na Tabela 22.

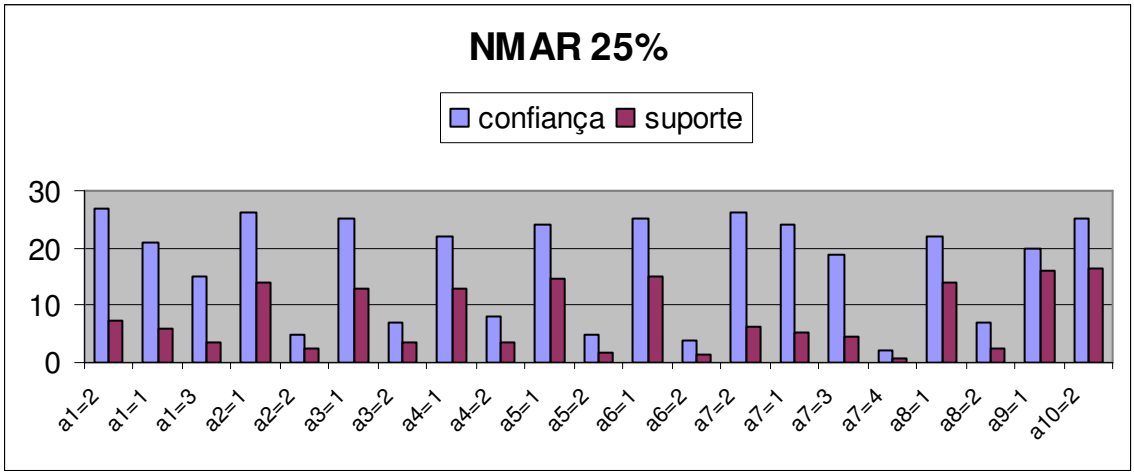


Gráfico 49 – *Wisconsin* – NMAR 25%

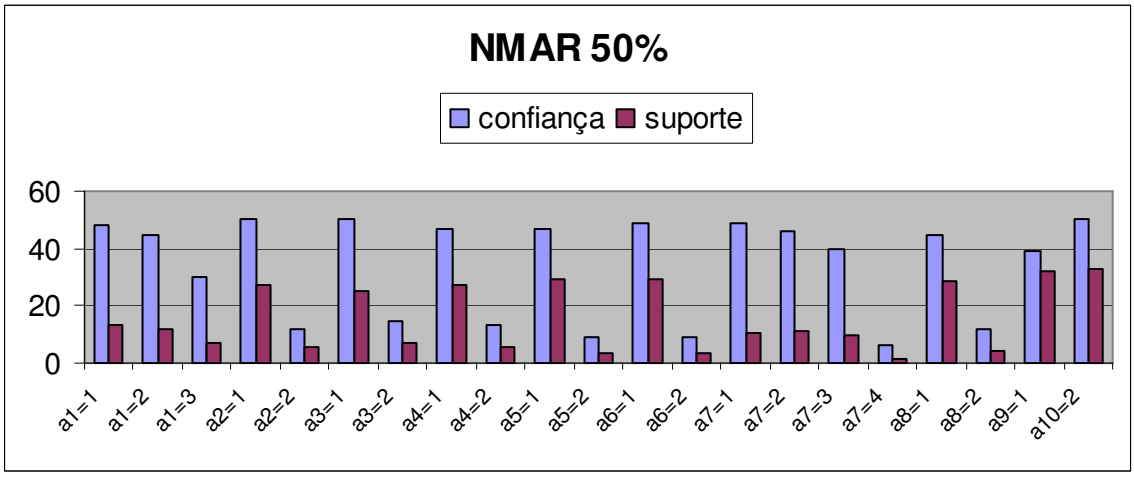


Gráfico 50 – *Wisconsin* – NMAR 50%

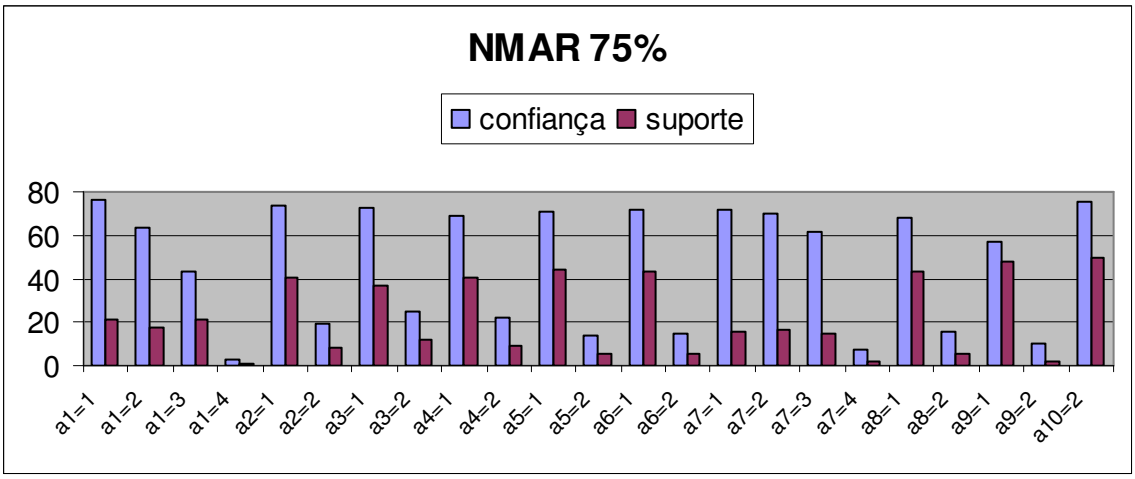


Gráfico 51 – *Wisconsin* – NMAR 75%

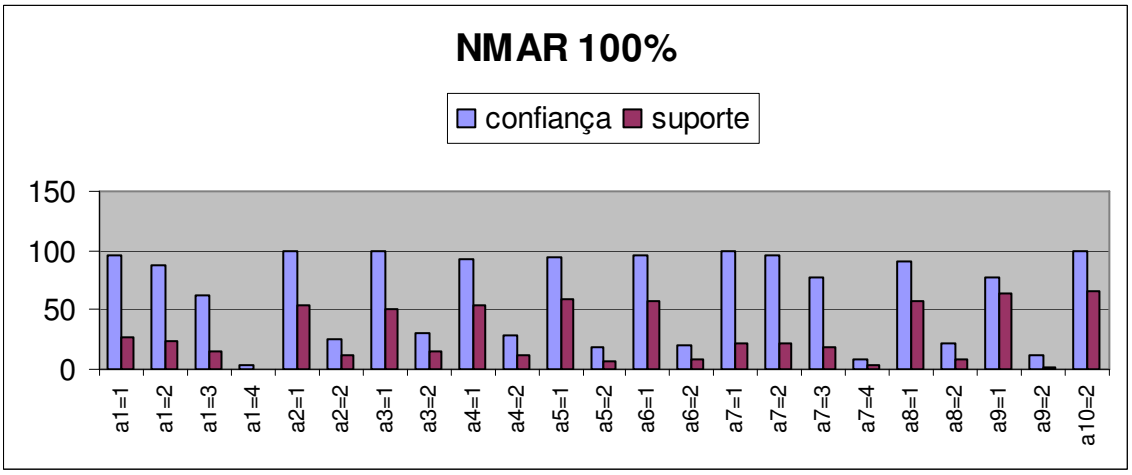


Gráfico 52 – *Wisconsin* – NMAR 100%

A aplicação do método RAIMA na base de dados *Wisconsin* obteve sucesso em 8 experimentos, na identificação dos mecanismos MCAR e NMAR. Na identificação do mecanismo MAR houve falha do método. Duas justificativas podem ser analisadas, a primeira é que o atributo utilizado para geração da ausência do mecanismo MAR (atributo a2) é altamente correlacionado ao atributo classe (atributo a10). A segunda é que o método RAIMA é sensível na identificação do mecanismo MAR quando a base de dados possui atributos altamente correlacionados ao atributo classe.

4.5 Avaliação dos resultados

Como dito, o método RAIMA foi aplicado em 3 bases de dados diferentes. Cada base de dados utilizada possui características específicas. A base de dados de Laminação a Frio possui atributos numéricos, média correlação entre os atributos e valores comportados para cada atributo, uma vez que os dados foram gerados a partir de um modelo teórico. A base de dados *Mushroom* possui atributos categóricos, baixa correlação entre os atributos e alguns atributos possuem uma distribuição de valores desbalanceada entre as categorias. Por fim, a base de dados *Wisconsin* possui atributos categóricos, alta correlação entre os atributos e alguns atributos também possuem uma distribuição de valores desbalanceada entre as categorias. Além disso, nesta base de dados foi necessário efetuar reagrupamento de categorias na tentativa de criar grupos coesos e com mesmo número de registros.

Para cada base de dados 12 experimentos foram realizados, totalizando 36 experimentos. Deste total, em apenas 4 experimentos não foi possível a identificação do mecanismo causador da ausência. Esta falha ocorreu na identificação do mecanismo MAR, na base de dados com atributos altamente correlacionados. Cabe ressaltar que para esse experimento o atributo causador da ausência possui alta correlação com o atributo classe. Nestes experimentos fica clara a confusão entre o mecanismo MAR e NMAR que, como apontado no próximo capítulo necessita de uma investigação mais detalhada, assim como a definição de um novo tipo de mecanismo: HMAR – *Hybrid Missing at Random*.

Com o objetivo de medir o desempenho do método RAIMA é proposto a seguir o índice RAIMA:

$$Índice\_RAIMA = \left[1 - \left(\frac{Número\_total\_experimentos\_com\_falha}{Número\_total\_experimentos}\right)\right] * 100\% \quad (27)$$

Note que o índice RAIMA mede o grau de sucesso do método em relação ao acerto na identificação do mecanismo de ausência numa respectiva base de dados. Na Tabela 23 o índice RAIMA para as 3 bases consideradas é apresentado.

Tabela 23

Índice RAIMA para cada base de dados

Base de dados	Número de experimentos com falha	Número Total de experimentos	Índice RAIMA
Mushroom	0	12	100%
Laminação a Frio	0	12	100%
Wisconsin	4	12	66,6%

Para ter uma noção clara da informação do índice RAIMA é necessário criar uma taxonomia para base de dados com o objetivo de identificar os tipos de base de dados em que o método RAIMA é mais apropriado.

4.6 Considerações finais

O resultado da avaliação experimental aponta que o método RAIMA é consistente, conseguiu identificar o mecanismo de ausência em quase todas as situações apresentadas e apresentou um bom índice de desempenho.

A limitação encontrada no método refere-se na identificação do mecanismo MAR quando os atributos são altamente correlacionados. Entretanto, atributos altamente correlacionados devem ser eliminados da base de dados, pois são atributos redundantes e no contexto deste trabalho evidenciaram a confusão entre os mecanismos MAR e NMAR.

No próximo capítulo são sugeridos tópicos de investigação para trabalhos futuros que poderão amenizar a limitação apresentada pelo método RAIMA. Além disso, são apresentadas outras sugestões de investigação que foram anotadas no decorrer do desenvolvimento deste trabalho.

5 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou uma ampla revisão bibliográfica referente a dados ausentes na área estatística e na área de mineração de dados. Muitos métodos já foram propostos para o tratamento de dados ausentes. Na maioria das vezes os métodos são aplicados a partir de suposições sobre o mecanismo de ausência. A identificação do mecanismo de ausência não é uma tarefa trivial, e não há na literatura relato de métodos capazes de identificar com precisão todos os mecanismos de ausência. Apenas para a identificação do mecanismo MCAR foram encontrados métodos de identificação (LITTLE, 1988). A falta de métodos para identificação do mecanismo de ausência serviu de estímulo para a proposta do método RAIMA. Portanto, a proposta de um método para identificação do mecanismo de ausência é uma iniciativa inovadora e uma contribuição na área de dados ausentes.

O método RAIMA mostrou-se consistente e alcançou desempenho favorável em quase todos os experimentos apresentados. A limitação evidenciada no método RAIMA refere-se a identificação do mecanismo MAR em base de dados com atributos altamente correlacionados, neste caso houve confusão entre os mecanismos MAR e NMAR.

Na aplicação do método RAIMA, três procedimentos mostraram-se essenciais para o sucesso do mesmo. O primeiro procedimento é a análise inicial dos dados. Neste trabalho esta análise foram utilizados histogramas de frequência. A partir da análise dos histogramas viu-se a necessidade ou não da utilização das técnicas de discretização. O segundo procedimento, aplicação de técnicas de discretização ou reagrupamento, que devem ser utilizadas quando os atributos são numéricos ou quando os atributos são categóricos e necessitam ser agrupados. Por fim, o procedimento de redução de dimensionalidade, para se obter uma representação reduzida da base de dados e regras mais representativas do domínio do problema. Nos experimentos deste trabalho foi decidido não eliminar os atributos apenas para demonstrar o problema identificado.

As regras de associação indicam as relações entre os itens de um banco de dados. O método RAIMA aplicou as regras de associação para descobrir relações entre os itens presentes e os itens ausentes em um banco de dados. Um dos grandes problemas enfrentados na análise das regras produzidas é a seleção das melhores regras. Este trabalho apresentou uma ampla revisão das medidas disponíveis para a avaliação da regras produzidas. As medidas amplamente utilizadas em regras de associação, confiança e suporte, nem sempre são adequadas para a seleção das melhores regras.

Finalmente, pode-se concluir que este trabalho é apenas um ponto inicial de várias frentes de pesquisa a serem investigadas, uma vez que vários assuntos referentes a dados ausentes foram levantados.

Uma vez que o método RAIMA parte do princípio que o valor ausente foi recuperado a partir das melhores técnicas disponíveis, o primeiro trabalho futuro sugerido é a implementação das técnicas de estimação ou imputação para aplicação do método RAIMA. A estimação ou imputação do valor ausente e a identificação do mecanismo de ausência podem trazer resultados satisfatórios ao processo de recuperação de dados ausentes, pois a identificação do mecanismo pode alimentar o processo de estimação ou imputação possibilitando a utilização da técnica mais apropriada para cada mecanismo. Assim, o processo de recuperação de dados ausentes pode ser comparado a um processo de aprendizado, o qual a identificação do mecanismo é um parâmetro que aperfeiçoa o processo de estimação ou imputação. Segundo POLETO (2006) somente quando as informações adicionais do processo de geração de dados ausentes estão disponíveis é possível identificar qual modelo é mais apropriado.

Outra sugestão para estudo futuro corresponde à definição do mecanismo híbrido de ausência. GRAHAM (2007) afirma que podem existir várias causas para a ausência, e casos puramente MAR ou puramente NMAR são apenas simplificações teóricas. Lança a hipótese da existência dos casos parcialmente MAR e NMAR, definido como mecanismo híbrido.

A definição de uma taxonomia para base de dados pode ajudar a identificar quais os tipos de base de dados são apropriadas para a utilização do método RAIMA, neste trabalho apenas a correlação de atributos foi explorada, porém, outras características podem ser investigadas em trabalhos futuros, tornando mais precisa a aplicação do método RAIMA.

Um estudo complementar que merece investigação é a visualização dos dados ausentes. TEMPL e FILZMOSER (2008) apresentam uma ferramenta de visualização da ausência que permite a exploração do dado ausente. Já WANG e WANG (2008) utilizam redes SOM para visualizar os padrões de ausências nas tarefas de classificação. A visualização da ausência pode auxiliar a identificação do mecanismo de ausência e de alguma forma melhorar o desempenho do método RAIMA.

Quanto à limitação do RAIMA na identificação do mecanismo MAR quando os atributos são altamente correlacionados, merece estudos na tentativa de buscar alternativas para lidar com atributos com essa característica no contexto de regras de associação. PLASSE et al. (2007) propõem um agrupamento de atributos correlacionados com o objetivo de diminuir o número de regras de associação.

Outra sugestão de estudo é a modificação do método RAIMA para lidar com padrões de ausência multivariados.

Além disso, efetuar avaliações experimentais em bases dados reais, que apresentam casos de *outliers*, dados desbalanceados, dados ruidosos e até mesmo dado ausente disfarçado (PEARSON, 2005a, 2005b).

Em relação às regras de associação, um trabalho experimental interessante é a comparação das diversas medidas de interesse listadas no Anexo I em relação ao índice *Jaccard*, que foi escolhido neste trabalho por medir a sobreposição dos casos do antecedente e consequente. Quanto à visualização das regras produzidas foram utilizados histogramas de frequência, porém na literatura vários trabalhos já foram publicados (NEVES, 2002). A comparação de técnicas de visualização das regras de associação na tentativa de escolher a melhor forma de visualização também é uma pesquisa que merece atenção.

REFERÊNCIAS

ABDELLA, M., MARWALA, T., Treatment of Missing Data Using Neural Networks and Genetic Algorithms. IEEE Proceedings of International Joint Conference on Neural Networks, 2005.

ACOCK, A., Working with Missing Values. Journal of Marriage and Family, v. 67, pp 1012-1028, 2005.

ACUÑA, E., RODRIGUEZ, C., The Treatment of missing values and its effect in the classifier accuracy. In Classification, Clustering and Data Mining Application, D. Banks, L. House, F. McMorris, P. Arabic, W. Gauls, Eds. Springer-Verlag, pp. 639-648, 2004.

AFIFI, A., ELASHOFF, R. Missing observations in multivariate statistics I: Review of the literature. Journal of the American Statistical Association, v. 61, pp. 595-604, 1966.

AGARWAL, S., Learning from Incomplete Data, 2001. Disponível em: <http://www.cs.ucsd.edu/users/elkan/254spring01/sagarwalrep.pdf> Acesso em: 30 jul 2007.

AGGARWAL, C., PARTHASARATHY, S., Mining Massively Incomplete Data Sets by Conceptual Reconstruction, In: Proceedings of the Seventh ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 227-232, 2001.

ALLAN, F.G., WISHART, J., A method of estimating the yield of missing plot in field experiments. J. Agric. Sci, v. 20, pp. 399-406, 1930.

AGRAWAL, R., IMIELINSKI, T., SWAMI, A., Mining Associations Between Sets of Items in Massive Databases. In: Proceedings of the ACM SIGMOD 1993 International Conference on Management of Data, pp. 207-216, Washington D.C., 1993.

AGRAWAL, R., SRIKANT, R., Fast algorithms for mining association rules. In 20th on Very Large DataBases Conference, pp. 487–499, Santiago, Chile, 1994.

ALLISON, P. D., Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods & Research, v.28, pp. 301-309, 2000.

ALLISON, P. D., Missing Data. Sage Publications, 2001.

AMEMIYA, T., Tobit Models: A survey. Journal of Econometrics, 24, pp. 3-61, 1984.

AZEVEDO, P.A., JORGE, A. M., Comparing Rule Measures for Predictive Association Rules. In: Proceedings of the 18th European conference on Machine Learning, pp. 510-517, 2007.

BATISTA, G. E. A. P. A., Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado, Tese de D. Sc., USP, São Paulo, SP, Brasil, 2003.

BATISTA, G. E. A. P. A., MONARD, M. C., A Study of K-Nearest Neighbour as a Model-Based Method to Treat Missing Data. In: Proceedings of the Argentine Symposium on Artificial Intelligence (ASAI'01), pp. 1–9, Buenos Aires, Argentina, 2001.

BATISTA, G. E. A. P. A., MONARD, M. C., An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, v.17, n. 5, pp. 519-533, 2003.

BEALE, E. M. L., LITTLE, R. J. A., Missing Values in Multivariate Analysis. Journal of the Royal Statistical Society, v. 37, n. 1, pp. 129-145, 1975.

BLAKE, C., KEOGH, E., MERZ, C. J., UCI repository of machine learning databases, 1999. Disponível em: <http://www.ics.uci.edu/~mlearn/MLRepository.html> Acesso em: 30 jul 2007.

BRIN, S., MOTWANI, R., ULLMAN, D., TSUR, S., Dynamic itemset counting and implication rules for market basket data. In Proceedings of the ACM SIGMOD Int'l Conference on Management of Data (ACM SIGMOD'97), pp. 265-276, 1997.

BROWN, M. L., KROS, J. F., Data Mining and the impact of missing data. In: Wang, J. (Author), Industrial Management & Data Systems, v. 103, n. 8, pp. 611-621 (11), Emerald Group Publishing Limited, 2003.

BUHI, E., R., GOODSON, P., NEILANDS, T. B., Out of Sigh, Not Out of Mind: Strategies for Handling Missing Data. American Journal Health Behaviour, v. 32, pp. 83-92, 2008.

CALDERS, T., GOETHALS, B., MAMPAEY, M., Mining Itemsets in the Presence of Missing Values. Proceedings of the 2007 ACM Symposium on Applied Computing, pp. 404-408, 2007.

CARPENTER, J., R., KENWARD, M., WHITE, I., R., Sensitivity analysis after multiple imputation under missing at random – a weighting approach. Statistical Methods in Medical Research, v. 16, pp. 259-275, 2007.

CARTWRIGHT, M., SHEPPERD, M.J., SONG, Q., Dealing with Missing Software Project Data. In Proceedings of the 9th International Symposium on Software Metrics, pp. 154-165, 2003.

CHANG, M. Y., Adjusting for Nonignorable Missing Data with Nonignorable Sampling Design. ASA Section on Survey Research Methods, pp. 2810-2814, 2005.

CHAWLA, N., KARAKOULAS, G. Learning from labeled and unlabeled data: An Empirical Study across Techniques and Domains. Journal of Artificial Intelligence Research, v. 23, pp. 331-366, 2005.

COLLINS, L. M., SCHAFER, J., L., KAM, C., A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. Psychological Methods, v. 6, n. 4, pp. 330-351, 2001.

COVER, T.M., HART, P.E., Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, v. 13, n. 1, pp. 21-27, 1967.

CRÉMILLEUX, B., RAGEL, A., BOSSON, J. L., An Interactive and Understandable Method to Treat Missing Values: Application to a Medical Data Set. In: Proceedings of the 5th International Conference on Information Systems Analysis and Synthesis (ISAS/SCI 99), pp. 137-144, 1999.

DELAVALLADE, T., DANG, T. H., Using Entropy to Impute Missing Data in a Classification Task. IEEE International Fuzzy Systems Conference, pp. 1-6, 2007.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B (Methodological), v. 39, n. 1, pp. 1-38, 1977.

DONDERS, A.R.T., HEIJDEN, G. J.M.G, STIJEN, T., Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, v. 59, pp. 1087-1091, 2006.

EFRON, B., Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, v. 89, n. 426, 1994.

ELDER, J. F. IV, PREGIBON, D., A Statistical Perspective on Knowledge Discovery in Databases. 1995.

FARHANGFAR, A., KURGAN, L., PEDRYCZ, W. Experimental analysis of methods for imputation values in databases. Intelligent computing. Conference N°2, ETATS-UNIS, v. 5421, pp. 172-182, 2004.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., From Data Mining to Knowledge Discovery in Databases, AI Magazine, American Association for Artificial Intelligence, pp. 37-54, 1996a.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, v. 39, n. 11, pp. 27-34, 1996b.

FORTES, I., MORA-LÓPEZ, L., TRIGUERO, F., Inductive learning models with missing values. Mathematical and Computer Modelling. v. 44, pp. 790-806, 2006.

FRANÇOIS, O., LERAY, P., Generation of Incomplete Test-Data using Bayesian Networks. In: EEE IJCNN, International Joint Conference on Neural Networks, 2007.

FRIEDMAN, N., Learning Belief Networks in the Presence of Missing Values and Hidden Variables. Fourteenth International Conference on Machine Learning (ICML97), 1997.

FUJIKAWA, Y. Efficient Algorithms for Dealing with Missing values in Knowledge Discovery. School of Knowledge Science – Japan Advanced Institute of Science and Techonology, Japan, 2001.

GARCIA, A. J. T., HRUSCHKA, E. R., Naive Bayes as an Imputation Tool for Classification Problems. Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05), 2005.

GARCÍA-LAENCINA, P.J., SANCHO-GÓMEZ, J., L., FIGUEIRA-VIDAL, A., R., Pattern Classification with Missing Values using Multitask Learning. IEEE International Joint Conference on Neural Networks, 2006.

GONÇALVES, E.C., Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf> Acesso em: 30 jan 2009.

GRAHAM, J.W., Course Session #14 Missing Data Analysis and Design, 2007. Disponível em <http://ies.ed.gov/ncer/whatsnew/conferences/rct-traininginstitute/presentations.asp> Acesso em: 30 jan 2009.

GORODETSKY, V., KARSAEV, O., SAMOILOV, V., Direct Mining of Rules from Data with Missing Values. Studies in Computational Intelligence, v. 6, pp. 233-264, 2005.

GRZYMALA-BUSSE, J. W, Rough Set Strategies to Data with Missing Attribute Values. Proceedings of the Workshop on Foundations and New Directions in Data Mining, associated with the third IEEE International Conference on Data Mining, November 19-22, Melbourne, FL, USA, 56-63, 2003.

GRZYMALA-BUSSE, J. W., HU, M., A comparison of Several Approaches to Missing Attribute Values in Data Mining, In Proceedings 2nd International Conference on Rough Sets and Current Trends in Computing RSTC, pp. 340-347, 2000.

GRZYMALA-BUSSE , J. W., SIDDHAYE, S., Rough Set Approaches to Rule Induction from Incomplete Data. Proceedings of the IPMU'2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System, Perugia, Italy, July 4-9, vol. 2, 923-930, 2004.

HAN, J., KAMBER, M., Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

HAN, J., PEI, J., YIN, Y., Mining frequent pattern without candidate generation. In Proceeding 2000 ACM-SIGMOD International Conference Management of Data (SIGMOD'00), pp. 1-12, 2000.

HARROD, L., LESSER, V., The Use of Propensity Scores to Adjust for Nonignorable Nonresponse Bias. Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 3109-3112, 2006.

HARTLEY, H. O., Maximum likelihood estimation from incompleta data. Biometrics, v. 14, pp. 174-198, 1957.

HARTLEY, H.O., HOCKING, R. The analysis of incomplete data. Biometrics, v 27, pp. 783-808, 1971.

HECKMAN, J., The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurements, v. 5, pp. 475-492, 1976.

HEDEKER, D., GIBBONS, R. D., Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies. Psychological Methods, v. 2, n. 1, pp. 64-78, 1997.

HEITJAN, D. F., Annotation: What can be done about Missing Data? Approaches to Imputation. American Journal of Public Health, v. 87, n. 4, pp. 548-549, 1997.

HEITJAN, D. F, BASU, S., Distinguishing "Missing at Random" and "Missing Completely at Random". The American Statistics, v. 50, n. 3, pp. 207-213, 1996.

HEITJAN, D. F, RUBIN, D. B., Ignorability and coarse data. The Annals of Statistics, 19(4), pp. 2244-2253, 1991.

HEWETT, R., Decision Making using Incomplete Data. IEEE International Conference on Systems, Man and Cybernetics, 2004.

HILDERMAN, R.J., HAMILTON, H.J., Knowledge Discovery and Interestingness Measures: A Survey, Technical Report CS 99-04, Department of Computer Science, University of Regina, Canada, 1999.

HORTON, N., LIPSITZ, S., Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. The American Statistician, v. 55, n. 3, pp. 244-254, 2001.

HRUSCHKA JR., E. R., Imputação Bayesiana no Contexto da Mineração de Dados. Tese de D. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2003.

HRUSCHKA, E. R., HRUSCHKA JR., E. R., EBECKEN, N. F. F., Missing values prediction with K2, Intelligent Data Analysis Journal, v. 6, n. 6, pp. 557-566, 2002.

HRUSCHKA, E. R., HRUSCHKA JR., E. R., EBECKEN, N. F. F., A Nearest-Neighbor Method as a Data Preparation Tool for a Clustering Genetic Algorithm. In: Anais do 18º Simpósio Brasileiro de Banco de Dados (SBBD), pp. 319-327, Manaus, AM, 2003a.

HRUSCHKA, E. R., HRUSCHKA JR., E. R., EBECKEN, N. F. F., Evaluating a Nearest-Neighbor Method to Substitute Continuous Missing Values. In: The 16th Australian Joint Conference on Artificial Intelligence - AI´03, Perth. Lecture Notes in Artificial Intelligence (LNAI 2903), v. 2903, pp. 723-734. Heidelberg, Springer-Verlag, 2003b.

HU, M., SALVUCCI, S. M., COHEN, M. P., Evaluation of some popular imputation Algoritms. In: The Survey Research Methods Section of the ASA, pp. 308-313, 1998.

HUA, M., PEI, J., Cleaning Disguised Missing Data: A Heuristic Approach. ACM KDD 07 Industrial and Govermment Track Paper, pp. 950-958, 2007.

HULSE, J., KHOSHGOFTAAR, T. M., SEIFFERT, C., A Comparison of Software Fault Imputation Procedures. Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA' 06), 2006.

INGRASSIA, S., DOMMA, F., Statistical Methods and Perspectives in Neural Network Training from Incomplete Data. Proceedings Atti XL Riunione Scientifica della Società Italiana di Statistica, pp. 353-364, 2000.

JACCARD, P., The distribution of the flora of the Alpine Zone. New Phytologist v. 11, pp. 37-50, 1912.

JAEGER, M., On Testing the Missing at Random Assumption. Machine Learning: ECML, v. 4212/2006, pp. 671-678, Springer Berlin, 2006.

JANSEN, I., HENS, N., MOLENBERGHS, G., AERTS, M., VERBEKE, G., KENWARD, M. G., The nature of sensitivity in monotone missing not at random models. Computational Statistics & Data Analysis, v. 50, pp. 830-858, 2006.

JAYNES, E.T., Information theory and statistical mechanics. Physical Review, 106(4), 1957.

JIANG, K., CHEN, H., YUAN, S., Classification for Incomplete Data Using Classifier Ensembles. IEEE International Conference on Neural Networks and Brain, v. 1, pp. 559-563, 2005.

JEREZ, J. M., MOLINA, I., SUBIRATS, J. L., FRANCO, L., Missing Data Imputation in Breast Cancer Prognosis. Proceedings of the 24th IASTED International Multi-Conference Biomedical Engineering, pp. 323-328, 2006.

JONSSON, P., WOHLIN, C., An Evaluation of k-Nearest Neighbour Imputation Using Likert Data. In: Proceedings of the 10th IEEE International Symposium on Software Metrics (METRICS'04), pp. 108-118, Chicago, USA, 2004.

JUNNINEN, H., NISKA, H., TUPPURRAINEN, K., RUUSKANEN, J., KOLEHMAINEN, M., Methods for Imputation of Missing Values in Air Quality Datasets. Journal of Atmospheric Environment v. 38, pp. 2895-2907, 2004.

KINNEAR, P., GRAY, C., SPSS for Windows made simple. Psychology Press, LTDA, 2000.

KLEMETTINEN, M., MANNILA, H., RONKAINEN, P., TOIVONEN, H., VERKANO, A., Finding interesting rules from large sets of discovered association rules. In R. Nabil et al. editors, Proceedings of 3rd International Conference on Information and Knowledge Management, pp. 401-407, 1994.

KONIG, T., FINKE, D., DAIMER, S., Ignoring the Non-ignorable?, European Union Politics, v. 6 (3), pp. 269-290, 2005.

LAIRD, N. M., Discussion of Informative drop-out in longitudinal data analysis by P.J. Diggle and M.G. Kenward. Applied Statistics, v. 43, pp. 84, 1994.

LATKOWSKI, R., On Decomposition for Incomplete Data. Fundamenta Informaticae, v. 54, pp. 1-16, IOS Press, 2003.

LATKOWSKI, R., MIKOLAJCZYK, M., Data Decomposition and Decision Rule Joining for Classification of Data with Missing Values. In: Rough Sets and Current Trends in Computing, RSCTC'2004, Springer, 2004.

LITTLE, R. J. A., A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, v. 83, pp. 1198-1202, 1988.

LITTLE, R. J. A., Regression with Missing X's: A Review, Journal of the American Statistical Association, v. 87, pp. 1227-1237, 1992.

LITTLE, R. J. A., Pattern-Mixture for Multivariate Incomplete Data. Journal of the American Statistical Association, v. 88, n. 421, pp. 125-134, 1993.

LITTLE, R. J. A., A class of Pattern-Mixture Models for Normal Incomplete Data. Biometrika, v. 81 (3) , pp. 471-483, 1994.

LITTLE, R. J. A., Biostatistical Analysis with Missing Data. Article for Encyclopedia of Biostatistics, P. Armitage and T. Colton, eds, Wiley: London, 1997.

LITTLE, R. J. A., RUBIN, D. B., Incomplete Data. Encyclopedia of the Statistical Sciences, v. 4, pp. 46-53, 1983.

LITTLE, R. J. A., RUBIN, D. B., Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics, 2nd Edition, 2002.

LITTLE, R. J. A., SCHENKER, N., Missing data. In: Handbook for Statistical Modeling in the Social and Behavioral Sciences, G. Arminger, C.C. Clogg and M. E. Sobel, eds, pp. 39-75, Plenum, New York, 1994.

LITTLE, R. J. A., WANG, Y., Pattern-Mixture Models for Multivariate Incomplete Data with Covariates. Biometrika, v. 52, pp. 98-111, 1996.

LI, J., CERCONE, N., Predicting Missing Attribute Values based on Frequent Itemset and RSFit. Technical Report CS-2006-13, School of Computer Science, University of Waterloo, 2006.

LIU, B., HSU, W., MA, Y., Pruning and summarizing the discovered associations. In Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD1999), 125–134, Aug, 1999.

LIU, H., HUSSAIN, F., TAN, C. L., DASH, M., "Discretization: An Enabling Technique", Data Mining and Knowledge Discovery, v. 6, pp. 393, Kluwer Academic Publishers, 2002.

LIU, P., LEI, L., Missing Data Treatment Methods and NBI Model. Proceedings on the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06), 2006.

LIU, P., LEI, L., WU, N., A Quantitative Study of the Effect of Missing Data in classifiers. In CIT'05: Proceedings of the 2005 Fifth International Conference on Computer and Information Technology, 2005.

LYNCH, S. M., Missing Data, 2003. Disponível em: <http://www.princeton.edu/~slynch/SOC_504/missingdata.pdf> Acesso em: 30 jul 2007.

MAGALHÃES, I. B., Avaliação de Redes Bayesianas para Imputação em Variáveis Qualitativas e Quantitativas. Tese de Doutorado, USP, São Paulo, SP, Brasil, 2007.

MAGNANI, M., Techniques for Dealing with Missing Data in Knowledge Discovery Tasks, 2004. Disponível em <http://magnanim.web.cs.unibo.it/index.html> Acesso em 30 jul 2007.

MAMPAEY, M., Association Rule Mining met Missing Values, Master Thesis, University of Antwerpe, Belgium, 2006.

MARLIN, B. ZEMEL, R. S., ROWEIS, S. T., Unsupervised Learning with Non-Ignorable Missing Data. In Proceeding of the Tenth International Workshop on Artificial Intelligence and Statistics, 2005.

MARLIN, B. ZEMEL, R. S., ROWEIS, S. T., SLANEY, M., Collaborative Filtering and the Missing at Random Assumption. In: Proceedings of the 23rd Conference Uncertainty in Artificial Intelligence, 2007.

MENG, X., Missing Data: Dial M for ???. Journal of the American Statistical Association, v. 95, n. 452, pp. 1325-1330, 2000.

MOLENGERGS, G., GOETGHEBEUR, E., LIPSITZ, S. R., KENWARD, M. G., Non-random missingness in Categorical data: strengths and limitations. American Statistical, v. 53, pp. 110-118, 1999.

MOLENBERGS, G., BEUNCKENS, C., SOTTO, C., KENWARD, M. G., Every Missing not at Random Model has got a Missing at Random Bodyguard. Technical Report 06103, IAP Statistics Network, Interuniversity Attraction Pole, 2002.

MOLENBERGHS G., THIJS, H., JANSEN I., BEUNCKENS, C., Analyzing incomplete longitudinal clinical trial data. Biostatistics, v. 5, n. 3, pp. 445-464, 2004.

MOONS K. G. M., DONDERS, R.A.R.T., STIJNEN, T., HARREL, F. E., Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology, v. 59, pp. 1092-1101, 2006.

MYERS, W. R., Handling Missing Data in Clinical Trials: an Overview. Drug Information Journal, v. 34, pp 525-533, 2000.

MYRTVEIT, I., STENSRUD, E., OLSSON, U. H., Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, IEEE Transactions on Software Engineering, v. 27, n. 11, Nov., 2001.

NAYAK, J. R., COOK, D. J., Approximate Association Rule Mining. In: Proceedings of the Florida Artificial Intelligence Research Symposium, pp. 259-263, 2001.

NELWAMONDO, F., V., MARWALA, T., Fuzzy Artmap and Neural Network approach to onling processing of inputs with missing values, In: SAO/NASA Astrophysics Data System, v. 705, 2007.

NELWAMONDO, F., V., MOHAMED, S., MARWALA, T., Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques. Current Science, v. 93, n. 11, pp. 1514-1521, 2007.

NEVES, J. M. P. M., Ambiente de pós-processamento para Regras de Associação, Dissertação de Mestrado, Universidade do Porto, Portugal, 2002.

NIE, N., HULL, C., JENKINS, J., STEINBRENNER, K., BENT, B., SPSS, 2nd Editon McGraw-Hill, New York, 1975.

NITTNER, T., TOUTENBURG, H., Identifying Missing Data Mechanisms in (2 x 2) – Contingency Tables. Institut fur Statistik Sonderforschungsbereich 386, Paper 373, 2004.

ORCHARD, T. WOODBURY, M. A., A Missing Information Principle: Theory and Applications, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, v. 1, pp. 697-715, 1972.

ORRE, R., BATE, A., NORÉN, G. N., SWAHN, E., ARNBORG, S., EDWARDS, I., R., A Bayesian Recurrent Neural Network for Unsupervised Pattern Recognition in Large Incomplete Data Sets, 2003.

PAWLAK, Z., Rough Sets. Intelligent Journal Computer and Information Science, v. 11, pp 341-356, 1982.

PEARSON, R., The Problem of Disguised Missing Data, SIGKDD Explorations, v. 8, n. 1, pp. 83-92, 2005a.

PEARSON, R., Mining Imperfect Data: Dealing with Contamination and Incomplete Records. SIAM, 2005b.

PENG, H., ZHU, S., Handling of incomplete data sets using ICA and SOM in data mining. Neural Comput & Applic v. 16, pp. 167-172, 2007.

PIATETSKY-SHAPIRO, G., Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases, pp229-248, 1991.

PLASSE, M., NIANG, N., SAPORTA, G., VILLEMINOT, A., LEBLOND, L., Combined use of Association Rules Mining and clustering methods to find relevant links between binary rare attributes in a large data set. Computational Statistics & Data Analysis, v. 52, n. 1, pp. 596-613, 2007.

POLETO, F. Z., Análise de dados categorizados com omissão. Dissertação de Mestrado, USP, São Paulo, SP, Brasil, 2006.

PREISSER, J. S., LOHMAN, K. K., RATHOUZ, P. J., Performed of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Statistics in Medicine, v. 21, pp. 3035-3054, 2002.

PYLE, D., Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999.

QUINLAN, J. R., C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: http://www.R-project.org. ISBN 3-900051-07-0, 2007.

RAGEL, A., CRÉMILLEUX, B., Treatment of Missing Values for Association Rules, In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 258-270, Melbourne, Australia, 1998.

RAGEL, A., CRÉMILLEUX, B., MVC – A Preprocessing Method to Deal With Missing Values, Knowledge-Based Systems, v. 12, n. 5–6 , pp. 285-291, 1999.

RAGHUNATHAN, T. E., What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data, Annual Review of Public Health, v. 25, pp. 99-117, 2004..

RIGGELSEN, C., FEELDERS, A., Learning Bayesian Network Models from Incomplete Data using Importance Sampling, In: Proceedings Tenth International Workshop on Artificial Intelligence and Statistics, 2005.

ROSENBAUM, P. R., RUBIN, D. B., The Central Role of the propensity score in observational studies for causal effects. Biometrika, v. 70, pp. 41-56, 1983.

RUBIN, D. B., Inference and missng data, Biometrika, v. 63(3), pp.581-592, 1976.

RUBIN, D. B., An Overview of Multiple Imputation, In: Proceedings of the Section on Survey Research Methods, pp. 79-84, American Statistical Association, 1988.

RUBIN, D. B., Multiple Imputation after 18+ years. Journal American Statistical Association, v. 91, pp. 473-489, 1996.

RUBIN, D. B., Conceptual, computational and inferential benefits of the missing data perspective in applied and theoretical statistical problem. Allgemeines Statistisches Archiv, v. 90, pp. 501-513, 2006.

SCHAFER, J. L., OLSEN, M. K., Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research, v. 33, pp. 545-571, 1998.

SCHAFER, J. L., GRAHAM, J. W., Missing Data: Our View of the State of the Art, Psycological Methods, v. 7, n. 2, pp. 147-177, 2002.

SCHEFFER, J., Dealing with Missing Data, Research Letters in the Information and Mathematical Sciences, v. 3, pp. 153-160, 2002.

SHADISH, W., Finding a solution for missing Data, 2002. Disponível em: <http://www.apa.org/monitor/julaug02/missingdata.html> Acesso em: 30 jul 2008.

SHALABI, L, A., NAJJAR, M., KAYED, A., A framework to Deal with Missing Data in Data sets. Journal of Computer Science, v. 2, pp. 740-745, 2006.

SHANNON, C. E., A Mathematical theory of Communication. Bell Syst. Tech. J. 27, pp. 379-423, 1948.

SCHONER, H., Working with Real-World Datasets: Preprocessing and prediction with large incomplete and heterogeneous datasets. Phd Thesis, Berlin University of Techonology, Berlin, 2004.

SMITH, A., Learning from Data Sets with Missing Labels, 2005. Disponível em: <http://www.cse.ucsd.edu/~atsmith/atsmith_rexam.pdf> Acesso em: 30 jul 2007.

SOARES, P. J. J., Análise Bayesiana de Dados Deficientemente Categorizados. Tese de D. Sc, Universidade Técnica de Lisboa, Lisboa, Portugal, 2004.

SOARES, J. A., Pré-Processamento em Mineração de Dados: Um estudo comparativo em Complementação. Tese de D. Sc, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2007.

SONG, Q., SHEPPERD, M., A new imputation method for small software project data sets. The Journal of System and Software, v. 80, pp. 51-62, 2007.

TAN, P.N., KUMAR, V., SRIVASTAVA, J., Selecting the Right Interestingness Measure for Association Patterns, In: Proceedings of the eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, pp. 32-41, 2002.

TEMPL, M., FILZMOSER, P., Visualization of Missing Values using the R-Package VIM, Forschungsberit CS-2008-1, 2008.

THIJS, H., MOLENBERGHS, G., MICHIELS, B., MICHIELS, B., VERBEKE, G., CURRAN, D., Strategies to fit pattern-mixture models. Biostatistics, v. 3, pp. 245-265, 2002.

TOIVONEN, H., KLEMETTINEN, M., RONKAINEN, P., HATONEN, K., MANNILA, H., Prunning and Grouping Discovered Association Rules. In MLNet Workshop on Statistics, Machine Learning and Discovery in Databases, pp. 47-52, 1995.

TRESP, V., NEUNEIER, R., AHMAD, S., Efficient Methods for Dealing with Missing Data in Supervised Learning. In: G. Tesauro, D. S. Touretzky and T. K. Leen, eds., “Advances in Neural Information Processing System 7”, MIT Press, Cambridge MA, 1995.

TSECHANSKY, M. S., PROVOST, F., Handling Missing Values when Applying Classification Models. Journal of Machine Learning Research, v. 8, pp. 1625-1657, 2007.

TSIKRIKTSIS, N., A review of techniques for treating missing data. In OM Survey Research. Journal of Operations Management, v. 24, pp. 53-62, 2005.

TWALA, B., CARTWRIGHT, M., SHEPPERD, M., Comparison of Various Methods for Handling Incomplete Data in Software Engineering Databases. In: 2005 International Symposium on Empirical Software Engineering, pp. 105-114, 2005.

TWALA, B., CARTWRIGHT, M., SHEPPERD, M., Ensemble of Missing Data Techniques to Improve Software Prediction Accuracy. In: Proceedings of the 28th International Conference on Software Engineering, 2006.

VIHAROS, Zs., J., MONOSTORI, L., VINCZE, T., Training and application of artificial neural networks with incomplete data. Lecture Notes of Artificial Intelligence, LNAI 2358. The Fifteenth International Conference on Industrial & Engineering Application of Artificial Intelligence & Expert Systems. Springer Computer Science, pp. 649-659, 2002.

WAGSTAFF, K., L., Clustering with missing values: No imputation required. Classification, Clustering and Data Mining Applications. In: Proceedings of the Meeting of the International Federation of Classification Societies, Springer, pp. 649-658, 2004.

WAGSTAFF, K., L., LAIDLER, V. G., Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy. Astronomical Data Analysis Software and System XIV ASP Conference Series, v. 30, 2005.

WANG, H., WANG, S., Visualization of the Critical Patterns of Missing Values in Classification Tasks. Lectures Notes in Computer Science, v. 4781/2007, 267-274, 2007.

WASITO, I., MIRKIN, B., Nearest neighbours in least-squares data imputation algorithms with different missing patterns. Computational Statistics & Data Analysis, v. 50, pp. 926-949, 2006.

WAYMAN, J. C., Multiple Imputation For Missing Data: What Is It And How Can I Use It? In: Proceedings of the Annual Meeting of the American Educational Research Association, Chicago, IL, 2003.

WEI, W., TANG, Y., A Generic Neural Network Approach for Filling Missing Data in Data Mining. IEEE International Conference on Systems, Man and Cybernetics, 2003.

WEISS, S., INDURKHYA, N., Decision-Rule Solutions for Data Mining with Missing Values. IBM Research Report RC-21783, 1999.

WEKA, Weka 3: Data Mining Software in Java. Disponível em http://www.cs.waikato.ac.nz/ml/weka/ em 30/07/2007.

WETTSHERECK, D., A KDDSE-independent PMML Visualizer. In Marko Bohamec, Dunja Mlandenic and Nada Lavrac, editors. ECML/PKDD02 Workshop on Integration Aspects of Data Mining, Decision Support and Meta-Learning, 2002.

WILLIAMS, D., LIAO, X., XUE, Y., CARIN, L., On Classification with Incomplete Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 29, n. 3, 2007.

WU, C., WUN, C, CHOU, H., Using Association Rules for Completing Missing Data. IEEE Proceedings of the Fourth International Conference on Hybrid Intelligent System, 2004.

YU, L., WANG, S., LAI, K., An Integrated Data Preparation Scheme for Neural Network Data Analysis. IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 2, 2006.

ZÁRATE, L. E., Um método para Análise da Laminação Tandem a Frio. Tese de Doutorado, UFMG, Belo Horizonte, MG, Brasil, 1998.

ZÁRATE, L. E., NOGUEIRA, B. M., SANTOS, T. R. A., Recuperação de Dados Ausentes através e Redes Neurais Artificiais – Estudo de Caso para uma Base de Dados Mercadológica, Congresso Brasileiro de Redes Neurais, CBRN 2005, Brasil, 2005.

ZÁRATE, L. E., NOGUEIRA, B. M., SANTOS, T. R. A., Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data. In: IEEE CIDM, 2007, Hawai. Proc. Symposium on Computational Intelligence and Data Mining, pp. 66-72, 2007.

ZÁRATE, L. E., NOGUEIRA, B. M., SANTOS, T. R. A., SONG, M. A. J., Techniques for Missing Data Recovering in Imbalanced Databases - Application in a Marketing Databases with Massive Missing Data. In: IEEE SMC, 2006, Taipei. Proc. of the 19th Int. Conf. on Systems, Man, and Cybernetics, v. 1. pp. 1-8, 2006.

ZHANG, S., QIN, Z., LING, C. X., SHENG, S., "Missing is Useful": Missing Values in Cost-Sensitive Decision Trees. IEEE Transactions on Knowledge and Data Enginneering, v. 17, n. 12, 2005.

ZOU, Y., AN, A., HUANG, X., Evaluation and Automatic Selection of Methods for Handling Missing Data. In Proceedings of the IEEE International Conference in Granular Computing, pp. 728-733, 2005.

Anexo A

Tabela 24

Domínio dos atributos - *Mushroom*

Atributo	Domínio
Class	e, p
cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
cap-color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
Bruises	bruises=t, no=f
Odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
gill-attachment	attached=a, descending=d, free=f, notched=n
gill-spacing	close=c, crowded=w, distant=d
gill-size	broad=b, narrow=n
gill-color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
stalk-shape	enlarging=e, tapering=t
stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
stalk-surface-above-ring	ibrous=f, scaly=y, silky=k, smooth=s
stalk-surface-below-ring	ibrous=f, scaly=y, silky=k, smooth=s
stalk-color-above-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p, red=e, white=w, yellow=y
veil-type	partial=p,universal=u
veil-color	brown=n,orange=o,white=w,yellow=y
ring-number	none=n,one=o,two=t
ring-type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
Habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

Anexo B

Tabela 25

Principais medidas de interesse e intervalos

Medida	Fórmula	*range*
Phi-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$	[-1,1]
Goodman-Kruskal's	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$	[0,1]
Odds ratio	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$	[0,∞[
Yule's Q	$\frac{P(A,B)P(\overline{A}\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A}\overline{B})+P(A,\overline{B})P(\overline{A},B)}=\frac{\alpha-1}{\alpha+1}$	[-1,1]
Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{A}\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A}\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}=\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$	[-1,1]
Kappa	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$	[-1,1]
Mutual Information	$\frac{\sum_i\sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$	[0,1]
J-Measure	$\max(P(A,B)\log(\frac{P(B\mid A)}{P(A)})+P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(\overline{A})}))$	[0,1]
Gini index	$\max(P(A)[P(B\mid A)^2+P(\overline{B}\mid A)^2]+P(\overline{A})[P(B\mid\overline{A})^2+P(\overline{B}\mid\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\mid B)^2+P(\overline{A}\mid B)^2]+P(\overline{B})[P(A\mid\overline{B})^2+P(\overline{A}\mid\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2)$	[0,1]
Support	$P(A,B)$	[0,1]
Confidence	$P(A\mid B)$	[0,1]

Laplace	$\frac{NP(A,B)+1}{NP(A)+2}$	[0,1]
Conviction	$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$	[0,5,∞[
Interest	$\frac{P(A,B)}{P(A)P(B)}$	[0,∞[
Cosine	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$	[0,1]
Piatetsky-Shapiro's	$P(A,B)-P(A)P(B)$	[-0,25, 0,25]
Certainty factor	$\max(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)})$	[-1,1]
Added Value	$\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$	[-0,5,1]
Colletive strength	$\frac{P(A,B)+P(\overline{A}\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{A}\overline{B})}$	[0,∞[
Jaccard	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$	[0,1]
Klosgen	$\sqrt{P(A,B)}\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$	[-0,12, 0,38]

(TAN, KUMAR e SRIASTAVA, 2002)

Livros Grátis

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

Milhares de Livros para Download:

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