Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Agrupamento de regras de associação no pré-processamento e no pós-processamento: o que vale mais a pena?

Veronica Oliveira de Carvalho
Fabiano Fernandes dos Santos
Solange Oliveira Rezende

Nº 381

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
Agosto/2012

Agrupamento de regras de associação no pré-processamento e no pós-processamento: o que vale mais a pena?[1]

Veronica Oliveira de Carvalho[1]
Fabiano Fernandes dos Santos[2]
Solange Oliveira Rezende[2]

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Departamento de Estatística, Matemática Aplicada e Computação
Rio Claro, SP - Brasil
[2]Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
São Carlos, SP - Brasil
e-mail: veronica@rc.unesp.br, {fabianof, solange}@icmc.usp.br

Resumo: Um dos problemas relacionados a técnica de associação se refere a quantidade de padrões extraídos. Geralmente o usuário é sobrecarregado com um grande volume de regras dificultando assim a identificação dos padrões relevantes a aplicação. Visando solucionar esse problema, várias abordagens têm sido propostas para mineração de associação, entre as quais encontra-se a de agrupamento. O agrupamento pode ser aplicado em duas etapas distintas do processo de mineração de dados, a saber: pré-processamento e pós-processamento. A utilização do agrupamento em cada uma das etapas possui suas particularidades, as quais influenciam os resultados obtidos. Nesse contexto, surge a pergunta: o que vale mais a pena? Visando responder essa pergunta, este relatório técnico apresenta um estudo divido em três partes: (i) a contribuição do agrupamento no pós-processamento de regras de associação, (ii) a contribuição do agrupamento no pré-processamento de regras de associação e (iii) uma comparação do agrupamento nas duas etapas considerando (i) e (ii). Para tanto, diversos processos e metodologias são propostas, gerando contribuições complementares à area de pesquisa em questão. Ao final de todo o estudo, apresenta-se a resposta: o que vale mais a pena?

Palavras-chave: regras de associação, pré-processamento, pós-processamento, agrupamento.

Agosto/2012

[1]Trabalho realizado com o apoio da FAPESP, processo 2010/07879-0.

Sumário

Lista de Figuras **iii**

Lista de Tabelas **vi**

1 Introdução **1**

2 Conceitos **4**
2.1 Regras de Associação 4
2.2 Agrupamento 4
2.3 Trabalhos Relacionados 5
2.3.1 Pós-processamento 6
2.3.2 Pré-processamento 7

3 Agrupamento de Regras de Associação no Pós-processamento **11**
3.1 Redução via Interesse 11
3.1.1 Experimentos 14
3.1.2 Resultados e Discussão 17
3.2 Descritivo/Representativo 25
3.2.1 Métodos de Rotulação 27
3.2.2 Metodologia de Avaliação 29
3.2.3 Experimentos 30
3.2.4 Resultados e Discussão 31
3.3 Considerações Finais 39

4 Agrupamento de Regras de Associação no Pré-processamento **42**
4.1 Avaliação das Organizações 43
4.2 Índices de Avaliação 48
4.2.1 Experimentos 52
4.2.2 Resultados e Discussões 53
4.3 Redução via Interesse 62
4.3.1 Experimentos 63
4.3.2 Resultados e Discussão 64
4.4 Descritivo/Representativo 66
4.4.1 Experimentos 67
4.4.2 Resultados e Discussão 68
4.5 Considerações Finais 71

5 Avaliação Comparativa entre o Pós-processamento e o Pré-processamento **74**
5.1 Comparação Objetiva: "Redução via Interesse" e "Descritivo/Representativo" 74
5.2 Comparação Subjetiva: Análise Especialista 77
5.3 Considerações Finais 92

6 Conclusão **93**

Referências **95**

Lista de Figuras

1 Visão geral do processo de agrupamento de regras de associação no pré-processamento e no pós-processamento. 2
2 Metodologia PAR-COM [Carvalho et al., 2011, Carvalho et al., 2012]. . . . 13
3 Distribuição das regras nos agrupamentos dos CD-P. 16
4 Distribuição das regras nos agrupamentos dos CD-NP. 17
5 Comportamento da PAR-COM, na organização Ward:J-RT, nos CD-P. . . 25
6 Comportamento da PAR-COM, na organização Ward:J-RT, nos CD-NP. . 26
7 Comportamento dos métodos LM-T e LM-PU, na organização Ward:J-RT, nos CD-P. 40
8 Comportamento dos métodos LM-T e LM-PU, na organização Ward:J-RT, nos CD-NP. 40
9 Reprodução da Figura 1(a), página 2: Visão geral do processo de agrupamento de regras de associação no pré-processamento. 43
10 Distribuição das transações nos agrupamentos dos CD-P. 46
11 Distribuição das transações nos agrupamentos dos CD-NP. 46
12 Distribuição das transações nos agrupamentos dos CD-P para um novo intervalo de k. 47
13 Distribuição das transações nos agrupamentos dos CD-NP para um novo intervalo de k. 47
14 Adaptação da metodologia PAR-COM [Carvalho et al., 2011, Carvalho et al., 2012] para o pré-processamento. 63
15 Comportamento da PAM-COM adaptada nos conjuntos de dados Adult e Income. 67
16 Comportamento da PAM-COM adaptada nos conjuntos de dados Groceries e Sup. 67
17 Comportamento dos métodos LM-T e LM-PU, na organização PAM:Agg, nos CD-P. 71
18 Comportamento dos métodos LM-T e LM-PU, na organização PAM:Agg, nos CD-NP. 72

Lista de Tabelas

1 Detalhes dos conjuntos de dados usados nos experimentos. 15
2 Configurações usadas para avaliar a PAR-COM. 17
3 Resultados do PAM, na PAR-COM, no conjunto de dados Adult. 19
4 Resultados do PAM, na PAR-COM, no conjunto de dados Income. 19
5 Resultados do PAM, na PAR-COM, no conjunto de dados Groceries. . . . 20
6 Resultados do PAM, na PAR-COM, no conjunto de dados Sup. 20
7 Resultados do Ward, na PAR-COM, no conjunto de dados Adult. 21
8 Resultados do Ward, na PAR-COM, no conjunto de dados Income. 21
9 Resultados do Ward, na PAR-COM, no conjunto de dados Groceries. . . . 22
10 Resultados do Ward, na PAR-COM, no conjunto de dados Sup. 22
11 Média de h, h' e R, nos k's entre 5-100, nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida). 24
12 Média de h, h' e R, nos k's entre 5-50, nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida). 26
13 Configurações usadas para avaliação dos métodos de rotulação no pós-processamento. 30
14 Resultados do PAM, nos LM's, no conjunto de dados Adult. 32
15 Resultados do PAM, nos LM's, no conjunto de dados Income. 32
16 Resultados do PAM, nos LM's, no conjunto de dados Groceries. 33
17 Resultados do PAM, nos LM's, no conjunto de dados Sup. 33
18 Resultados do Ward, nos LM's, no conjunto de dados Adult. 34
19 Resultados do Ward, nos LM's, no conjunto de dados Income. 34
20 Resultados do Ward, nos LM's, no conjunto de dados Groceries. 35
21 Resultados do Ward, nos LM's, no conjunto de dados Sup. 35
22 Média de P e RF nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida) em cada LM. 37
23 Média de P e RF nos diferentes tipos de dado nas organizações que se destacaram na Seção 3.1.2 em cada LM. 39
24 Organizações e métodos de rotulação adequados ao pós-processamento referente aos aspectos "Redução via Interesse" e "Descritivo/Representativo". 41
25 Configurações usadas para selecionar as organizações mais adequadas ao pré-processamento segundo o critério de distribuição das transações. 45
26 Síntese dos índices propostos. 52
27 Configurações usadas para avaliar a abordagem de agrupamento no pré-processamento segundo os índices propostos. 53
28 Resultados do PAM, segundo os índices propostos, no conjunto de dados Adult. 54
29 Resultados do PAM, segundo os índices propostos, no conjunto de dados Income. 54
30 Resultados do PAM, segundo os índices propostos, no conjunto de dados Groceries. 55
31 Resultados do PAM, segundo os índices propostos, no conjunto de dados Sup. 55

32 Resultados do Ward, segundo os índices propostos, no conjunto de dados Adult. 55
33 Resultados do Ward, segundo os índices propostos, no conjunto de dados Income. 56
34 Resultados do Ward, segundo os índices propostos, no conjunto de dados Groceries. 56
35 Resultados do Ward, segundo os índices propostos, no conjunto de dados Sup. 56
36 Média dos índices propostos nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida). 61
37 Média dos índices propostos nos diferentes tipos de dado na organização PAM:Agg. 61
38 Configurações usadas para aplicação da PAR-COM adaptada. 64
39 Resultados da organização PAM:Agg, na PAR-COM adaptada, nos conjuntos de dados Adult e Income. 65
40 Resultados da organização PAM:Agg, na PAR-COM adaptada, nos conjuntos de dados Groceries e Sup. 65
41 Média de h, h', R e $RNovas$ nos diferentes tipos de dado na organização PAM:Agg. 66
42 Configurações usadas para avaliação dos métodos de rotulação no pré-processamento. 68
43 Resultados da organização PAM:Agg, nos LM's, nos conjuntos de dados Adult e Income. 69
44 Resultados da organização PAM:Agg, nos LM's, nos conjuntos de dados Groceries e Sup. 69
45 Média de P e RF nos diferentes tipos de dado na organização PAM:Agg em cada LM. 70
46 Organizações e métodos de rotulação adequados ao pré-processamento referente as particularidades da abordagem e aos aspectos "Redução via Interesse" e "Descritivo/Representativo". 73
47 Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Redução via Interesse". 75
48 Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Descritivo/Representativo". 76
49 Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Redução via Interesse" e intervalo de k padronizado. 77
50 Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Descritivo/Representativo" e intervalo de k padronizado. . . 77
51 Agrupamentos selecionados para comparação subjetiva das abordagens de pós-processamento e pré-processamento. 78
52 Descritores dos agrupamentos selecionados para o conjunto de dados Adult. 80
53 Descritores dos agrupamentos selecionados para o conjunto de dados Income. 81
54 Descritores dos agrupamentos selecionados para o conjunto de dados Groceries. 82
55 Descritores dos agrupamentos selecionados para o conjunto de dados Sup. . 83

56 Resultados da PAR-COM/PAR-COM adaptada nos agrupamentos selecionados e apresentados na Tabela 51. 87
57 Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Adult. 90
58 Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Income. 90
59 Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Groceries. 91
60 Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Sup. 91
61 Síntese dos resultados das comparações objetiva e subjetiva realizadas entre as abordagens de pós-processamento e pré-processamento. 92

Lista de Abreviaturas

Af - Afinidade

Agg - Medida de similaridade Aggarwal

CD-NP - Conjunto de Dados Não Padronizados

CD-P - Conjunto de Dados Padronizados

conf - Confiança

conf-min - Confiança Mínima

DUPI - *Directing the User to what is Potentially Interesting*

GD - Grupo de Dados

GR - Grupo de Regras

$\mathbf{I}_{A-I-N-RcP}$ - Índice da proporção de regras "antigas" não existentes em RcP entre as h-top regras interessantes contidas em RcNP

$\mathbf{I}_{A-Rcp}$ - Índice da proporção de regras "antigas" em RcP

$\mathbf{I}_{C-I}$ - Índice da proporção de regras em comum entre as h-top regras interessantes em RcP e as h-top regras interessantes em RcNP

$\mathbf{I}_{NC-I-RcP}$ - Índice da proporção de grupos do agrupamento gerador de RcP que contém as h-top regras interessantes contidas em RcP

$\mathbf{I}_{N-I-RcP}$ - Índice da proporção de regras "novas" entre as h-top regras interessantes contidas em RcP

$\mathbf{I}_{N-RcP}$ - Índice da proporção de regras "'novas" em RcP

$\mathbf{I}_{NR-RcP}$ - Índice da proporção de regras obtidas em RcP em relação as obtidas em RcNP

$\mathbf{I}_{R-A-RcP}$ - índice da proporção de regras "antigas" em RcP que se repetem

$\mathbf{I}_{R-N-RcP}$ - Índice da proporção de regras "novas" em RcP que se repetem

J-RI - *Jaccard with Rules by Items*

J-RT - *Jaccard with Rules by Transactions*

LHS - *Left Hand Side*

LM - *Labeling Methods*

LM-M - *Labeling Method Medoid*

LM-PU - *Labeling Method Popescul e Ungar*

LM-S - *Labeling Method Sahar*

LM-T - *Labeling Method Transaction*

MT - Mineração de Textos

P - Medida Precisão (*Precision*)

PAM - Algoritmo *Partitioning Around Medoids*

PAR-COM - Post-processing Association Rules with Clustering and Objective Measures

PIK - *Potentially Interesting Knowledge*

P-J - Medida de similaridade *Plasse Jaccard*

P-RR - Medida de similaridade *Plasse Russel Rao*

RA - Regras de Associação

RcNP - Conjunto de Regras obtido a partir de um Conjunto de dados Não Particionado

RcP - Conjunto de Regras obtido a partir de um Conjunto de dados Particionado

RES - *Reducing the Exploration Space*

RF - Medida distinção (*Repetition Frequency*)

RHS - *Right Hand Side*

RI - Recuperação de Informação

SD - *Structuring the Domain*

sim - Similaridade

sup - Suporte

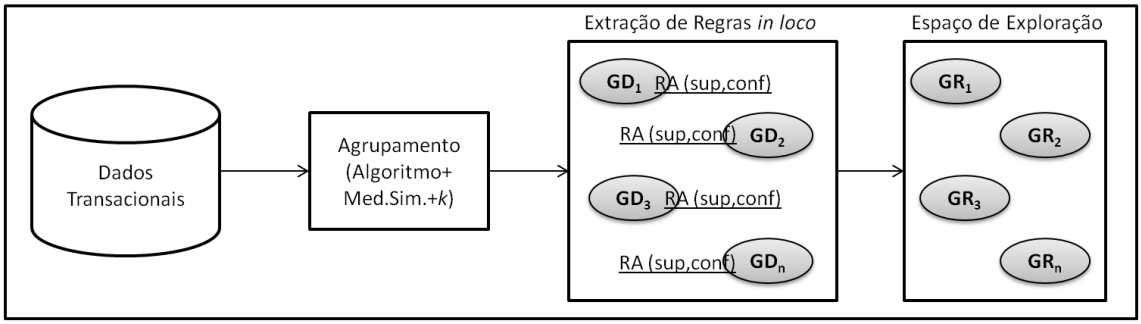
sup-min - Suporte Mínimo

1 Introdução

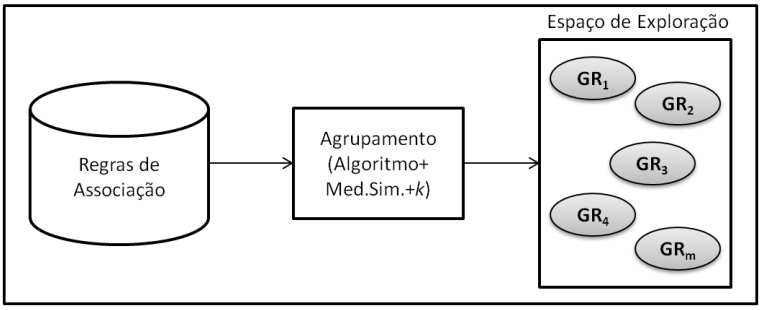
Devido a grande quantidade de regras que se obtém em um processo de mineração de associação, três linhas vem sendo adotadas para solucionar o problema: (i) pré-processar os dados antes de se extrair as regras; (ii) considerar restrições a serem atendidas durante o processo de extração; (iii) explorar as regras, após a sua obtenção, considerando os interesses do domínio. Cada uma dessas três linhas atuam em uma das etapas do processo de mineração de dados, a saber: pré-processamento (item (i)), extração de padrões (item (ii)) e pós-processamento (item (iii)). Entretanto, todas visam o mesmo objetivo: encontrar os padrões associativos relevantes da aplicação. Dentro de cada uma dessas linhas encontram-se várias abordagens.

Uma abordagem bastante utilizada tanto no pré-processamento como no pós-processamento é o agrupamento (*clustering*). O processo de agrupamento nas referidas etapas é apresentado na Figura 1. É importante mencionar que durante todo o relatório a etapa de pós-processamento será abordada antes da de pré-processamento por ser a mais usual e ter sido primeiramente por nós explorada. Como observa-se na Figura 1(b), no pós-processamento o conjunto de regras de associação, obtido *a priori*, é agrupado e as regras divididas em m grupos (GR_1,GR_2,GR_3,GR_4,...,GR_m) (destacados em cinza). No caso da Figura 1(a), referente ao pré-processamento, os dados são inicialmente agrupados em n grupos (GD_1,GD_2,GD_3,...,GD_n). A partir desse agrupamento inicial é que as regras de associação são então extraídas dentro de cada grupo, obtendo-se, assim como no pós-processamento, n grupos de regras (GR_1,GR_2,GR_3,...,GR_n) (destacados em cinza). Contudo, a interpretação do espaço de exploração é diferente em cada uma das etapas. No pós-processamento cada grupo contém um subconjunto das regras existentes no conjunto a partir do qual o particionamento foi realizado. Já no pré-processamento cada grupo contém um conjunto próprio de regras, ou seja, vários conjuntos de regras existem e, portanto, não há efetivamente um agrupamento. Entretanto, a fim de possibilitar a comparação das duas abordagens, este trabalho considera que o espaço de exploração no pré-processamento pode ser visto como um agrupamento, no qual as regras extraídas em todos os grupos compõem o conjunto a partir do qual o mesmo seria obtido. Desse modo, cada grupo, assim como no pós-processamento, conteria um subconjunto das regras totais.

Em geral, a intenção de se aplicar agrupamento, independente da etapa, é organizar as regras em grupos que contenham, de algum modo, conhecimento similar. No caso do pós-processamento, o objetivo é melhorar a apresentação dos padrões obtidos a fim de fornecer ao usuário uma visão do domínio a ser explorado e, consequente, facilitar sua busca por padrões interessantes [Reynolds et al., 2006, Sahar, 2002]. No caso do pré-processamento, o objetivo é obter regras potencialmente interessantes que não seriam



(a) Pré-processamento



(b) Pós-processamento

Figura 1: Visão geral do processo de agrupamento de regras de associação no pré-processamento e no pós-processamento.

extraídas a partir de conjuntos não particionado por não terem suporte suficiente sem, contudo, sobrecarregar o usuário com uma grande quantidade de padrões. Isso porque, para que esses mesmos padrões sejam descobertos a partir de conjuntos não particionado deve-se configurar o suporte mínimo com um valor muito baixo e abaixo do utilizado nos conjuntos particionado, fazendo com que o número de regras cresça rapidamente.

Uma vez que as duas abordagens existem, surge a pergunta: o que vale mais a pena? Entretanto, para que essa questão seja respondida, é necessário que inicialmente se investigue cada uma separadamente a fim de verificar se elas por si só ajudam ou não o usuário. Identificadas as vantagens e desvantagens de cada uma é possível comparar as duas e chegar a resposta final. Como já mencionado, cada abordagem possui suas particularidades. No caso do pós-processamento, o agrupamento se torna útil ao usuário se ele o apóia durante o processo de exploração. Esse apoio pode ser obtido se: (i) de todos os grupos apresentados destacaram-se aqueles que contêm os padrões potencialmente mais interessantes; (ii) os grupos contiverem bons descritores que forneçam uma visão dos tópicos a serem explorados. Nesse caso, por meio de (i) o usuário não necessitaria explorar todo o espaço de busca, uma vez que apenas um subconjunto dos grupos seria examinado. Além disso, por meio de (ii) o usuário poderia direcionar sua busca em função dos descritores associados aos grupos, útil quando o usuário não tem, *a priori*,

ideia por onde começar. Em relação ao pré-processamento, considera-se como válidos os mesmos argumentos do pós-processamento. Entretanto, existem outras particularidades a serem investigadas como se as regras identificadas no conjunto particionado realmente representam padrões mais interessantes e específicos do domínio.

Considerando o exposto, este relatório disponibiliza na Seção 2 os conceitos necessários ao entendimento do conteúdo aqui apresentado. Em seguida, apresenta na Seção 3 uma proposta para avaliar a contribuição do agrupamento no pós-processamento e na Seção 4 uma proposta para avaliar a contribuição do agrupamento no pré-processamento. Com base nos resultados obtidos nessas respectivas seções, a Seção 5 apresenta uma comparação entre as duas abordagens e tenta responder: o que vale mais a pena?

2 Conceitos

Visando uma melhor compreensão deste trabalho, essa seção apresenta uma breve síntese dos tópicos contidos neste relatório técnico (Seções 2.1 e 2.2), assim como uma descrição dos trabalhos relacionados (Seção 2.3).

2.1 Regras de Associação

A ideia de minerar regras de associação surgiu da análise de dados de cestas de compras em que regras do tipo "um cliente que compra os produtos x_1,x_2,...,x_n também irá comprar o produto y com probabilidade c%" são geradas. Entretanto, as regras de associação não estão restritas a análises de dependência no contexto de aplicações de varejo, uma vez que elas são aplicadas com sucesso a uma ampla gama de problemas [Semenova et al., 2001, Aggelis, 2004, Rajasekar and Weng, 2009, Changguo et al., 2009].

Uma regra de associação é definida da maneira descrita a seguir (definição baseada nos trabalhos de [Agrawal and Srikant, 1994, Yang, 2005] e adaptada para a notação utilizada neste trabalho): Seja D uma base de dados composta por um conjunto de itens $A = \{a_1, ..., a_m\}$, ordenados lexicograficamente, e por um conjunto de transações $T = \{t_1, ..., t_n\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens (chamado *itemset*) tal que $t_i \subseteq A$. É dito que uma transação t_i suporta um item x se $x \in t_i$. É dito que uma transação t_i suporta um *itemset* X se $X \subseteq t_i$. O ***suporte*** $P(X)$ de um *itemset* X representa a probabilidade da ocorrência do evento X. A regra de associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \varnothing$, LHS representa o lado esquerdo (*Left Hand Side*) da gera e RHS o lado direito (*Right Hand Side*) da regra. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com ***confiança*** $conf$ e ***suporte*** sup, onde $P(LHS\ RHS)$ representa o ***suporte*** da regra (a probabilidade da ocorrência da transação $LHS \cup RHS$) e $P(RHS|LHS)$ a ***confiança*** da regra (a probabilidade condicional de RHS dado LHS).

Existem atualmente diversos algoritmos que geram regras de associação, dentre os quais se destaca o *Apriori* [Agrawal and Srikant, 1994]. Apesar da diversidade de algoritmos, todos geram, teoricamente, as mesmas regras [Zheng et al., 2001].

2.2 Agrupamento

O objetivo da tarefa de agrupamento (*clustering*) é particionar objetos de dados (padrões, entidades, instâncias) em *clusters* (grupos, subconjuntos, categorias) [Xu and Wunsch, 2008]. Um *cluster* é um agregado de pontos em um espaço d-dimensional de atributos de modo que a distância entre quaisquer dois pontos em um *cluster* é menor

que a distância entre qualquer ponto no *cluster* e qualquer ponto não pertence ao *cluster* [Xu and Wunsch, 2008]. Segundo essa definição, um *cluster* é descrito em termos de homogeneidade interna e separação externa [Jain and Dubes, 1988, P. Hansen, 1997], isto é, objetos de dados em um mesmo *cluster* devem ser similares uns aos outros, enquanto objetos de dados em *clusters* diferentes devem ser dissimilares uns dos outros. O processo de agrupamento pode ser dividido em diversas etapas, entre elas o "projeto e/ou seleção do algoritmo de agrupamento" [Xu and Wunsch, 2008]. Nessa etapa determina-se uma medida de proximidade (similaridade ou dissimilaridade) adequada, uma vez que os objetos são agrupados em *clusters* considerando suas semelhanças. Quase todos os algoritmos de agrupamento são explicitamente ou implicitamente ligados a uma definição particular de medida de proximidade. Uma vez que a medida de proximidade é determinada o agrupamento pode ser construído. A medida de proximidade comumente utilizada para atributos contínuos é a distância Euclidiana e para atributos discretos o Coeficiente de Casamento Simples e/ou a Jaccard. Mais medidas podem ser encontradas em [Xu and Wunsch, 2008].

Existem muitos algoritmos de agrupamento na literatura, os quais podem ser classificados de diferentes formas, sendo a classificação mais usual em Hierárquico ou Particional [Tan et al., 2006]. Um agrupamento particional é a divisão do conjunto de dados em subconjuntos (*clusters*) não sobrepostos tal que um determinado dado esteja em apenas um *cluster*. Se forem adicionados *subclusters* aos *clusters*, obtém-se então um agrupamento hierárquico. Nesse caso, os *clusters* são aninhados e organizados como uma árvore (conhecida como dendograma). Cada nó (*cluster*) na árvore (exceto as folhas) representa a união de seus filhos (*subclusters*) e a raiz representa um *cluster* contendo todos os dados. Normalmente, mas não sempre, as folhas de uma árvore são *clusters* contendo um único dado. Assim, um agrupamento particional pode ser obtido através de um agrupamento hierárquico cortando-se a árvore em um determinado nível. O algoritmo particional mais tradicional é o *K-means*. Dentre os hierárquicos destacam-se o *Single Linkage*, o *Complete Linkage*, o *Average Linkage* e o *Ward Linkage* [Tan et al., 2006].

2.3 Trabalhos Relacionados

Como mencionado anteriormente, o agrupamento pode ser utilizado em etapas distintas do processo de mineração de dados a fim de facilitar a exploração de conjuntos de regras de associação. Desse modo, são apresentados a seguir os trabalhos revisados e relacionados aos objetivos propostos em função da etapa que o agrupamento é empregado.

2.3.1 Pós-processamento

Os trabalhos que empregam o agrupamento nessa etapa tem como objetivo melhorar a apresentação dos padrões, obtidos *a priori*, a fim de fornecer ao usuário uma visão do domínio a ser explorado e, consequente, facilitar sua busca por padrões interessantes [Reynolds et al., 2006, Sahar, 2002]. Nesse contexto, visando a estruturação do conhecimento, diferentes estratégias de agrupamento vêm sendo utilizadas para pós-processar regras de associação.

[Reynolds et al., 2006] propõem agrupar regras de classificação parcial obtidas por dois algoritmos por eles propostos. Nesse caso, todas as regras geradas contêm o mesmo consequente, ou seja, o agrupamento se dá em relação ao antecedente das regras. Embora o tipo de regra por eles considerada não seja de associação, a ideia do trabalho é a mesma dos trabalhos relacionados a associação: a única diferença é que todas as regras contêm o mesmo consequente. Para realizar o agrupamento os autores propõem a utilização de algoritmos particionais (K-means, PAM, CLARANS) e hierárquicos (AGNES) usando a Jaccard como medida de similaridade. Nesse caso, a Jaccard entre duas regras r_1 e r_2, apresentada na Equação 1, é calculada considerando as transações t em comum que as regras cobrem. Essa medida de similaridade é referenciada aqui como J-RT (Jaccard with Rules by Transactions). Uma regra cobre uma transação t se todos os itens da regra estão contidos em t.

$$\text{J-RT}(r_1,r_2)=\frac{|\{t \textit{ cobertas por } r_1\} \cap \{t \textit{ cobertas por } r_2\}|}{|\{t \textit{ cobertas por } r_1\} \cup \{t \textit{ cobertas por } r_2\}|} \tag{1}$$

[Jorge, 2004] propõe agrupar as regras de associação por meio da utilização de algoritmos hierárquicos (Single Linkage, Complete Linkage, Average Linkage) usando a Jaccard como medida de similaridade. Nesse caso, a Jaccard entre duas regras r_1 e r_2, apresentada na Equação 2, é calculada considerando os itens que as regras compartilham. Essa medida de similaridade é referenciada aqui como J-RI (Jaccard with Rules by Items).

$$\text{J-RI}(r_1,r_2)=\frac{|\{\textit{itens em } r_1\} \cap \{\textit{itens em } r_2\}|}{|\{\textit{itens em } r_1\} \cup \{\textit{itens em } r_2\}|} \tag{2}$$

[Toivonen et al., 1995] propõem uma medida de similaridade baseada em transação e utiliza um algoritmo baseado em densidade para realizar o agrupamento das regras. Nesse trabalho os autores consideram que todas as regras possuem o mesmo consequente, ou seja, assim como em [Reynolds et al., 2006] o agrupamento se dá em relação ao antecedente das regras. [Sahar, 2002] também propõe um medida de similaridade baseada em transação com base na proposta de [Toivonen et al., 1995], embora utilize um algoritmo hierárquico para agrupar as regras. Entretanto, não é mencionado qual o algoritmo uti-

lizado e, diferentemente de [Toivonen et al., 1995], estende-se o agrupamento para regras que possuem consequentes distintos.

Considerando os objetivos da Seção 1, o agrupamento se torna útil ao usuário se ele o apoia durante o processo de exploração, apoio esse a ser obtido por meio do direcionamento aos padrões potencialmente interessantes e/ou por meio de bons descritores que forneçam uma visão dos tópicos a serem explorados. Entretanto, todos os trabalhos acima descritos estão preocupados somente com a organização do domínio e não apoiam todo o processo. Em geral, cada trabalho usa apenas uma família de algoritmos juntamente com uma única medida de similaridade para realizar o agrupamento, assim como um único método de rotulação para descrever os resultados extraídos. [Reynolds et al., 2006] e [Jorge, 2004] selecionam como descritores, em cada grupo, os itens contidos na regra mais similar a todas as outras regras do grupo (o medóide do grupo). [Toivonen et al., 1995] não mencionam como os descritores são encontrados, mas fornecem indícios que indicam que os descritores selecionados representam os itens mais frequentes do grupo. Por outro lado, [Sahar, 2002] propõe uma abordagem para sumarizar cada grupo de regras encontrando, para tanto, padrões $a \Rightarrow c$ que cubram todas as regras contidas no grupo; nesse caso, a e c são itens do domínio e um padrão $a \Rightarrow c$ cobre uma regra $A \Rightarrow C$ se $a \in A$ e $c \in C$. Como observa-se, embora a abordagem não seja usada para definir descritores, a ideia pode ser utilizada para essa finalidade (vide Seção 3.2).

Embora muitos métodos tenham sido propostos para determinar descritores em agrupamentos de documentos nas tarefas de Mineração de Texto (MT) e Recuperação de Informação (RI), como em [Moura and Rezende, 2010, Lopes et al., 2007, Kashyap et al., 2005, Fung et al., 2003, Glover et al., 2002, Popescul and Ungar, 2000, Larsen and Aone, 1999, Cutting et al., 1992], os trabalhos relacionados a agrupamentos de regras de associação não vêm explorando essa questão, nem mesmo a avaliação dos métodos existentes. Entretanto, muitos dos métodos aplicados a documentos são similares aos utilizados em agrupamentos de regras de associação (vide Seção 3.2).

2.3.2 Pré-processamento

Outra linha que alguns pesquisadores propõem para descobrir e/ou facilitar a busca por padrões interessantes é utilizar o agrupamento na etapa de pré-processamento. Nesse caso, têm-se como objetivo a obtenção de regras interessantes que não seriam extraídas a partir de conjuntos não particionado por não terem suporte suficiente sem, contudo, sobrecarregar o usuário com uma grande quantidade de padrões. A ideia é realizar inicialmente um agrupamento do conjunto de dados e, a partir de cada grupo gerado, obter as regras de associação. [Koh and Pears, 2008] afirmam que desse modo cada grupo de re-

gras expressa suas próprias associações sem a interferência dos outros grupos que contêm diferentes padrões de relacionamentos. Existem diversos trabalhos nessa linha, os quais são descritos a seguir.

[Plasse et al., 2007] propõem segmentar os itens (variáveis/atributos) contidos no conjunto de dados de modo a construir grupos homogêneos de itens e então minerar as regras dentro de cada grupo. O objetivo é encontrar associações relevantes entre itens raros (menos frequentes) que não seriam descobertas se o processo fosse realizado considerando todo o conjunto de dados, principalmente em domínios com dados esparsos. Por outro lado, os autores demonstram que se o conjunto é agrupado é possível reduzir tanto o número de regras geradas quanto a complexidade das mesmas em relação ao conjunto não agrupado. Contudo, esses resultados são dependentes dos algoritmos de agrupamento e das medidas de similaridade utilizadas. Os autores utilizam vários algoritmos hierárquicos (Single Linkage, Complete Linkage, Average Linkage, Ward Linkage, Varclus) com diversas medidas de similaridade (Jaccard, Russel Rao, Dice, Ochiai, Pearson) para apresentar o estudo. Entretanto, não fica claro como as regras são extraídas dos grupos, já que é necessário que se tenha um conjunto de transações, ou seja, não explicita-se como as transações são distribuídas ao longo dos clusters, uma vez que o agrupamento é realizado sob os itens.

Dentre as medidas utilizadas em [Plasse et al., 2007] destacam-se, para este trabalho, a Jaccard e a Russel Rao, apresentadas, respectivamente, nas Equações 3 e 4 – a Russel Rao por ter apresentado um bom desempenho nos experimentos realizados por [Plasse et al., 2007] e a Jaccard por ser utilizada pela medida descrita a seguir, Agg. A medida Jaccard entre dois itens i_1 e i_2, referenciada aqui como P-J (Plasse Jaccard), é computada considerando as transações t em comum que os itens cobrem. Um item cobre uma transação t se o item estiver contido em t. Já a medida Russel Rao entre dois itens i_1 e i_2, referenciada aqui como P-RR (Plasse Russel Rao), é calculada considerando as transações t que os itens cobrem simultaneamente. Em outras palavras, essa medida equivale ao suporte de $i_1 \cap i_2$. Na Equação 4, N_t representa o número total de transações.

$$\text{P-J}(i_1,i_2)=\frac{|\{t \textit{ cobertas por } i_1\} \cap \{t \textit{ cobertas por } i_2\}|}{|\{t \textit{ cobertas por } i_1\} \cup \{t \textit{ cobertas por } i_2\}|} \tag{3}$$

$$\text{P-RR}(i_1,i_2)=\frac{|\{t \textit{ cobertas por } i_1\} \cap \{t \textit{ cobertas por } i_2\}|}{N_t} \tag{4}$$

[Aggarwal et al., 2002] propõem um algoritmo para segmentar o conjunto de dados com o objetivo de descobrir associações localizadas em pequenos segmentos (subconjuntos) dos dados. Os autores justificam que essa abordagem tem um impacto considerável na extração de regras, uma que vez que padrões que não podem ser identificados no conjunto como um todo podem ser geralmente descobertos em segmentos individuais. De certo

modo os autores têm por objetivo o mesmo proposto em [Plasse et al., 2007], o de não perder associações menos frequentes quando todo o conjunto de dados é considerado.

Diferente da proposta de [Plasse et al., 2007], [Aggarwal et al., 2002] agrupam transações usando uma medida de similaridade por eles proposta, apresentada na Equação 5. Essa medida de similaridade é referenciada aqui como Agg (*Agg*arwal). Nesse caso, a similaridade entre duas transações t_1 e t_2 é calculada por meio da média das afinidades (Af) existentes entre os itens contidos nas transações. Observa-se que a afinidade (Af) equivale a Jaccard entre itens, ou seja, a medida P-J. Portanto, após computar-se a P-J entre os itens m contidos em t_1 e os itens n contidos em t_2, calcula-se então a média entre os valores obtidos. Nota-se que quanto mais afinidade os itens entre as transações tiverem mais similares serão as transações. A partir desse agrupamento é que as regras são então extraídas. Além disso, diferente de [Plasse et al., 2007], o número de regras obtidas por meio do processo de segmentação é maior do que quando comparado com o conjunto não segmentado, porém menor ao número de regras do conjunto não segmentado caso o suporte mínimo fosse muito baixo para contemplar as mesmas associações. Uma limitação da abordagem é a quantidade de parâmetros que precisa ser configurada para executar o algoritmo (cinco parâmetros mais duas constantes).

$$\text{Agg}(t_1,t_2)=\frac{\sum_{p=1}^{m}\sum_{q=1}^{n} Af(i_p,j_q)}{m*n},\ em\ que \tag{5}$$

$$\text{Af(i,j)}=\frac{sup(\{i,j\})}{sup(\{i\})+sup(\{j\})-sup(\{i,j\})}$$

[Koh and Pears, 2008] propõem uma abordagem para realizar o agrupamento de transações e então minerar um conjunto de regras de associação raras. O algoritmo por eles proposto, denominado *Apriori Inverse*, é aplicado sobre os grupos de transações a fim de verificar se é possível extrair regras raras antes não descobertas considerando todo o conjunto de dados. A ideia é similar a dos trabalhos acima descritos, de que subconjuntos de regras que não apareceriam no todo tendem a aparecer em segmentos individuais. Ao final do estudo eles demonstram que o agrupamento contribui significativamente para a descoberta de novas associações. Para realizar o agrupamento, a abordagem por eles proposta encontra inicialmente k sementes (centróides), onde k refere-se a quantidade de *itemsets* frequentes contidos nas transações que atendem determinadas condições. Cada semente compõe um grupo. Após a geração das sementes, cada transação é alocada ao grupo (semente) mais similar considerando a medida apresentada na Equação 6. Nota-se que a equação é uma adaptação do coeficiente Jaccard e, portanto, compara-se a quantidade de itens em comum entre a transação (t) e o centróide do grupo (c_k). Nesse caso, quanto maior for a sobreposição de t em relação a semente (composta por um

conjunto de itens) maior é o valor da similaridade (Sim=similaridade). Nessa abordagem o usuário necessita configurar dois parâmetros adicionais além do suporte mínimo, o que dificulta um pouco a exploração.

$$Sim(t, c_k) = \frac{|t \cap c_k|}{|t \cup c_k| - |t \cap c_k| + 1} \tag{6}$$

Existem outros trabalhos que abordam o agrupamento de transações, embora não relacionados a extração de regras de associação, que poderiam ser utilizados para tanto. Em [Wang et al., 1999] a similaridade entre as transações é medida em termos da quantidade de *itemsets* frequentes que elas compartilham. Nesse caso, o agrupamento é realizado por meio de uma abordagem por eles proposta que funciona do seguinte modo: para cada transação lida verifica-se o custo de adicionar a transação à um dos grupos existentes ou à um novo grupo. O custo é determinado em função da quantidade de *itemsets* frequentes e não frequentes que o grupo irá conter com a adição da transação. O grupo que apresentar o menor custo é o selecionado para receber a nova transação. Assim, ao adicionar-se uma transação a um grupo a quantidade de *itemsets* frequentes deve aumentar e a de infrequentes diminuir. Uma limitação da abordagem é que ela considera somente *itemsets* frequentes de tamanho um para calcular o custo. Trabalhos similares ao de [Wang et al., 1999] são encontrados em [Yun et al., 2001, Yang and Padmanabhan, 2005]. Já em [Wang and Karypis, 2004] os autores propõem uma abordagem que encontra para cada transação o *itemset* frequente de maior tamanho que a representa. Em seguida, as transações que apresentam o mesmo *itemset* são agrupadas.

Em [D'Enza et al., 2005] os autores propõem uma estratégia para identificar, *a priori*, os itens potencialmente interessantes para comporem os antecedentes e os consequentes das regras de associação sem efetivamente gerá-las. Para tanto, a abordagem divide-se em dois passos: o agrupamento das transações e a seleção dos itens de interesse. Para realizar o agrupamento os autores propõem o uso do K-means incremental e como medida de similaridade a Jaccard entre transações, apresentada na Equação 7. Essa medida de similaridade é referenciada aqui como Denza. Nesse caso, a similaridade entre duas transações t_1 e t_2 é calculada considerando os itens que as transações compartilham. Após o agrupamento, estatísticas são aplicadas sobre os grupos para identificar os itens relevantes à aplicação.

$$\text{Denza}(t_1,t_2)=\frac{|\{itens\ em\ t_1\} \cap \{itens\ em\ t_2\}|}{|\{itens\ em\ t_1\} \cup \{itens\ em\ t_2\}|} \tag{7}$$

Por fim, diferentemente do pós-processamento, nenhum dos trabalhos acima descritos cobrem o aspecto relacionado aos métodos de rotulação.

3 Agrupamento de Regras de Associação no Pós-processamento

Como mencionado na Seção 1, o objetivo do agrupamento na etapa de pós-processamento é melhorar a apresentação dos padrões obtidos, a fim de fornecer ao usuário uma visão do domínio a ser explorado e, consequente, facilitar sua busca por padrões interessantes [Reynolds et al., 2006, Sahar, 2002]. Desse modo, o agrupamento se torna útil ao usuário se: (R-i) de todos os grupos apresentados destacarem-se aqueles que contêm os padrões potencialmente mais interessantes, reduzindo assim o espaço de exploração; (R-ii) os grupos contiverem bons descritores que forneçam uma visão dos tópicos a serem explorados.

Nesse contexto, essa seção apresenta duas metodologias para atender os requisitos (R-i) e (R-ii) e, consequentemente, avaliar a contribuição do agrupamento na etapa de pós-processamento. A primeira, descrita na Seção 3.1, visa reduzir o espaço de exploração direcionando o usuário para o que é potencialmente interessante, atendendo o requisito (R-i). A segunda, descrita na Seção 3.2, visa avaliar métodos de rotulação, atendendo o requisito (R-ii). Desse modo, as metodologias avaliam o agrupamento sob dois aspectos: "Redução via Interesse" e "Descritivo/Representativo".

3.1 Redução via Interesse

Para que o agrupamento de regras seja útil é essencial que o usuário seja direcionado para um espaço de exploração reduzido, o qual contenha o conhecimento interessante do domínio a ser explorado. Entretanto, como observa-se nos trabalhos da Seção 2.3.1, não existe uma metodologia voltada para esse fim. Assim, essa seção apresenta uma metodologia que atenda ao requisito (R-i) apresentado anteriormente.

Muitas abordagens de pós-processamento têm sido propostas para tratar o problema da grande quantidade de regras de associação que são obtidas nos processos de extração. Essas abordagens tentam apoiar o usuário [Baesens et al., 2000, Jorge, 2004, Natarajan and Shekar, 2005, Zhao et al., 2009]: (i) reduzindo o espaço de exploração (**RES**: *R*educing the *E*xploration *S*pace), como no caso das abordagens de consulta (*Querying*), poda (*Pruning*) e sumarização (*Summarizing*); (ii) direcionando o usuário para o que é potencialmente interessante (**DUPI**: *D*irecting the *U*ser to what is *P*otentially *I*nteresting), como no caso das medidas de avaliação (*Evaluation Measures*); (iii) estruturando o domínio (**SD**: *S*tructuring the *D*omain), como no caso do agrupamento.

Uma das abordagens mais utilizadas para estimar a interessabilidade de uma regra é a aplicação de medidas de avaliação, as quais são usualmente classificadas como objetivas

ou subjetivas. Medidas objetivas são mais gerais e independentes do domínio no qual o processo de mineração é realizado e, portanto, são mais usadas. Nessa abordagem, as regras são ranqueadas de acordo com uma dada medida e uma lista ordenada do conhecimento potencialmente interessante é apresentada ao usuário. Embora essa abordagem **DUPI** identifique o conhecimento potencialmente interessante do domínio, ela não reduz nem organiza a coleção de regras, tornando o entendimento do domínio uma tarefa difícil.

Por outro lado, o agrupamento é uma abordagem relevante relacionada a **SD**, uma vez que organiza as regras em grupos que contêm, de algum modo, conhecimento similar. Esses grupos melhoram a apresentação dos padrões extraídos, fornecendo ao usuário uma visão do domínio a ser explorado [Reynolds et al., 2006, Sahar, 2002]. Entretanto, essa abordagem não reduz o espaço de exploração nem direciona o usuário para o conhecimento interessante, dificultando sua busca por padrões relevantes. O agrupamento pode ser realizado: (i) considerando-se um critério do usuário; (ii) por meio do uso de um algoritmo de agrupamento. Enquanto que em (i) o usuário descreve como os grupos serão formados, em (ii) o usuário "deixa as regras falarem por si mesmas" [Natarajan and Shekar, 2005].

Por fim, em relação as abordagens **RES**, vale ressaltar que as mesmas também apresentam algumas deficiências: nas "abordagens de consulta" deve-se ter uma ideia dos tipos de relacionamentos que se tem interesse em explorar; nas de "sumarização" deve-se fornecer algum conhecimento extra para que o processo ocorra, como uma taxonomia; nas de "poda" conhecimentos interessantes podem ser perdidos.

Considerando os argumentos expostos, propôs-se em [Carvalho et al., 2011, Carvalho et al., 2012] uma metodologia, denominada PAR-COM (*P*ost-processing *A*ssociation *R*ules with *C*lustering and *O*bjective *M*easures), que visa reduzir o espaço de exploração de regras de associação direcionando o usuário para o que é potencialmente interessante. Para tanto, a metodologia combina agrupamento (**SD**) e medidas objetivas (**DUPI**), abordando, consequentemente, tanto os conceitos das abordagens **RES** como das **DUPI**. A metodologia foi desenvolvida sob duas hipóteses:

- Existe um subconjunto de grupos que contém todas as h-top regras interessantes, de tal modo que um pequeno número de grupos deve ser explorado. As h-top regras interessantes são as h regras que contêm os maiores valores referentes a uma medida objetiva, onde h é um número a ser escolhido.

- Se algumas regras dentro de um grupo expressam conhecimento interessante, então as outras regras dentro do mesmo grupo também tendem a expressar conhecimento interessante. Essa suposição baseia-se no conceito de agrupamento: uma coleção de objetos similares uns aos outros. Desse modo, se as regras são similares em relação a uma medida de similaridade, uma regra interessante dentro de um grupo tende a

indicar que suas regras similares também podem ser potencialmente interessantes.

Nesse contexto, a PAR-COM consegue reduzir o espaço de exploração direcionando o usuário para os grupos idealmente interessantes. Como consequência, a PAR-COM permite a descoberta adicional de conhecimento interessante que venha a existir dentro desses grupos. A metodologia PAR-COM, apresentada na Figura 2, é descrita como segue:

Passo A: o valor de uma medida objetiva é computado para todas as regras do conjunto de regras de associação.

Passo B: as h-top regras são selecionadas considerando os valores computados.

Passo C: após selecionar um algoritmo de agrupamento e uma medida de similaridade, o conjunto de regras é agrupado.

Passo D: uma busca é realizada para encontrar os grupos que contêm uma ou mais das h-top regras selecionadas no Passo B. Esses grupos são os que contêm o conhecimento potencialmente interessante do domínio (PIK: *P*otentially *I*nteresting *K*nowledge). Quanto mais h-top regras um grupo contiver mais interessante ele será.

Passo E: apenas os m primeiros grupos mais interessantes são apresentados ao usuário, o qual é direcionado para um espaço de exploração que contém o PIK do domínio, onde m é um valor a ser escolhido.

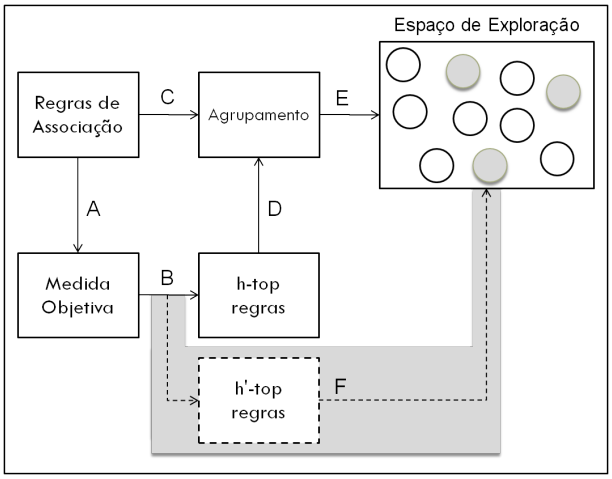


Figura 2: Metodologia PAR-COM [Carvalho et al., 2011, Carvalho et al., 2012].

O Passo F, destacado em cinza na Figura 2, é usado, nas próximas seções, para validar a metodologia. Esse passo considera todas as h'-top regras interessantes a serem selecionadas também no Passo B. As h'-top regras são as primeiras h regras que imediatamente seguem as h-top regras previamente selecionadas. Assim, o objetivo desse

passo é demonstrar que os m grupos a serem apresentados ao usuário realmente contém o PIK. Para esse propósito, uma busca é realizada para verificar se esses m grupos contêm uma ou mais das h'-top regras. É esperado que esses m grupos cubram todas as h'-top regras, uma vez que por definição um grupo é uma coleção de objetos similares. Assim, como mencionado nas hipóteses, se as regras são similares em relação a uma medida de similaridade, uma regra interessante dentro de um grupo tende a indicar que suas regras similares são também potencialmente interessantes. É importante mencionar que a metodologia não apóia a exploração como uma lista ordenada, que é o caso quando medidas objetivas são utilizadas. Por essa razão, a metodologia permite a descoberta adicional de conhecimento interessante dentro dos m grupos. Em [Carvalho et al., 2012] apresenta-se um estudo de caso em que compara-se a metodologia proposta com os resultados obtidos com uma das abordagens **DUPI**, mais especificamente medidas objetivas, uma vez que a PAR-COM combina **SD+DUPI** a fim de obter um espaço de exploração reduzido (**RES**) potencialmente interessante.

3.1.1 Experimentos

Quatro conjuntos de dados foram usados nos experimentos para validar a metodologia, os quais encontram-se descritos na Tabela 1. Esses conjuntos são utilizados em todos os experimentos apresentados neste relatório ao longo das seções. Os três primeiros encontram-se disponíveis no Projeto R para Computação Estatística[2] por meio do pacote "arules"[3]. O último foi fornecido por um supermercado localizado na cidade de São Carlos. Todas as transações em Adult e Income contêm o mesmo número de itens (referenciados como conjunto de dados padronizados (CD-P)), diferentemente do Groceries e Sup (referenciados como conjunto de dados não padronizados (CD-NP)), em que cada transação contêm uma quantidade distinta de itens. O objetivo é realizar a avaliação em diferentes tipos de dado. As regras foram extraídas usando uma implementação do *Apriori* desenvolvida por Christian Borgelt[4] com um número mínimo de 2 itens e um número máximo de 5 itens por regra. Com o conjunto Adult 6508 regras foram obtidas usando um suporte mínimo (sup-min) de 10% e uma confiança mínima (conf-min) de 50%; com o Income 3714 regras considerando sup-min=17% e conf-min=50%; com o Groceries 2050 regras considerando sup-min=0.5% e conf-min=0.5%; com o Sup 7588 regras considerando sup-min=0.7% e conf-min=0.5%. Os valores dos parâmetros foram escolhidos experimentalmente.

Uma vez que os trabalhos descritos na Seção 2.3.1 apenas usam uma família de algo-

[2]http://www.r-project.org/.
[3]http://cran.r-project.org/web/packages/arules/index.html.
[4]http://www.borgelt.net/apriori.html.

Tabela 1: Detalhes dos conjuntos de dados usados nos experimentos.

Conjunto de Dados	# de transações	# de itens distintos	Descrição
Adult	48842	115	Esse conjunto é uma versão R pré-processada para a tarefa de associação da base "Adult" disponível em [Frank and Asuncion, 2010]. Foi originalmente utilizada para predizer se a renda excedia 50K/yr considerando dados do censo.
Income	6876	50	Esse conjunto é uma versão R pré-processada para a tarefa de associação da base "Marketing" disponível em [Hastie et al., 2009]. Foi originalmente utilizada para predizer a renda familiar anual a partir de atributos demográficos.
Groceries	9835	169	Esse conjunto também encontra-se disponível no R e contém transações referentes a um mês de vendas de um armazém.
Sup	1716	1939	Esse conjunto contém transações referentes a um dia de vendas de um supermercado localizado na cidade de São Carlos, São Paulo, Brasil.

ritmos de agrupamento e uma medida de similaridade para agrupar regras de associação, decidiu-se realizar os experimentos com um algoritmo de cada família e com as duas medidas de similaridade mais usadas para regras de associação (J-RI e J-RT (Equações 2 e 1 (Seção 2.3.1))). O algoritmo PAM (*P*artitioning *A*round *M*edoids) foi o escolhido dentre os particionais e o Ward dentre os hierárquicos. No caso particional, um algoritmo baseado em medóide foi selecionado, uma vez que o objetivo é agrupar as regras mais similares em um grupo; assim, o ideal é que o elemento representativo do grupo seja uma regra e não a média como no K-means (mesmo argumento apresentando em [Reynolds et al., 2006]). No caso hierárquico, os algoritmos tradicionais foram aplicados (Single, Complete, Average e Ward) e o que obteve o melhor desempenho foi considerado. Nesse caso, para que a seleção fosse realizada, considerou-se como critério o comportamento da distribuição das regras nos grupos. Isso porque, é importante que as regras de um dado agrupamento estejam distribuídas proporcionalmente entre os grupos, uma vez que não é interessante que existam muitos grupos com poucas regras e poucos grupos com muitas regras, o que pode afetar significativamente os resultados. Em relação a k, executou-se o PAM com valores entre 5 e 100 em passos de 5. Os dendogramas gerados pelo Ward foram cortados nos mesmos valores (5 a 100 em passos de 5). Os valores foram escolhidos com base nos resultados apresentados em [Carvalho et al., 2012].

As Figuras 3 e 4 apresentam o comportamento da distribuição das regras nos agrupamentos considerados em cada um dos tipos de dado. O eixo x de cada gráfico representa o intervalo de k. O eixo y a porcentagem de grupos do agrupamento que contém, no mínimo, 0.5% do total das regras. Desse modo, quanto maior a porcentagem, melhor a organização (algoritmo+medida) distribui as regras nos grupos do agrupamento. A fim

de auxiliar a interpretação dos gráficos considere a Figura 3(a). Pode-se observar que, para $k = 5$, apenas 60% dos grupos da organização Average:J-RI (3 grupos (60% de 5 grupos)) contém, pelo menos, 0.5% do total das regras, não tendo, portanto, contribuído para um bom particionamento das regras. Em outras palavras, 2 dos 5 grupos contêm um número inexpressivo de regras, fazendo com que o agrupamento fique desbalanceado. Avaliando os gráficos observa-se que dentre os hierárquicos, o que melhor distribui é o Ward, como mencionado anteriormente. Com exceção ao conjunto Groceries na medida J-RI, o PAM apresenta comportamento semelhante ao Ward.

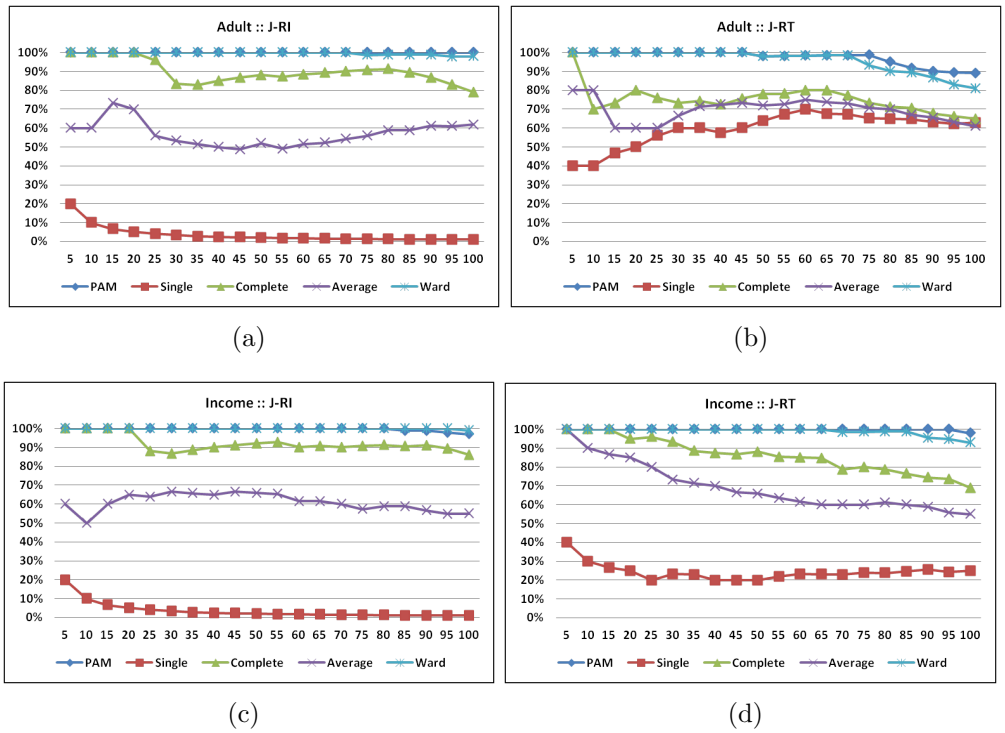


Figura 3: Distribuição das regras nos agrupamentos dos CD-P.

Para aplicar a PAR-COM foi necessário escolher uma medida objetiva (Passo A) e os valores de h (Passo B) e m (Passo E). O h foi definido para assumir 0.5% do total de regras do conjunto de modo a tornar a comparação entre os mesmos mais justa; assim, cada conjunto contém seus próprios valores que são proporcionais em todos eles. O m foi definido como no máximo 3, ou seja, se apenas um grupo contiver todas as h-top regras somente ele será selecionado; se apenas dois grupos cobrirem todas as h-top regras somente os dois serão selecionados; e assim sucessivamente até no máximo 3 grupos. Uma vez que espera-se que qualquer medida produza bons resultados, em vez de escolher uma medida específica, considerou-se a média do ranqueamento obtida por meio de 18 medidas objetivas (veja Tabela 2) como segue: (i) computa-se o valor das 18 medidas para cada regra; (ii) cada regra recebe 18 ID's, cada um correspondente a sua posição em um dos

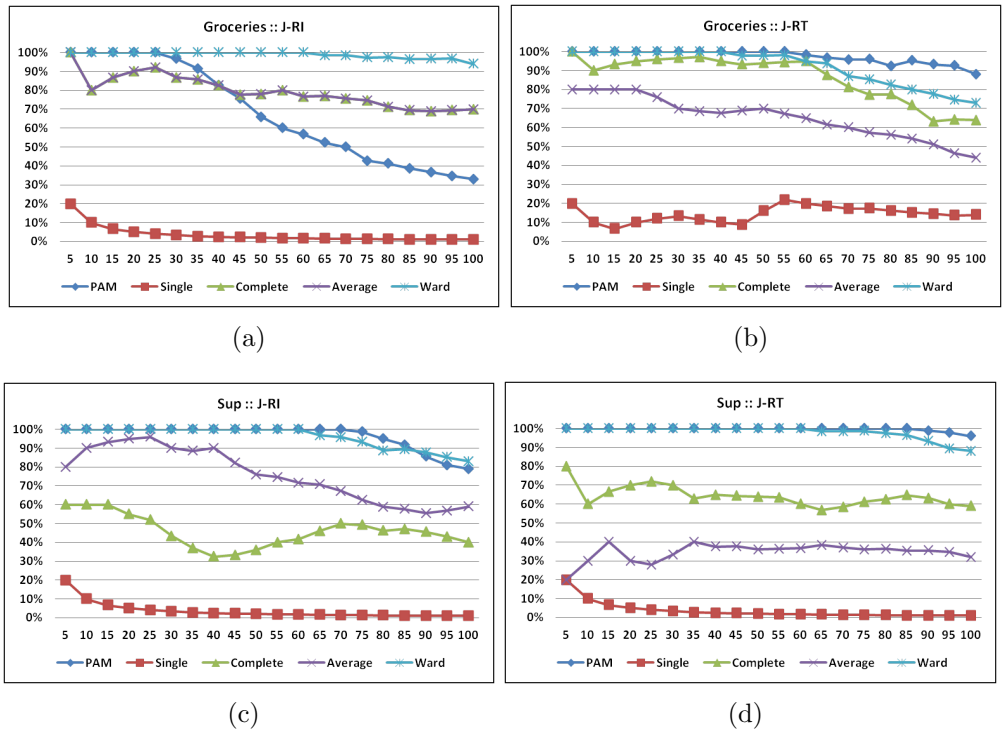


Figura 4: Distribuição das regras nos agrupamentos dos CD-NP.

ranqueamentos relacionados a uma determinada medida; (iii) a média é então calculada com base nas posições dos ranqueamentos (ID's). Detalhes sobre as medidas podem ser encontrados em [Tan et al., 2004]. Diante do apresentado, a Tabela 2 apresenta as configurações aplicadas para avaliar a PAR-COM. É importante mencionar que os valores foram escolhidos com base nos resultados apresentados em [Carvalho et al., 2012].

Tabela 2: Configurações usadas para avaliar a PAR-COM.

Conjuntos de Dados	Adult; Income; Groceries; Sup
Algoritmos	PAM; Ward Linkage
Medidas de Similaridade	J-RI; J-RT
k	5 a 100, passos de 5
h	0.5% do total de regras
m	≤ 3
Medidas Objetivas	Média dos ranqueamentos das medidas Added Value, Certainty Factor, Collective Strength, Confidence, Conviction, IS, ϕ-coefficient, Gini Index, J-Measure, Kappa, Klosgen, λ, Laplace, Lift, Mutual Information (assimétrica), Novelty, Support, Odds Ratio

3.1.2 Resultados e Discussão

Considerando as configurações apresentadas na Tabela 2, a PAR-COM foi aplicada e os resultados encontram-se apresentados nas Tabelas 3 a 10. Os resultados foram agrupados

por algoritmo para cada conjunto de dados. As Tabelas 3 e 7 apresentam os resultados para o conjunto Adult, as Tabelas 4 e 8 para o Income, as Tabelas 5 e 9 para o Groceries e as Tabelas 6 e 10 para o Sup. Cada tabela apresenta os resultados de ambas as medidas de similaridade consideradas, i.e., J-RI e J-RT. Cada coluna apresenta o resultado referente a um dado valor de k. Em relação as linhas tem-se que: (i) h-top e h'-top apresentam a porcentagem de h-top e h'-top regras contidas nos m primeiros clusters interessantes; (ii) R a porcentagem de redução do espaço de exploração; (iii) m a quantidade de clusters necessários para recuperar as h-top regras ($m \leq 3$).

De modo a auxiliar a interpretação das tabelas, considere a organização PAM:J-RI da Tabela 3. Pode-se observar, para $k = 5$, que: (i) os 2 primeiros clusters interessantes (m=2) contém 100% das 33-top regras (h=33 (0.5%*6508)); (ii) os 2 primeiros clusters interessantes (m=2) contém 100% das 33'-top-regras; assim, pode-se constatar, por meio do passo de validação (Passo F), que esses 2 clusters são idealmente os 2 subconjuntos mais interessantes; (iii) os 2 primeiros clusters interessantes cobrem 38% (100%-62%) das regras, o que leva a uma redução de 62% do espaço de exploração; em outras palavras, se o usuário explorar esses 2 clusters, ele irá explorar apenas 38% do espaço das regras.

Tabela 3: Resultados do PAM, na PAR-COM, no conjunto de dados Adult.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	97%	97%	91%	82%	82%	61%	58%	55%	55%	67%	67%	70%	70%	64%	61%	61%	61%	70%	67%
h'-top	100%	97%	100%	82%	73%	64%	61%	42%	39%	36%	39%	39%	42%	42%	36%	36%	39%	33%	45%	42%
R	62%	73%	76%	82%	85%	88%	90%	92%	92%	93%	94%	94%	94%	94%	95%	95%	95%	95%	95%	96%
m	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
h'-top	91%	91%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%
R	69%	86%	94%	94%	95%	95%	95%	95%	95%	95%	95%	95%	96%	96%	96%	96%	96%	96%	96%	96%
m	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2

Tabela 4: Resultados do PAM, na PAR-COM, no conjunto de dados Income.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	89%	74%	89%	84%	74%	63%	63%	63%	63%	63%	63%	53%	53%	53%	42%	53%	42%	47%	47%
h'-top	100%	74%	74%	74%	68%	74%	63%	47%	58%	53%	42%	42%	37%	37%	58%	58%	32%	37%	21%	26%
R	30%	70%	76%	79%	84%	87%	89%	90%	91%	91%	92%	92%	94%	95%	95%	95%	95%	96%	96%	96%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	95%	89%	89%	89%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%	79%
h'-top	89%	79%	74%	74%	68%	68%	68%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%	47%
R	37%	67%	74%	84%	88%	88%	89%	91%	91%	92%	94%	94%	95%	95%	95%	96%	96%	96%	96%	96%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 5: Resultados do PAM, na PAR-COM, no conjunto de dados Groceries.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	100%	100%	82%	91%	73%	64%	82%	82%	73%	73%	73%	73%	73%	73%	73%	64%	64%	64%	64%
h'-top	82%	64%	55%	27%	27%	18%	18%	27%	18%	27%	27%	27%	18%	18%	18%	18%	18%	18%	18%	18%
R	28%	63%	69%	71%	85%	86%	85%	87%	87%	88%	88%	88%	89%	89%	89%	90%	90%	90%	90%	90%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	100%	82%	82%	73%	73%	73%	73%	64%	64%	73%	64%	73%	73%	73%	73%	64%	64%	64%	64%	64%
h'-top	45%	27%	9%	9%	36%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%
R	40%	64%	76%	82%	86%	91%	93%	95%	95%	95%	95%	96%	96%	96%	96%	97%	97%	98%	98%	98%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 6: Resultados do PAM, na PAR-COM, no conjunto de dados Sup.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	95%	79%	74%	66%	50%	47%	50%	47%	45%	45%	39%	34%	29%	29%	32%	29%	26%	34%	32%	18%
h'-top	79%	74%	50%	45%	47%	45%	39%	39%	37%	29%	26%	24%	29%	16%	11%	11%	8%	16%	8%	5%
R	43%	61%	80%	80%	87%	88%	89%	90%	93%	93%	94%	95%	95%	96%	96%	96%	95%	93%	97%	98%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	84%	42%	45%	37%	32%	34%	37%	29%	34%	37%	32%	32%	37%	34%	32%	32%	32%	34%	29%	29%
h'-top	74%	53%	34%	37%	29%	21%	13%	29%	32%	24%	21%	24%	21%	16%	13%	16%	13%	18%	11%	11%
R	36%	69%	80%	83%	86%	89%	91%	90%	91%	92%	93%	94%	94%	95%	95%	96%	95%	96%	96%	96%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 7: Resultados do Ward, na PAR-COM, no conjunto de dados Adult.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	100%	97%	82%	76%	73%	73%	70%	70%	70%	67%	55%	55%	55%	55%	55%	55%	55%	45%	45%
h'-top	97%	85%	39%	27%	27%	21%	21%	18%	18%	18%	18%	15%	15%	15%	15%	15%	15%	15%	12%	12%
R	43%	60%	74%	83%	85%	92%	92%	93%	93%	93%	94%	96%	96%	96%	96%	96%	96%	96%	98%	98%
m	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
h'-top	100%	91%	91%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%	42%
R	77%	92%	92%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%	96%
m	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2

Tabela 8: Resultados do Ward, na PAR-COM, no conjunto de dados Income.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	79%	68%	68%	63%	63%	63%	63%	63%	63%	63%	53%	42%	42%	42%	37%	37%	37%	32%	32%
h'-top	95%	79%	53%	53%	32%	32%	47%	47%	47%	47%	47%	32%	32%	11%	11%	21%	21%	21%	16%	16%
R	23%	68%	78%	83%	86%	88%	91%	92%	92%	92%	92%	93%	95%	96%	96%	96%	96%	96%	97%	97%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	100%	95%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	79%	79%	79%	79%	79%	79%
h'-top	100%	79%	74%	74%	74%	74%	74%	74%	74%	74%	74%	74%	74%	74%	47%	47%	47%	47%	47%	47%
R	27%	66%	80%	90%	90%	90%	90%	90%	92%	92%	94%	94%	94%	94%	96%	96%	96%	97%	97%	97%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 9: Resultados do Ward, na PAR-COM, no conjunto de dados Groceries.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	100%	82%	73%	73%	73%	73%	73%	64%	55%	55%	45%	45%	45%	45%	45%	36%	36%	36%	36%	36%
h'-top	100%	100%	91%	18%	18%	18%	18%	18%	9%	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
R	16%	30%	48%	65%	71%	72%	78%	84%	87%	89%	91%	92%	93%	93%	94%	95%	95%	96%	97%	97%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	100%	100%	82%	82%	82%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	55%	55%	55%
h'-top	27%	27%	18%	18%	18%	18%	18%	9%	9%	9%	9%	9%	9%	9%	9%	9%	9%	0%	0%	0%
R	13%	19%	44%	53%	55%	63%	68%	71%	74%	78%	78%	79%	80%	80%	80%	81%	81%	91%	92%	92%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 10: Resultados do Ward, na PAR-COM, no conjunto de dados Sup.

k	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
J-RI																				
h-top	82%	68%	55%	53%	47%	45%	42%	42%	42%	42%	39%	39%	39%	37%	37%	37%	37%	29%	29%	29%
h'-top	82%	53%	39%	32%	26%	26%	24%	24%	24%	24%	16%	16%	16%	16%	16%	16%	16%	16%	16%	16%
R	9%	32%	41%	47%	55%	61%	89%	89%	89%	89%	93%	93%	93%	94%	94%	94%	94%	94%	94%	95%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
J-RT																				
h-top	87%	79%	74%	63%	53%	53%	47%	47%	47%	47%	47%	47%	47%	47%	47%	45%	45%	45%	45%	45%
h'-top	71%	61%	45%	39%	37%	37%	24%	24%	24%	24%	24%	24%	24%	24%	24%	18%	18%	16%	16%	16%
R	71%	75%	84%	89%	92%	92%	95%	95%	95%	95%	95%	96%	96%	96%	96%	97%	97%	97%	97%	97%
m	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Com base nesses dados, de modo a identificar as organizações que apresentaram os melhores resultados, uma análise baseada na média de cada critério (h, h', R), considerando os diferentes tipos de dado, independente do conjunto, foi realizada. A Tabela 11 apresenta os resultados. Cada média foi obtida a partir dos resultados dos experimentos relacionados a configuração apresentada. O valor h de 68% da configuração CD-P:PAM:J-RI, por exemplo, foi obtido por meio da média dos valores de h das configurações PAM:Adult:J-RI e PAM:Income:J-RI das Tabelas 3 e 4. Para facilitar o entedimento, esses valores encontram-se sublinhados nas respectivas tabelas. As médias mais altas, referentes a cada um dos critérios (h, h', R), encontram-se marcadas com "*" em cada algoritmo. Para a configuração CD-P:PAM, por exemplo, a média mais representativa para h é a relacionada a medida J-RT (91%); já em relação a h' J-RI (54%). Desse modo, é possível visualizar, para cada configuração tipo-de-dado+algoritmo, a medida de similaridade que apresenta o melhor desempenho. Por fim, é importante mencionar que os resultados são determinísticos e, portanto, nenhum teste estatístico precisa ser realizado para checar se existe diferença significativa entre as médias. Pode-se observar que:

- em relação ao algoritmo PAM nos CD-P, a medida de similaridade que apresenta o melhor resultado em h é a J-RT, em h' J-RI e em R J-RT. Entretanto, considerando os três critérios simultaneamente, uma vez que os valores de h' encontram-se muito próximos (54% x 52%), tem-se que a medida J-RT é a que se ajusta mais adequadamente ao PAM.

- em relação ao algoritmo Ward nos CD-P, a medida de similaridade que apresenta o melhor resultado em todos os critérios é a J-RT. Portanto, tem-se que a medida J-RT é a que se ajusta mais adequadamente ao Ward.

- em relação ao algoritmo PAM nos CD-NP, a medida de similaridade que apresenta o melhor resultado tanto em h quanto em h' é a J-RI e em R J-RT. Entretanto, considerando os três critérios simultaneamente, uma vez que os valores de R encontram-se muito próximos (85% x 89%), tem-se que a medida J-RI é a que se ajusta mais adequadamente ao PAM.

- em relação ao algoritmo Ward nos CD-NP, a medida de similaridade que apresenta o melhor resultado tanto em h quanto em R é a J-RT e em h' J-RI. Entretanto, considerando os três critérios simultaneamente, uma vez que os valores de h' encontram-se muito próximos (23% x 21%), tem-se que a medida J-RT é a que se ajusta mais adequadamente ao Ward.

Diante do exposto, pode-se notar que para os CD-P a medida que se apresenta mais adequada, independente do algoritmo, é a J-RT. Por outro lado, nos CD-NP não há um

Tabela 11: Média de h, h' e R, nos k's entre 5-100, nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida).

Tipo de Dado	Algoritmo	Medida	h	h'	R
CD-P	PAM	J-RI	68%	54%*	88%
		J-RT	91%*	52%	91%*
	Ward	J-RI	62%	32%	88%
		J-RT	93%*	59%*	91%*
CD-NP	PAM	J-RI	61%*	30%*	85%
		J-RT	54%	19%	89%*
	Ward	J-RI	50%	23%*	78%
		J-RT	61%*	21%	80%*

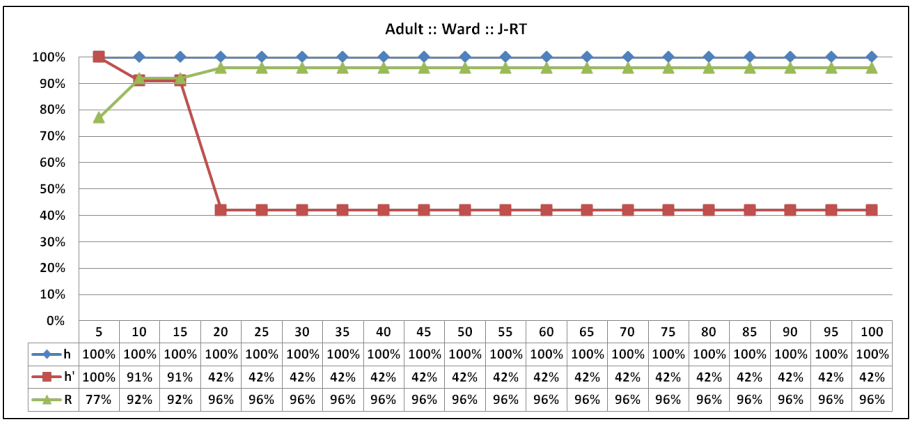
padrão, uma vez que a medida J-RI se apresenta mais adequada ao PAM e a J-RT ao Ward. Desse modo, a fim de identificar a melhor organização para cada um dos tipos de dado, as seguintes comparações foram realizadas: PAM:J-RT e Ward:J-RT nos CD-P e PAM:J-RI e Ward:J-RT nos CD-NP. Pode-se observar que (vide Tabela 11):

- nos CD-P, comparando-se PAM:J-RT com Ward:J-RT tem-se que ambas as organizações apresentam resultados muito próximos, tendo a organização Ward:J-RT apresentado um desempenho um pouco melhor. Ambas as organizações também apresentam comportamento similar em relação a distribuição das regras nos agrupamentos (Figuras 3(b) e 3(d) (página 16)).

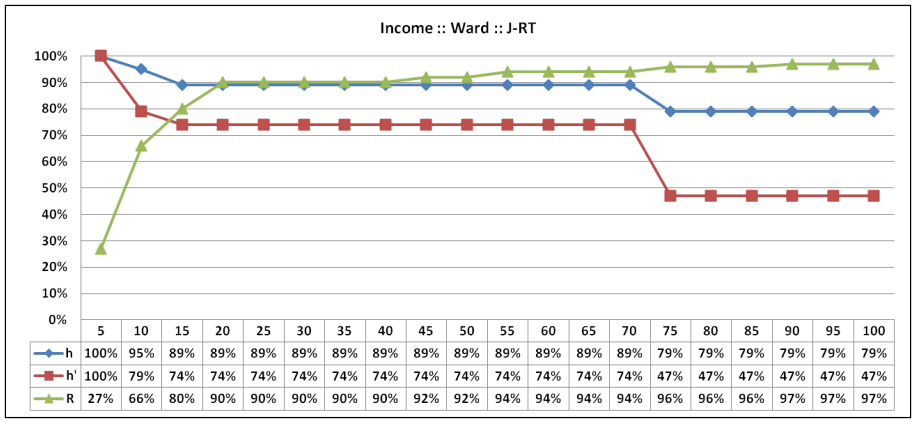
- nos CD-NP, comparando-se PAM:J-RI com Ward:J-RT tem-se que ambas as organizações apresentam resultados muito próximos, tendo a organização PAM:J-RI apresentado um desempenho um pouco melhor. Entretanto, o Ward:J-RT apresenta uma melhor distribuição das regras nos agrupamentos em relação ao PAM:J-RI (Figuras 4(a), 4(b), 4(c) e 4(d) (página 17)). Portanto, a organização Ward:J-RT se apresenta mais interessante.

Considerando o apresentado, indica-se a utilização da PAR-COM com a organização Ward:J-RT para ambos os tipos de dado, ou seja, CD-P e CD-NP. As Figuras 5 e 6 apresentam o comportamento da PAR-COM, em cada um dos conjuntos, na organização indicada. É possível observar, tanto na Tabela 11 quanto nas figuras, que o desempenho da PAR-COM nos CD-P é melhor do que nos CD-NP. Entretanto, em ambos os casos, **nota-se que a PAR-COM auxilia o usuário tanto na identificação de conhecimentos interessantes (h, h') quanto na redução do espaço de exploração (R). Portanto, por meio da PAR-COM, é possível afirmar que o agrupamento no pós-processamento é útil ao usuário**.

Por fim, considerando a organização Ward:J-RT, observa-se nas Figuras 3 e 4 (páginas 16 e 17) que: (i) em relação ao conjunto Adult (Figura 3(b)), a distribuição se mantém nos 100% até k=45; (ii) em relação ao conjunto Income (Figura 3(d)), a distribuição se



(a)



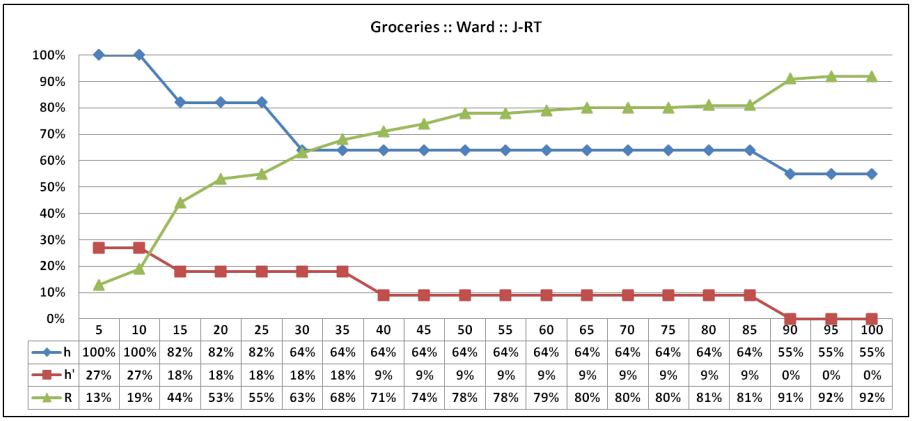
(b)

Figura 5: Comportamento da PAR-COM, na organização Ward:J-RT, nos CD-P.

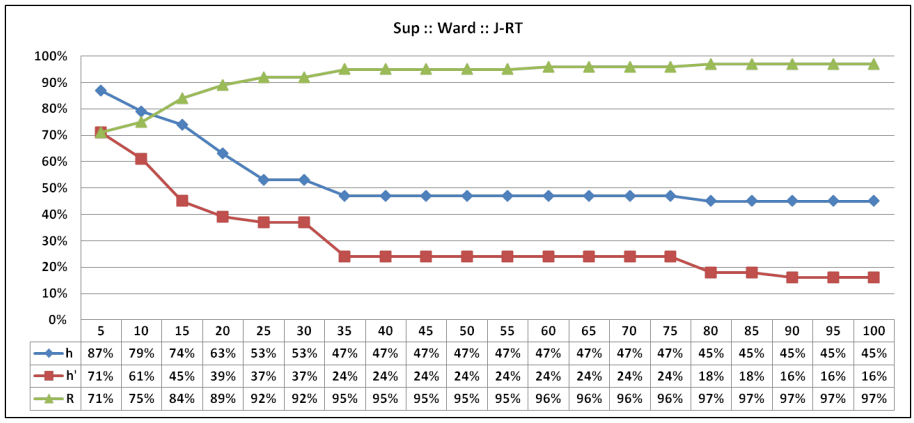
mantém nos 100% até k=65; (iii) em relação ao conjunto Groceries (Figura 4(b)), a distribuição se mantém nos 100% até k=40; (iv) em relação ao conjunto Sup (Figura 4(d)), a distribuição se mantém nos 100% até k=60. Desse modo, a fim de garantir uma comparação mais justa entre as organizações, no que se refere a distribuição das regras, optou-se, a partir dos próximos experimentos, por se utilizar um intervalo de k entre 5 e 50 em passos de 5. Além disso, como observa-se na Tabela 12, as conclusões acima obtidas se mantém as mesmas para k entre 5 e 50. A Tabela 12 foi obtida do mesmo modo que a Tabela 11, porém considerando k entre 5 e 50. A interpretação da tabela é a mesma da referente aos valores de k entre 5 e 100 (Tabela 11).

3.2 Descritivo/Representativo

Para que o agrupamento de regras seja útil ao usuário é essencial que os grupos sejam representados por descritores que forneçam uma visão dos tópicos contidos no espaço de exploração, ajudando o usuário a guiar sua busca. Encontrar bons descritores é um



(a)



(b)

Figura 6: Comportamento da PAR-COM, na organização Ward:J-RT, nos CD-NP.

Tabela 12: Média de h, h' e R, nos k's entre 5-50, nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida).

Tipo de Dado	Algoritmo	Medida	h	h'	R
CD-P	PAM	J-RI	77%	69%*	81%
		J-RT	92%*	59%	86%*
	Ward	J-RI	75%	45%	80%
		J-RT	95%*	67%*	87%*
CD-NP	PAM	J-RI	72%*	42%*	78%
		J-RT	58%	26%	81%*
	Ward	J-RI	62%	38%*	62%
		J-RT	68%*	28%	71%*

assunto relevante em muitas tarefas de mineração de Texto (MT) e Recuperação de Informação (RI) (veja algumas aplicações em [Manning et al., 2009]). É necessário, por exemplo, que bons descritores sejam apresentados aos usuários a fim de facilitar análises exploratórias, úteis quando o usuário não tem ideia por onde começar. Embora muitos métodos tenham sido propostos para rotular clusters de documentos em MT e RI, os artigos relacionados a agrupamento de regras de associação ainda não exploraram esse tópico, nem mesmo a avaliação dos métodos existentes. Desse modo, a análise de diferentes métodos de rotulação para agrupamentos de regras de associação, assim como em outras tarefas, também é relevante, uma vez que é necessário que se identifique os métodos que apresentam os melhores resultados. Por fim, a integração de bons métodos de rotulação com outras metodologias pode permitir que o agrupamento de regras de associação se torne uma poderosa ferramenta de pós-processamento. A integração com a PAR-COM, por exemplo, pode permitir a identificação dos tópicos potencialmente interessantes do domínio.

Nesse contexto, a fim de atender o requisito (R-ii), essa seção analisa alguns métodos de rotulação de modo a identificar: **(a)** os métodos mais adequados para serem utilizados em agrupamentos de regras de associação; **(b)** as organizações que fornecem os melhores resultados, uma vez que o desempenho dos métodos é afetado pelas mesmas; **(c)** como consequência de **(b)**, a organização que melhor estrutura o conhecimento. Uma vez que não existe uma metodologia para avaliar descritores em agrupamentos de regras de associação duas medidas, Precision (P) e Repetition Frequency (RF), foram propostas e usadas. O ideal é que os descritores de cada grupo representem o mais precisamente possível o conhecimento do seu próprio grupo (medida P, Seção 3.2.2) e seja o mais diferente possível dos descritores dos outros grupos (medida RF, Seção 3.2.2).

3.2.1 Métodos de Rotulação

Com o objetivo de analisar alguns métodos de rotulação (Labeling Methods (LM)) para agrupamentos de regras de associação em relação ao comportamento dos mesmos no que se refere a precisão (P) e a distinção (RF), quatro métodos foram selecionados. Esses métodos representam as idéias de muitos dos métodos previamente descritos e citados na Seção 2.3.1 (tanto para regras de associação (RA) como para documentos (MT e RI)). A fim de compreender os métodos, considere um agrupamento composto por três grupos de regras de associação: C_1={r_1: coffee $\Rightarrow$ butter; r_2: milk $\Rightarrow$ coffee; r_3: milk & butter $\Rightarrow$ coffee}; C_2={r_1: butter $\Rightarrow$ coffee; r_2: milk $\Rightarrow$ butter}; C_3={r_1: butter $\Rightarrow$ milk; r_2: coffee $\Rightarrow$ milk}. O exemplo é meramente ilustrativo. Os quatro métodos abaixo descritos são **LM-M**, **LM-T**, **LM-S** e **LM-PU**.

Em **LM-M** (*L*abeling *M*ethod *M*edoid) os descritores de cada grupo são construídos pelos itens contidos na regra do grupo que seja mais similar a todas as outras regras do grupo (o medóide do grupo). Para tanto, computa-se a similaridade acumulada (a_s) de cada regra considerando sua similaridade em relação a todas as outras regras e aquela com o maior valor é selecionada. Considerando C_1 do exemplo acima e que r_1 cobre $\{t_1, t_3, t_5, t_7\}$, r_2 $\{t_1, t_3, t_5, t_7, t_9\}$, r_3 $\{t_3, t_5, t_7\}$, as similaridades s(r_1,r_2)=s(r_2,r_1)=$\frac{4}{5}$ = 0.8, s(r_1,r_3)=s(r_3,r_1)=$\frac{3}{4}$ = 0.75, s(r_2,r_3)=s(r_3,r_2)=$\frac{3}{5}$ = 0.6, considerando a medida J-RT (Equação 1 (Seção 2.3.1, página 6)), são obtidas e as seguintes a_s encontradas: a_s(r_1)=s(r_1,r_2)+s(r_1,r_3)=1.55; a_s(r_2)=s(r_2,r_1)+s(r_2,r_3)=1.40; a_s(r_3)=s(r_3,r_1) + s(r_3, r_2) = 1.35. Assim, r_1 é selecionada e os descritores de C_1 passam a ser {coffee, butter}. As similaridades entre as regras podem ser obtidas por meio de qualquer medida de similaridade, como as apresentadas na Seção 2.3.1. A similaridade escolhida foi a J-RT, uma vez que é a mais utilizada pelos trabalhos para computar LM-M. Citam-se como trabalhos relacionados a essa ideia [Reynolds et al., 2006, Jorge, 2004] do contexto *RA* e [Kashyap et al., 2005, Larsen and Aone, 1999, Cutting et al., 1992] do contexto *MT* e *RI*.

Em **LM-T** (*L*abeling *M*ethod *T*ransaction) os descritores de cada grupo são construídos pelos itens contidos na regra do grupo que cobre o maior número de transações. Uma regra cobre uma transação t se todos os itens da regra estão contidos em t. Em outras palavras, a regra a ser selecionada em cada grupo é aquela que apresenta o maior valor de suporte. Considerando C_1 do exemplo acima e que r_1 cobre $\{t_1, t_3, t_5, t_7\}$, r_2 $\{t_3, t_5, t_7\}$, r_3 $\{t_1, t_3, t_5, t_7, t_9\}$, r_3 é selecionada e os descritores de C_1 passam a ser {milk, butter, coffee}. Cita-se como trabalho relacionado a essa ideia [Fung et al., 2003] do contexto *MT* e *RI*.

Em **LM-S** (*L*abeling *M*ethod *S*ahar devido a sua referência a [Sahar, 2002])[5] os descritores de cada grupo são construídos como segue: (i) considere um conjunto $I = \{i_1, ..., i_m\}$ contendo todos os itens distintos do grupo, um conjunto $R = \{r_1, ..., r_n\}$ contendo todos os possíveis relacionamentos $a \Rightarrow c$, onde $a, c \in I$ – cada um desses relacionamentos representa um padrão de regra; (ii) o número de regras que cada padrão $r_i \in R$ cobre é computado (N_c); um padrão $a \Rightarrow c$ cobre uma regra $A \Rightarrow C$ se $a \in A$ e $c \in C$; (iii) o padrão com a maior cobertura é selecionado; em caso de empate, todos os padrões empatados são selecionados; (iv) todos os padrões selecionados compõem o conjunto $P \subseteq R$; (v) ao final, todos os itens distintos em P passam a ser os descritores do grupo. Considerando C_1 do exemplo acima tem-se que: I={coffee, butter, milk}, R=$\{r_1 : coffee \Rightarrow butter$; $r_2 : butter \Rightarrow coffee$; $r_3 : coffee \Rightarrow milk$; $r_4 : milk \Rightarrow coffee$; $r_5 : butter \Rightarrow milk$; $r_6 : milk \Rightarrow butter\}$, N_c=$\{r_1 : 1, r_2 : 1, r_3 : 0, r_4 : 2, r_5 : 0, r_6 : 0\}$ e P=$\{r_4\}$. Assim, os descritores de C_1 passam a ser {milk, coffee}.

[5]Versão simplificada do processo descrito em [Sahar, 2002] do contexto *RA* e explicada na Seção 2.3.1.

Em **LM-PU** (*L*abeling *M*ethod *P*opescul e *U*ngar devido a sua referência a [Popescul and Ungar, 2000]) os descritores de cada grupo são construídos pelos N itens do grupo que apresentam o melhor equilíbrio entre frequência e valor preditivo. Formalmente tem-se que: $f(i_n|C_n) * \frac{f(i_n|C_n)}{f(i_n)}$. A primeira parte da equação, $f(i_n|C_n)$, mede a frequência f de cada item i_n em seu próprio grupo C_n. A segunda parte da equação, $\frac{f(i_n|C_n)}{f(i_n)}$, mede a frequência f de cada item i_n em seu próprio grupo C_n dividido pela frequência do item em todos os grupos. Os itens i_n são todos os itens distintos que aparecem nas regras do grupo. Cada vez que um item i_n ocorre em uma regra sua frequência é incrementada em um. Desse modo, os descritores são construídos pelos N itens mais frequentes em seus próprios grupos e infrequentes nos outros grupos. Considerando C_1 do exemplo acima, seus itens distintos {coffee, butter, milk} e $N = 1$ tem-se que: coffee=$3 * \frac{3}{5}$=1.8; butter=$2 * \frac{2}{5}$=0.8; milk=$2 * \frac{2}{5}$=0.8. Assim, o descritor de C_1 passa a ser {coffee}. Citam-se como trabalhos relacionados a essa ideia [Toivonen et al., 1995] do contexto RA e [Lopes et al., 2007, Treeratpituk and Callan, 2006, Glover et al., 2002, Popescul and Ungar, 2000] do contexto MT e RI.

3.2.2 Metodologia de Avaliação

De modo a avaliar a precisão (P) e a distinção (RF) dos quatro métodos de rotulação acima apresentados, duas medidas, apresentadas nas Equações 8 e 9, foram propostas. Ambas as medidas variam entre 0 e 1. A fim de compreender as medidas considere um agrupamento composto por três grupos de regras de associação: C_1={r_1: coffee $\Rightarrow$ butter; r_2: milk $\Rightarrow$ butter} com os descritores {coffee, butter, milk}; C_2={r_1: butter $\Rightarrow$ coffee; r_2: milk $\Rightarrow$ coffee} com o descritor {milk}; C_3={r_1: butter $\Rightarrow$ milk; r_2: coffee $\Rightarrow$ milk} com os descritores {butter, milk}. O exemplo é meramente ilustrativo.

$$P(C) = \frac{\sum_{i=1}^{N} P(C_i)}{N}, \; onde\ N = \#\ de\ grupos; \tag{8}$$

$$P(C_i) = \frac{\#\{regras\ cobertas\ em\ C_i\ pelos\ descritores\ de\ C_i\}}{\#\{regras\ em\ C_i\}}$$

$$RF(C) = 1 - \frac{\#\{descritores\ distintos\ que\ se\ repetem\ nos\ grupos\}}{\#\{descritores\ distintos\ nos\ grupos\}} \tag{9}$$

A Precisão (P de Precision), na Equação 8, mede o quanto o método de rotulação gera descritores que realmente representam as regras contidas nos grupos. Essa medida é uma adaptação da Recall utilizada em Recuperação da Informação (veja [Manning et al., 2009]). Entretanto, nesse caso, os itens relevantes a serem recuperados são todas as regras em um dado grupo. Considerando o exemplo acima, o método ilustrativo tem uma P de 0.83 ($P(C) = \frac{\frac{2}{2}+\frac{1}{2}+\frac{2}{2}}{3}$), uma vez que os descritores de C_2 representam apenas

uma das duas regras existentes no grupo. Considera-se que uma regra é representada (coberta) por um conjunto de descritores se a regra contém pelo menos um dos descritores. Assim, espera-se que um bom método contenha uma alta precisão. Entretanto, não é suficiente se ter precisão se os descritores aparecerem repetidamente entre os grupos. Desse modo, a Frequência de Repetição (RF de Repetition Frequency), apresentada na Equação 9, mede o quanto os diferentes descritores, presentes em todos os grupos, não se repetem. Considerando o exemplo acima, o método ilustrativo tem uma RF de 0.33 ($RF(C) = 1 - \frac{2}{3}$): apenas um (coffee) dos três diferentes descritores (coffee, butter, milk) presentes nos grupos não se repete. Quanto maior o valor de RF, melhor o método, i.e., menos repetição implica em um melhor desempenho.

É importante mencionar que ambas as medidas apenas levam em consideração os descritores dos grupos e as regras contidas nos mesmos, embora alguns métodos também façam uso de outras informações para se obter os descritores de um grupo como o número de transações. Entretanto, o objetivo aqui é exatamente identificar os métodos que fornecem os melhores resultados independentemente do modo como os descritores foram obtidos: o importante é que os descritores sejam precisos (P) e distintos (RF), que são os critérios que consideramos ser relevantes para agrupamentos de regras de associação.

3.2.3 Experimentos

Experimentos foram realizados para avaliar os métodos de rotulação em relação a precisão e a distinção por meio das medidas P e RF. Para tanto, os mesmos conjuntos de dados apresentados na Tabela 1 (página 15) foram utilizados, permitindo que os métodos fossem avaliados em diferentes tipos de dado. Além disso, as mesmas medidas de similaridade foram consideradas. Por outro lado, diferentemente da Seção 3.1.1, em que quatro algoritmos hierárquicos foram executados e o que obteve melhor desempenho considerado, somente o Ward, dentre os hierárquicos, foi executado. Isso porque visa-se, ao final, uma integração dos dois aspectos: "Redução via Interesse" e "Descritivo/Representativo". Em relação ao valor de k, como mencionado no final da Seção 3.1.2, optou-se por variá-lo entre 5 e 50 considerando-se passos de 5. Diante do exposto, a Tabela 13 apresenta as configurações usadas nos experimentos.

Tabela 13: Configurações usadas para avaliação dos métodos de rotulação no pós-processamento.

Conjuntos de Dados	Adult; Income; Groceries; Sup
Algoritmos	PAM; Ward Linkage
Medidas de Similaridade	J-RI; J-RT
k	5 a 50, passos de 5

Considerando as configurações da Tabela 13, os quatro métodos de rotulação (LM-

M; LM-T; LM-S; LM-PU) foram aplicados as diferentes organizações. Em relação aos métodos de rotulação, LM-M e LM-T selecionam apenas uma regra para compor os descritores, LM-S uma ou mais regras, em caso de empate, e LM-PU os 5 itens que apresentam o melhor equilíbrio entre frequência e predição. Assim, na média, todos os métodos de rotulação geram a mesma quantidade de descritores por grupo. Ao final, o desempenho de cada método de rotulação foi avaliado via P e RF, cujos resultados encontram-se apresentados na próxima seção. É importante lembrar que o objetivo das medidas é avaliar, respectivamente, o quanto o método pode encontrar descritores que representam o mais precisamente possível o conhecimento contido em seus próprios grupos e como os descritores estão distribuídos ao longo dos grupos. O ideal é identificar métodos que tenham altos valores para ambas as medidas.

3.2.4 Resultados e Discussão

O desempenho dos métodos de rotulação, avaliados via P e RF, encontram-se apresentados nas Tabelas 14 a 21 agrupados por algoritmo para cada conjunto de dados. As Tabelas 14 e 18 apresentam os resultados para o conjunto Adult, as Tabelas 15 e 19 para o Income, as Tabelas 16 e 20 para o Groceries e as Tabelas 17 e 21 para o Sup. Cada tabela possui 8 colunas principais, cada 4 relacionadas a uma medida de similaridade. Cada uma dessas 4 colunas corresponde a um método de rotulação. Para cada método os valores de P e RF são apresentados. Cada linha apresenta o desempenho dos métodos para um dado valor de k. De modo a facilitar a interpretação das tabelas, considere a Tabela 14. Pode-se observar que, para $k = 5$, em J-RI o método LM-M tem uma P de 0.999, o que significa que quase todas as regras (99.90%) são cobertas pelos descritores obtidos, e RF de 0.273, o que significa que apenas 27.30% dos rótulos distintos não se repetem.

Tabela 14: Resultados do PAM, nos LM's, no conjunto de dados Adult.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.999	0.273	0.906	0.200	0.894	0.500	0.999	0.250	0.985	0.400	0.856	0.667	0.890	0.625	0.987	0.600
10	0.999	0.091	0.965	0.375	0.961	0.250	1.000	0.143	0.992	0.529	0.862	0.750	0.956	0.733	0.997	0.500
15	1.000	0.000	0.963	0.222	0.967	0.100	1.000	0.263	0.995	0.450	0.945	0.765	0.974	0.474	0.999	0.542
20	1.000	0.077	0.980	0.300	0.973	0.182	1.000	0.200	0.995	0.350	0.964	0.526	0.975	0.450	0.999	0.480
25	1.000	0.143	0.984	0.273	0.979	0.250	1.000	0.182	0.995	0.333	0.973	0.579	0.980	0.429	0.999	0.400
30	1.000	0.077	0.988	0.167	0.984	0.167	1.000	0.136	0.998	0.286	0.981	0.316	0.984	0.286	0.999	0.360
35	1.000	0.067	0.989	0.167	0.990	0.000	1.000	0.125	0.999	0.273	0.978	0.333	0.990	0.273	1.000	0.333
40	1.000	0.188	0.988	0.077	0.987	0.154	1.000	0.125	0.999	0.304	0.982	0.318	0.993	0.304	1.000	0.414
45	1.000	0.176	0.991	0.000	0.988	0.071	1.000	0.192	0.999	0.261	0.984	0.273	0.993	0.261	1.000	0.400
50	1.000	0.059	0.994	0.000	0.989	0.133	1.000	0.120	0.999	0.250	0.987	0.261	0.995	0.250	1.000	0.400

Tabela 15: Resultados do PAM, nos LM's, no conjunto de dados Income.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.991	0.417	0.844	0.400	0.889	0.750	0.994	0.533	0.978	0.786	0.785	0.833	0.830	0.714	0.987	0.706
10	0.998	0.313	0.892	0.600	0.909	0.583	0.999	0.381	0.989	0.375	0.841	0.375	0.911	0.643	0.998	0.429
15	0.998	0.250	0.927	0.455	0.942	0.417	0.999	0.261	0.995	0.294	0.858	0.364	0.949	0.438	0.998	0.391
20	0.998	0.063	0.939	0.250	0.954	0.231	1.000	0.208	0.996	0.300	0.892	0.467	0.948	0.421	1.000	0.375
25	0.999	0.059	0.954	0.357	0.971	0.375	0.999	0.080	0.997	0.286	0.907	0.412	0.968	0.400	0.999	0.346
30	0.999	0.059	0.980	0.083	0.984	0.235	1.000	0.038	0.998	0.348	0.960	0.409	0.986	0.435	0.999	0.346
35	0.999	0.200	0.979	0.286	0.980	0.250	1.000	0.077	0.999	0.304	0.969	0.348	0.987	0.348	1.000	0.308
40	0.999	0.200	0.983	0.333	0.983	0.389	1.000	0.038	0.999	0.333	0.979	0.417	0.993	0.375	1.000	0.269
45	0.999	0.200	0.987	0.313	0.983	0.294	0.999	0.038	1.000	0.320	0.985	0.360	0.995	0.360	1.000	0.231
50	0.999	0.150	0.985	0.333	0.993	0.111	1.000	0.000	1.000	0.320	0.988	0.333	0.996	0.320	1.000	0.231

Tabela 16: Resultados do PAM, nos LM's, no conjunto de dados Groceries.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.922	0.800	0.666	0.857	0.778	0.750	0.953	0.588	0.767	0.625	0.599	0.400	0.638	0.500	0.851	0.765
10	0.967	0.667	0.776	0.444	0.858	0.667	0.985	0.647	0.896	0.733	0.697	0.600	0.771	0.667	0.937	0.750
15	0.981	0.632	0.809	0.500	0.866	0.500	0.993	0.600	0.951	0.722	0.802	0.643	0.857	0.688	0.970	0.844
20	0.989	0.500	0.841	0.429	0.951	0.606	0.995	0.604	0.969	0.783	0.841	0.737	0.916	0.783	0.986	0.839
25	0.994	0.476	0.884	0.286	0.973	0.184	0.997	0.526	0.986	0.704	0.870	0.739	0.946	0.778	0.993	0.677
30	0.996	0.409	0.893	0.353	1.000	0.220	0.999	0.464	0.990	0.633	0.895	0.704	0.961	0.700	0.994	0.582
35	0.999	0.400	0.901	0.474	0.978	0.195	0.999	0.382	0.995	0.697	0.905	0.767	0.971	0.697	0.991	0.507
40	0.999	0.360	0.920	0.350	0.991	0.275	1.000	0.379	0.996	0.703	0.922	0.833	0.967	0.703	0.996	0.425
45	1.000	0.407	0.930	0.435	0.996	0.220	1.000	0.357	0.997	0.632	0.915	0.676	0.969	0.632	0.995	0.431
50	1.000	0.367	0.939	0.481	0.996	0.146	1.000	0.327	0.996	0.641	0.919	0.667	0.968	0.636	0.995	0.370

Tabela 17: Resultados do PAM, nos LM's, no conjunto de dados Sup.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.895	0.667	0.715	0.667	0.776	0.714	0.924	0.625	0.771	0.769	0.541	0.667	0.590	0.714	0.826	0.778
10	0.952	0.600	0.763	0.556	0.811	0.333	0.962	0.417	0.797	0.556	0.528	0.600	0.645	0.667	0.845	0.688
15	0.976	0.600	0.780	0.455	0.813	0.455	0.983	0.486	0.821	0.524	0.583	0.750	0.679	0.722	0.863	0.711
20	0.980	0.500	0.826	0.429	0.865	0.357	0.985	0.615	0.845	0.423	0.614	0.667	0.669	0.444	0.886	0.705
25	0.983	0.455	0.886	0.471	0.927	0.353	0.988	0.603	0.876	0.429	0.659	0.571	0.742	0.565	0.902	0.676
30	0.986	0.440	0.896	0.500	0.930	0.400	0.990	0.571	0.893	0.469	0.684	0.429	0.770	0.643	0.913	0.682
35	0.990	0.444	0.897	0.500	0.946	0.435	0.992	0.563	0.900	0.528	0.695	0.458	0.792	0.636	0.924	0.725
40	0.988	0.467	0.908	0.556	0.940	0.440	0.992	0.568	0.915	0.526	0.700	0.556	0.815	0.556	0.932	0.719
45	0.991	0.500	0.904	0.417	0.932	0.348	0.993	0.571	0.927	0.571	0.720	0.625	0.834	0.650	0.944	0.788
50	0.993	0.529	0.913	0.481	0.935	0.370	0.994	0.574	0.931	0.558	0.771	0.571	0.853	0.548	0.956	0.754

Tabela 18: Resultados do Ward, nos LM's, no conjunto de dados Adult.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.992	0.500	0.751	0.400	0.863	0.500	0.982	0.500	0.966	0.500	0.801	0.833	0.874	0.556	0.976	0.667
10	0.997	0.412	0.840	0.600	0.922	0.583	0.994	0.450	0.986	0.529	0.904	0.818	0.958	0.333	0.995	0.524
15	0.998	0.421	0.857	0.643	0.966	0.588	0.995	0.500	0.995	0.450	0.956	0.733	0.986	0.368	0.996	0.500
20	0.998	0.400	0.902	0.500	0.978	0.526	0.986	0.500	0.995	0.300	0.970	0.588	0.986	0.350	0.998	0.417
25	0.999	0.350	0.921	0.471	0.982	0.474	0.989	0.423	0.995	0.286	0.971	0.579	0.990	0.381	0.994	0.360
30	0.999	0.381	0.938	0.474	0.989	0.429	0.988	0.407	0.998	0.238	0.977	0.526	0.992	0.333	0.995	0.320
35	0.999	0.333	0.946	0.421	0.990	0.429	0.989	0.407	0.998	0.273	0.981	0.476	0.996	0.273	0.997	0.346
40	0.999	0.286	0.956	0.263	0.991	0.333	0.991	0.429	0.999	0.261	0.988	0.381	0.996	0.261	0.998	0.333
45	0.999	0.190	0.963	0.200	0.991	0.190	0.992	0.321	0.999	0.217	0.985	0.333	0.996	0.217	0.998	0.357
50	0.999	0.190	0.969	0.200	0.992	0.190	0.993	0.321	0.999	0.250	0.990	0.364	0.997	0.250	0.998	0.357

Tabela 19: Resultados do Ward, nos LM's, no conjunto de dados Income.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.980	0.615	0.840	0.833	0.886	0.857	0.985	0.563	0.941	0.769	0.693	0.833	0.822	0.714	0.968	0.667
10	0.990	0.294	0.897	0.727	0.923	0.846	0.994	0.333	0.968	0.500	0.807	0.727	0.896	0.667	0.988	0.435
15	0.993	0.333	0.907	0.538	0.941	0.400	0.997	0.375	0.978	0.278	0.869	0.733	0.935	0.529	1.000	0.462
20	0.993	0.368	0.923	0.467	0.964	0.412	0.999	0.320	0.987	0.250	0.907	0.471	0.951	0.500	1.000	0.385
25	0.994	0.263	0.934	0.313	0.971	0.294	0.999	0.320	0.993	0.348	0.947	0.571	0.972	0.522	0.999	0.385
30	0.995	0.300	0.940	0.313	0.976	0.333	0.999	0.280	0.992	0.348	0.954	0.478	0.977	0.391	1.000	0.385
35	0.996	0.333	0.944	0.353	0.979	0.400	0.999	0.240	0.993	0.348	0.961	0.391	0.981	0.391	1.000	0.385
40	0.997	0.238	0.950	0.333	0.984	0.200	0.999	0.231	0.994	0.348	0.967	0.348	0.984	0.391	1.000	0.385
45	0.997	0.273	0.954	0.368	0.985	0.238	0.999	0.231	0.995	0.261	0.971	0.261	0.986	0.304	1.000	0.308
50	0.997	0.273	0.958	0.316	0.986	0.238	0.999	0.231	0.997	0.250	0.979	0.250	0.990	0.292	1.000	0.269

Tabela 20: Resultados do Ward, nos LM's, no conjunto de dados Groceries.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.827	0.778	0.734	0.833	0.794	0.667	0.871	0.571	0.776	0.667	0.658	0.833	0.721	0.600	0.869	0.778
10	0.912	0.857	0.849	0.818	0.867	0.818	0.931	0.652	0.909	0.643	0.868	0.909	0.907	0.818	0.939	0.619
15	0.941	0.909	0.863	0.875	0.867	0.875	0.954	0.645	0.900	0.750	0.841	0.938	0.900	0.850	0.958	0.655
20	0.954	0.920	0.904	0.905	0.917	0.864	0.965	0.684	0.931	0.792	0.880	0.905	0.887	0.913	0.964	0.613
25	0.965	0.862	0.933	0.885	0.939	0.815	0.974	0.595	0.963	0.828	0.946	0.846	0.968	0.786	0.972	0.531
30	0.975	0.806	0.953	0.867	0.968	0.833	0.978	0.565	0.965	0.781	0.952	0.793	0.967	0.806	0.974	0.514
35	0.979	0.806	0.959	0.853	0.974	0.736	0.980	0.529	0.959	0.667	0.948	0.688	0.958	0.667	0.977	0.487
40	0.972	0.750	0.947	0.842	0.979	0.632	0.985	0.509	0.964	0.588	0.955	0.647	0.966	0.618	0.979	0.512
45	0.978	0.698	0.954	0.805	0.983	0.656	0.989	0.509	0.973	0.632	0.962	0.676	0.975	0.658	0.982	0.465
50	0.973	0.644	0.960	0.795	0.998	0.625	0.990	0.533	0.961	0.625	0.959	0.667	0.992	0.583	0.985	0.500

Tabela 21: Resultados do Ward, nos LM's, no conjunto de dados Sup.

	J-RI								J-RT							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.890	0.556	0.706	0.667	0.764	0.833	0.900	0.600	0.771	0.769	0.541	0.667	0.590	0.714	0.826	0.778
10	0.947	0.733	0.855	0.818	0.892	0.833	0.961	0.550	0.797	0.556	0.528	0.600	0.645	0.667	0.845	0.688
15	0.954	0.750	0.879	0.875	0.918	0.750	0.968	0.571	0.821	0.524	0.583	0.750	0.679	0.722	0.863	0.711
20	0.966	0.636	0.926	0.850	0.951	0.750	0.975	0.583	0.845	0.423	0.614	0.667	0.669	0.444	0.886	0.705
25	0.975	0.704	0.938	0.880	0.970	0.808	0.980	0.561	0.876	0.429	0.659	0.571	0.742	0.565	0.902	0.676
30	0.979	0.750	0.947	0.900	0.964	0.800	0.983	0.587	0.893	0.469	0.684	0.429	0.770	0.643	0.913	0.682
35	0.968	0.784	0.935	0.914	0.960	0.833	0.981	0.600	0.900	0.528	0.695	0.458	0.792	0.636	0.924	0.725
40	0.976	0.818	0.937	0.900	0.957	0.854	0.984	0.552	0.915	0.526	0.700	0.556	0.815	0.556	0.932	0.719
45	0.982	0.813	0.955	0.911	0.975	0.870	0.986	0.525	0.927	0.571	0.720	0.625	0.834	0.650	0.944	0.788
50	0.986	0.827	0.969	0.900	0.981	0.880	0.987	0.516	0.931	0.558	0.771	0.571	0.853	0.548	0.956	0.754

Com base nesses dados, de modo a identificar os métodos mais adequados para agrupamentos de regras de associação e as organizações que fornecem os melhores resultados, uma análise baseada na média de cada medida (P, RF), considerando os diferentes tipos de dado, independente do conjunto, foi realizada. A Tabela 22 apresenta os resultados. Cada média foi obtida a partir dos resultados dos experimentos relacionados a configuração apresentada. O valor P de 0.999 de LM-M da configuração CD-P:PAM:J-RI, por exemplo, foi obtido por meio da média dos valores de P de LM-M das configurações PAM:Adult:J-RI e PAM:Income:J-RI das Tabelas 14 e 15. Para facilitar o entendimento, esses valores encontram-se sublinhados nas respectivas tabelas. As médias mais altas, referentes a cada uma das medidas (P, RF), encontram-se marcadas com "*" em cada algoritmo. Para a configuração CD-P:PAM, por exemplo, a média mais representativa para P é a relacionada aos métodos LM-M e LM-PU na medida J-RI (0.999); já em relação a RF ao método LM-T na medida J-RT (0.455). A cada "*" destaca-se na tabela o par P/RF referente a marcação – no caso do exemplo anterior, P/RF de LM-M, LM-PU e LM-T. A medida com "*" no método sublinhado indica o valor que fez com que o método fosse selecionado – no caso do exemplo anterior, em LM-M P (0.999), em LM-PU P (0.999) e em LM-T RF (0.455). Desse modo, é possível visualizar, para cada configuração tipo-de-dado+algoritmo, o método que apresenta o melhor desempenho e em qual medida a condição acontece. Por fim, é importante mencionar que os resultados são determinísticos e, portanto, nenhum teste estatístico precisa ser realizado para checar se existe diferença significativa entre as médias. Pode-se observar que:

- em relação ao algoritmo PAM nos CD-P, os métodos que apresentam os melhores resultados em P são o LM-M e o LM-PU e em RF o LM-T. Entretanto, nota-se que em todos os métodos selecionados P apresenta altos valores e RF, em LM-M e LM-PU, valores muito baixos. Portanto, LM-T se torna o mais indicado quando se busca um equilíbrio entre P e RF, uma vez que ele melhora RF enquanto mantém um bom valor para P. Observa-se também que essa condição ocorre com o uso da medida J-RT.

- em relação ao algoritmo Ward nos CD-P, o método que apresenta o melhor resultado em P é o LM-M e em RF o LM-T. Entretanto, nota-se que em ambos os métodos selecionados P apresenta altos valores e RF valores não tão altos. Portanto, LM-T se torna o mais indicado quando se busca um equilíbrio entre P e RF, uma vez que ele melhora RF enquanto mantém um bom valor para P. Observa-se também que essa condição ocorre com o uso da medida J-RT.

- em relação ao algoritmo PAM nos CD-NP, o método que apresenta o melhor resultado tanto em P quanto em RF é o LM-PU, sendo, portanto, o mais indicado

nesse caso. Em relação as medidas de similaridade, uma vez que P apresenta valores acima de 0.9 em ambas, dá-se preferência a medida J-RT, uma vez que a mesma apresenta RF próxima de 0.7.

- em relação ao algoritmo Ward nos CD-NP, o método que apresenta o melhor resultado em P é o LM-PU e em RF o LM-T. Entretanto, nota-se que em ambos os métodos selecionados P apresenta altos valores e RF, em LM-PU, um valor não tão alto. Portanto, LM-T se torna o mais indicado quando se busca um equilíbrio entre P e RF, uma vez que ele melhora RF enquanto mantém um bom valor para P. Observa-se também que essa condição ocorre com o uso da medida J-RI.

Diante do exposto, pode-se notar que:

- em relação aos métodos de rotulação tem-se que:
  - nos CD-P o método que se mostra mais adequado é o LM-T;
  - nos CD-NP os métodos que se mostram mais adequados são o LM-T e o LM-PU.
- em todos os casos acima destacados P apresenta bons valores tendo, portanto, a medida RF influenciado na escolha dos métodos (LM-T; LM-PU).
- com exceção da configuração CD-NP:Ward, os métodos selecionados (LM-T; LM-PU) são mais indicados quando utilizados com a medida J-RT.
- em relação aos algoritmos tem-se que:
  - comparando-se os pares P/RF da Tabela 22 do PAM com o Ward nos CD-P, observa-se que o Ward apresenta um melhor desempenho;
  - comparando-se os pares P/RF da Tabela 22 do PAM com o Ward nos CD-NP, observa-se que o Ward apresenta um desempenho um pouco melhor.

Tabela 22: Média de P e RF nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida) em cada LM.

Tipo de Dado	Algoritmo	Medida	LM-M		LM-T		LM-S		LM-PU	
			P	RF	P	RF	P	RF	P	RF
CD-P	PAM	J-RI	0.999*	0.153	0.961	0.260	0.965	0.272	0.999*	0.170
		J-RT	0.995	0.355	0.934	0.455*	0.965	0.427	0.998	0.403
	Ward	J-RI	0.996*	0.338	0.915	0.437	0.963	0.423	0.993	0.369
		J-RT	0.988	0.350	0.929	0.535*	0.963	0.401	0.995	0.412
CD-NP	PAM	J-RI	0.979	0.511	0.852	0.482	0.913	0.398	0.986*	0.523
		J-RT	0.911	0.611	0.743	0.633	0.818	0.646	0.935	0.671*
	Ward	J-RI	0.955	0.770	0.905	0.855*	0.931	0.787	0.966*	0.572
		J-RT	0.899	0.616	0.773	0.690	0.832	0.672	0.929	0.645

Por fim, analisando somente as organizações que se destacaram na Seção 3.1.2 referentes a utilização da PAR-COM, apresenta-se novamente na Tabela 23 alguns dados da Tabela 22. As médias mais altas, referentes a cada uma das medidas (P, RF), encontram-se marcadas com "*" em cada organização. Para a organização PAM:J-RT nos CD-P, por exemplo, a média mais representativa para P é a relacionada ao método LM-PU (0.998); já em relação a RF ao método LM-T (0.455). Como anteriormente, a cada "*" destaca-se na tabela o par P/RF referente a marcação, tendo a medida com "*" contribuído para que o método fosse selecionado. Pode-se observar que:

- em relação a organização PAM:J-RT nos CD-P, o método que apresenta o melhor resultado em P é o LM-PU e em RF o LM-T. Uma vez que em ambos os métodos selecionados P apresenta um valor acima de 0.9, considerou-se que LM-T apresenta um desempenho um pouco melhor baseando-se em RF.

- em relação a organização Ward:J-RT nos CD-P, o método que apresenta o melhor resultado em P é o LM-PU e em RF o LM-T. Uma vez que em ambos os métodos selecionados P apresenta um valor acima de 0.9, considerou-se que LM-T apresenta um melhor desempenho baseando-se em RF.

- em relação a organização PAM:J-RI nos CD-NP, o método que apresenta o melhor resultado tanto em P quanto em RF é o LM-PU.

- em relação a organização Ward:J-RT nos CD-NP, o método que apresenta o melhor resultado em P é o LM-PU e em RF o LM-T. Uma vez que em ambos os métodos selecionados RF apresenta valores próximos a 0.7, considerou-se que LM-PU apresenta um melhor desempenho baseando-se em P.

- nos CD-P a organização Ward:J-RT é a que apresenta o melhor desempenho por meio do método LM-T considerando a medida RF (0.934/0.455 x 0.998/0.403 x 0.929/0.535 x 0.995/0.412), uma vez que P se mantém estável. Porém, nota-se que tanto o LM-T quanto o LM-PU se destacam nos CD-P, apresentando desempenhos similares.

- nos CD-NP a organização Ward:J-RT é a que apresenta o melhor desempenho por meio do método LM-PU (0.986/0.523 x 0.773/0.690 x 0.929/0.645); no caso da organização Ward:J-RT opta-se pelo LM-PU devido a grande diferença em P (0.773 x 0.929) e a pequena diferença em RF (0.690 x 0.645).

Considerando as discussões acima apresentadas, tem-se que: (i) nos CD-P o método que se mostra mais adequado para agrupamentos de regras de associação é o LM-T,

Tabela 23: Média de P e RF nos diferentes tipos de dado nas organizações que se destacaram na Seção 3.1.2 em cada LM.

Tipo de Dado	Algoritmo	Medida	LM-M		LM-T		LM-S		LM-PU	
			P	RF	P	RF	P	RF	P	RF
CD-P	PAM	J-RT	0.995	0.355	0.934	0.455*	0.965	0.427	0.998*	0.403
	Ward	J-RT	0.988	0.350	0.929	0.535*	0.963	0.401	0.995*	0.412
CD-NP	PAM	J-RI	0.979	0.511	0.852	0.482	0.913	0.398	0.986*	0.523*
	Ward	J-RT	0.899	0.616	0.773	0.690*	0.832	0.672	0.929*	0.645

embora o método LM-PU também tenha apresentado um bom desempenho; (ii) nos CD-NP o método que se mostra mais adequado para agrupamentos de regras de associação é o LM-PU; (iii) os métodos apresentam melhores resultados quando obtém-se o agrupamento por meio do Ward; (iv) J-RT se mostra uma boa medida de similaridade a ser utilizada com o Ward; (v) como consequência de (iii) e (iv), nota-se que a configuração Ward:J-RT representa a organização que melhor separa o conhecimento do domínio – pode-se inferir que o domínio encontra-se bem separado se a organização, juntamente com um método de rotulação adequado, fornece bons descritores. Essas conclusões abrangem os três objetivos mencionados na Seção 3.2 (letras **(a)** a **(c)**). As Figuras 7 e 8 apresentam o comportamento dos métodos LM-T e LM-PU, na organização Ward:J-RT, tanto nos CD-P quanto nos CD-NP. É importante notar que embora P se mantenha estável e apresente altos valores em todos os valores de k o mesmo não ocorre com RF, cuja tendência é diminuir conforme o número de clusters aumenta. Entretanto, **nota-se que os descritores obtidos pelos métodos LM-T e LM-PU fornecem ao usuário uma visão dos tópicos a serem explorados com uma boa precisão (P) e uma razoável distinção (RF) em ambos os tipos de dado. Portanto, por meio desses dois métodos (LM-T, LM-PU), é possível afirmar que o agrupamento no pós-processamento é útil ao usuário.**

Por fim, observa-se que a maioria dos trabalhos da literatura relacionados a agrupamentos de regras de associação, descritos na Seção 2.3.1, basicamente exploram o método LM-M, o qual não se destaca em nenhuma das organizações. Além disso, o LM-M apresenta um alto custo computacional, sendo o LM-T e o LM-PU os dois menos custosos. Desse modo, acredita-se, com base no que foi exposto, que outros métodos de MT e RI poderiam ser adaptados para agrupamentos de regras de associação, uma vez que o LM-T e o LM-PU apresentaram um bom desempenho.

3.3 Considerações Finais

A Tabela 24 apresenta as conclusões obtidas nas Seções 3.1 e 3.2 referente aos aspectos abordados: "Redução via Interesse" e "Descritivo/Representativo". Uma vez que visa-se a integração dos dois aspectos, de modo a atender, simultaneamente, os requisitos (R-i) e (R-

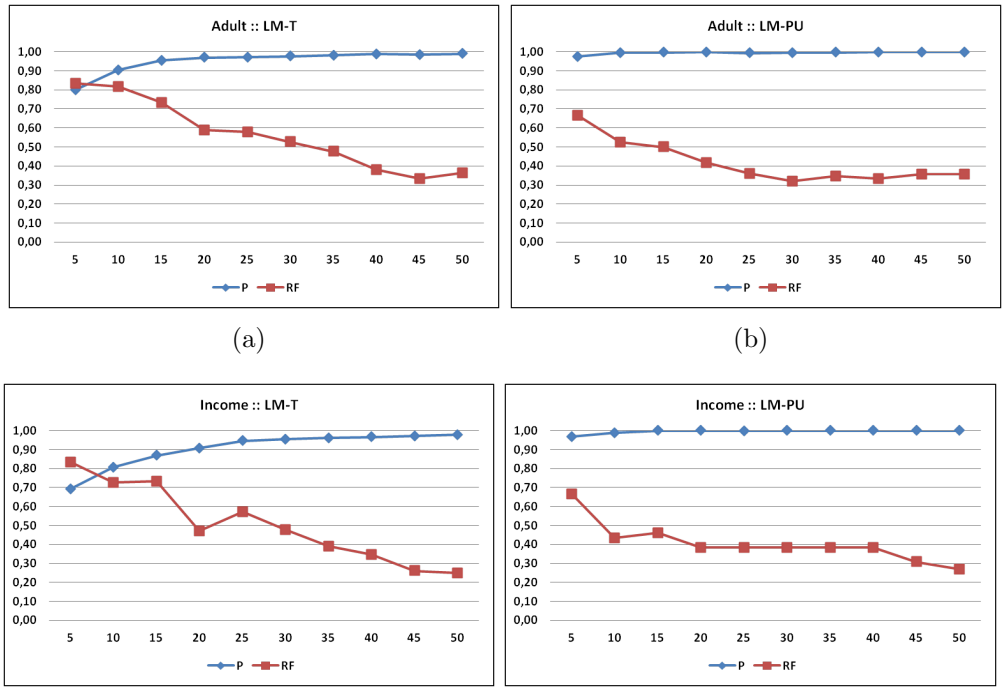


Figura 7: Comportamento dos métodos LM-T e LM-PU, na organização Ward:J-RT, nos CD-P.

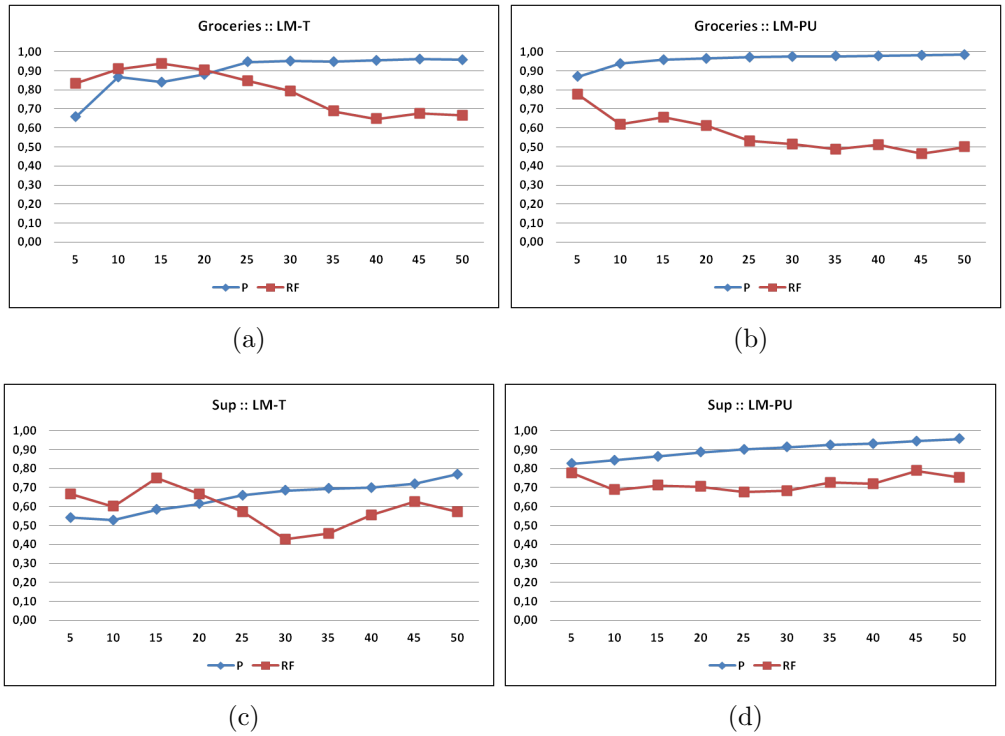


Figura 8: Comportamento dos métodos LM-T e LM-PU, na organização Ward:J-RT, nos CD-NP.

ii), tem-se que: (i) independente do tipo de dado utilizado (CD-P; CD-NP) a organização que se destaca nos dois aspectos é a Ward:J-RT, sendo, portanto, a mais indicada para agrupamentos de regras de associação na etapa de pós-processamento; (ii) em relação aos métodos de rotulação, tem-se que o LM-T é mais indicado para CD-P e o LM-PU para CD-NP. Desse modo, uma vez que há um *matching* perfeito entre os resultados dos dois aspectos, **o agrupamento de regras no pós-processamento possibilita que: (i) via PAR-COM, o espaço de exploração seja reduzido direcionando o usuário para o que é potencialmente interessante; (ii) via LM-T e LM-PU, bons descritores sejam obtidos de modo a fornecer uma visão dos tópicos a serem explorados. Portanto, considerando o exposto, pode-se afirmar que o agrupamento no pós-processamento apóia o usuário.**

Tabela 24: Organizações e métodos de rotulação adequados ao pós-processamento referente aos aspectos "Redução via Interesse" e "Descritivo/Representativo".

"Redução via Interesse"	
CD-P	Ward:J-RT
CD-NP	Ward:J-RT
"Descritivo/Representativo"	
CD-P	LM-T Ward:J-RT
CD-NP	LM-PU Ward:J-RT

Por fim, destacam-se as contribuições obtidas, por meio desse estudo, em relação ao agrupamento de regras de associação na etapa de pós-processamento:

- proposta da metodologia PAR-COM;
- proposta das medidas P e RF para se avaliar métodos de rotulação;
- identificação da organização mais adequada para se reduzir o espaço de exploração, identificar o conhecimento potencialmente interessante e obter bons descritores;
- identificação dos métodos de rotulação mais adequados ao contexto apresentado;
- identificação da possibilidade de se integrar a PAR-COM com os métodos de rotulação adequados a cada tipo de dado de modo a permitir a identificação dos tópicos potencialmente mais interessantes do domínio.

4 Agrupamento de Regras de Associação no Pré-processamento

Como mencionado na Seção 1, o objetivo do agrupamento na etapa de pré-processamento é obter regras potencialmente interessantes que não seriam extraídas a partir de conjuntos não particionado por não terem suporte suficiente sem, contudo, sobrecarregar o usuário com uma grande quantidade de padrões. Isso porque, para que esses mesmos padrões sejam descobertos a partir de conjuntos não particionado deve-se configurar o suporte mínimo com um valor muito baixo e abaixo do utilizado nos conjuntos particionado, fazendo com que o número de regras cresça rapidamente. Caso o usuário configure sup-min em 10%, por exemplo, para um dado conjunto particionado, esse valor deverá estar bem abaixo no conjunto não particionado para que os mesmos padrões sejam obtidos, gerando-se uma quantidade maior ou igual de padrões nesse último caso em relação ao primeiro. Assim, no pré-processamento, particiona-se inicialmente os dados e extrai-se as regras dentro de cada grupo do agrupamento, fazendo com que cada grupo expresse suas próprias associações [Koh and Pears, 2008]. Portanto, o agrupamento se torna útil ao usuário, assim como no pós-processamento, se: (R-i) de todos os grupos apresentados destacaram-se aqueles que contêm os padrões potencialmente mais interessantes, reduzindo assim o espaço de exploração; (R-ii) os grupos contiverem bons descritores que forneçam uma visão dos tópicos a serem explorados. Entretanto, diferentemente do pós-processamento, existem algumas particularidades a serem investigadas, a saber:

Questão 1 Existe sobreposição entre os conjuntos de regras obtidos a partir de dados particionado (RcP[6]), ou seja, extraídos no pré-processamento, em relação aos obtidos a partir de dados não particionado (RcNP[7]), ou seja, extraídos via processo tradicional? Considera-se que em RcP encontram-se as regras extraídas em todos os grupos. Desse modo, como observa-se na Figura 9, os grupos resultantes são vistos como integrantes de um único agrupamento (vide explicação na Seção 1).

Questão 2 Existe sobreposição entre as regras em RcNP e RcP em relação a interessabilidade do conhecimento? Ou seja, os conjuntos de regras obtidos a partir de dados particionado contêm, de fato, padrões mais interessantes e específicos do domínio em relação aos obtidos a partir de dados não particionado?

Questão 3 Qual o comportamento do processo em relação ao número de regras obtidas em RcP? Ou seja, o número de regras em RcP é maior ou menor do que o número de regras em RcNP?

[6]RcP = conjunto de *R*egras obtido a partir de um *C*onjunto de dados *P*articionado.
[7]RcNP = conjunto de *R*egras obtido a partir de um *C*onjunto de dados não *P*articionado.

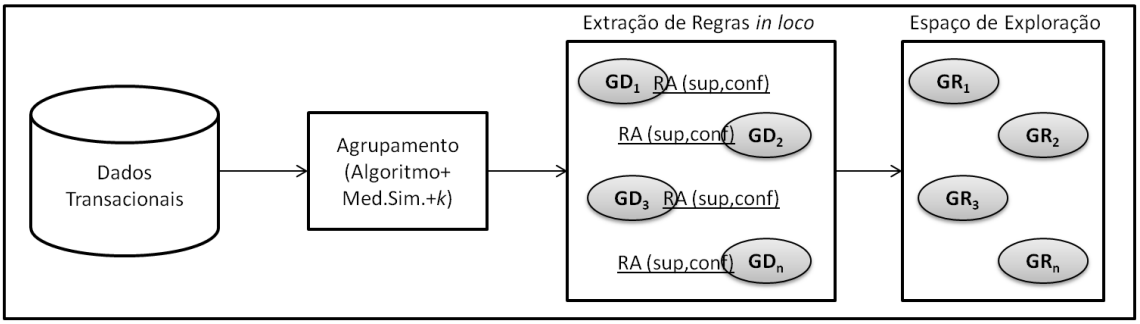


Figura 9: Reprodução da Figura 1(a), página 2: Visão geral do processo de agrupamento de regras de associação no pré-processamento.

Para que essas perguntas sejam respondidas, vários índices foram propostos, os quais encontram-se descritos na Seção 4.2. Contudo, para que esses índices sejam computados, a fim de checar se o agrupamento no pré-processamento auxilia ou não o processo de exploração, é necessário que se identifique, inicialmente, as organizações mais adequadas a essa etapa. Desse modo, a Seção 4.1 apresenta uma avaliação sobre várias organizações. Somente após avaliar a contribuição do agrupamento no pré-processamento segundo os índices propostos (Seção 4.2), considerando a avaliação das organizações da Seção 4.1, é que uma análise em relação aos requisitos (R-i) e (R-ii) é realizada. Assim, apresentam-se na Seção 4.3 os resultados referentes ao requisito (R-i) e na Seção 4.4 os resultados referentes ao requisito (R-ii).

4.1 Avaliação das Organizações

Como mencionado anteriormente, antes de se avaliar a contribuição do agrupamento no pré-processamento, segundo os índices propostos e apresentados na Seção 4.2, é necessário que se identifique, inicialmente, as organizações mais adequadas a essa etapa, i.e, as melhores combinações entre algoritmos e medidas de similaridade. Desse modo, experimentos foram realizados segundo o processo da Figura 9. Para tanto, definiu-se os algoritmos e as medidas de similaridade a serem utilizadas, assim como o intervalo de k. Os algoritmos inicialmente selecionados foram os mesmos utilizados no pós-processamento (PAM, Single, Complete, Average e Ward). Porém, em relação as medidas de similaridade, considerou-se, nesse caso, as descritas na Seção 2.3.2 (Equações 3, 4, 5 e 7), utilizadas nos trabalhos relacionados a agrupamentos de regras de associação no pré-processamento. Em relação a essas medidas tem-se que:

- P-J (Equação 3, página 8) e P-RR (Equação 4, página 8) medem a similaridade existente entre os itens distintos do conjunto de dados. Desse modo, após o agrupamento, cada grupo contém um subconjunto dos itens distintos. Assim, é necessário

que as transações do conjunto sejam distribuídas entre os grupos a fim de possibilitar a extração das regras de associação. Para que essa distribuição fosse realizada utilizou-se a Equação 6 (página 10) da Seção 2.3.2, atribuindo cada transação ao grupo em que a mesma obtivesse o maior valor de similaridade.

- Agg (Equação 5, página 9) e Denza (Equação 7, página 10) medem a similaridade existente entre as transações do conjunto de dados. Desse modo, após o agrupamento, cada grupo já contém um subconjunto das transações. Assim, diferentemente das medidas anteriores, não é necessário que distribuições sejam realizadas para que a extração das regras de associação seja executada.

Em relação a k, considerou-se os valores entre 5 e 25 em passos de 5. A justificativa desse escolha é apresentada mais adiante. Uma outra questão importante a ser definida se referia aos valores de suporte e confiança mínimos a serem configurados durante a extração de regras em cada grupo, como observa-se na Figura 9. Visando automatizar a configuração do suporte em cada um dos grupos, o seguinte procedimento foi adotado: (i) encontrar os 1-itemsets do grupo com seus respectivos suportes, (ii) calcular a média desses suportes e, ao final, (iii) utilizar esse suporte médio como suporte mínimo do grupo. Desse modo, garantiria-se que o suporte de todos os grupos, em todos os experimentos, seriam obtidos da mesma maneira, levando, posteriormente, a uma comparação mais justa. Em relação a confiança, optou-se por utilizar a mesma em todos os grupos de um dado agrupamento.

Diante do exposto, para que as melhores organizações (algoritmo+medida) pudessem ser selecionadas, para que a abordagem de pré-processamento fosse avaliada segundos os índices propostos (Seção 4.2), os mesmos quatro conjuntos de dados utilizados nos pós-processamento foram aqui considerados. Em relação ao Adult, a fim de acelerar a execução dos experimentos, uma amostra do conjunto foi obtida como descrito em [Hahsler et al., 2012], obtendo-se, ao final, 9210 transações. Vale ressaltar que a fim de padronizar os experimentos em ambas as abordagens, a extração de padrões foi aqui realizada considerando-se também um número mínimo de 2 itens e um número máximo de 5 itens por regra.

Para que a seleção das organizações mais adequadas ao pré-processamento fosse realizada, considerou-se como critério o comportamento da distribuição das transações nos grupos. Isso porque, é importante que as transações de um dado agrupamento estejam distribuídas proporcionalmente entre os grupos, uma vez que não é interessante que existam muitos grupos com poucas transações e poucos grupos com muitas transações, o que pode afetar significativamente os resultados. Além disso, uma vez que as regras são obtidas dentro de cada grupo, se a quantidade de transações em um dado grupo for muito

pequena (i) o número de regras a ser obtido pode se tornar muito grande e (ii) o usuário pode acabar sendo sobrecarregado com regras que não expressam nenhum conhecimento interessante, por serem simplesmente combinações dos itens distintos contidos no grupo. Desse modo, considerou-se que regras seriam extraídas de um grupo apenas se o mesmo contivesse, no mínimo, 0.5% do total das transações.

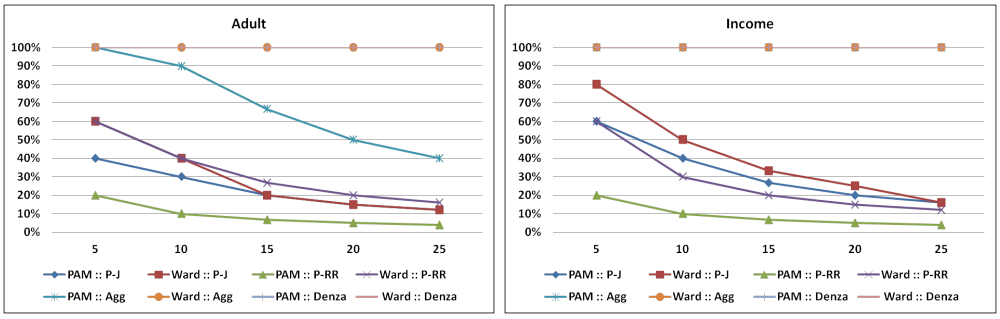
Nesse contexto, a Tabela 25 apresenta as configurações utilizadas para se selecionar, segundo o critério de distribuição acima, as organizações a serem consideradas. Em relação aos algoritmos hierárquicos, apresenta-se, na tabela, apenas aquele que obteve o melhor desempenho, nesse caso, o Ward.

Tabela 25: Configurações usadas para selecionar as organizações mais adequadas ao pré-processamento segundo o critério de distribuição das transações.

Conjuntos de Dados	Adult; Income; Groceries; Sup
Algoritmos	PAM; Ward Linkage
Medidas de Similaridade	P-J; P-RR; Agg; Denza
k	5 a 25, passos de 5

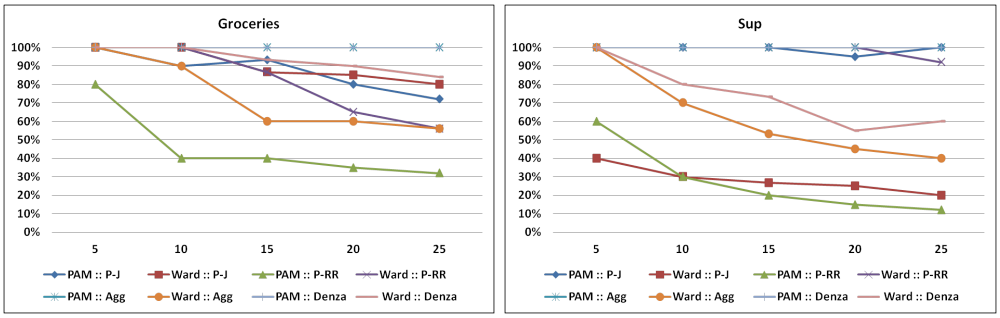
As Figuras 10 e 11 apresentam o comportamento da distribuição das transações nos agrupamentos considerados, após o processo da Figura 9, em cada um dos tipos de dado. O eixo x de cada gráfico representa o intervalo de k. O eixo y a porcentagem de grupos do agrupamento que contém, no mínimo, 0.5% do total das transações. Desse modo, quanto maior a porcentagem, melhor a organização (algoritmo+medida) distribui as transações nos grupos do agrupamento. A fim de auxiliar a interpretação dos gráficos considere a Figura 10(a). Pode-se observar que, para $k = 5$, apenas 20% dos grupos da organização PAM:P-RR (1 grupo (20% de 5 grupos)) contém, pelo menos, 0.5% do total das transações, não tendo, portanto, contribuído para um bom particionamento do conjunto de dados. Em outras palavras, 4 dos 5 grupos contém um número inexpressivo de transações, fazendo com que o agrupamento fique desbalanceado. Avaliando os gráficos tem-se que:

- no conjunto Adult, as organizações que melhor distribuíram os dados (100% de utilização dos grupos) foram Ward:Agg, PAM:Denza e Ward:Denza;
- no conjunto Income, as organizações que melhor distribuíram os dados (100% de utilização dos grupos) foram PAM:Agg, Ward:Agg, PAM:Denza e Ward:Denza;
- no conjunto Groceries, as organizações que melhor distribuíram os dados (100% de utilização dos grupos) foram PAM:Agg e PAM:Denza;
- no conjunto Sup, as organizações que melhor distribuíram os dados (100% de utilização dos grupos) foram PAM:Agg e PAM:Denza.



(a) (b)

Figura 10: Distribuição das transações nos agrupamentos dos CD-P.



(a) (b)

Figura 11: Distribuição das transações nos agrupamentos dos CD-NP.

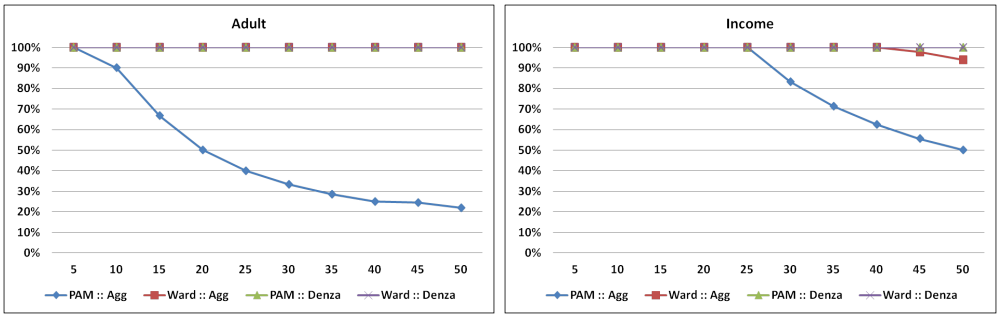
Diante do exposto, pode-se notar que: (i) as medidas mais adequadas para se realizar o agrupamento são aquelas relacionadas a similaridade existente entre as transações, nesse caso, Agg e Denza; (ii) as organizações mais apropriadas ao pré-processamento, em decorrência de (i), a serem consideradas são PAM:Agg, Ward:Agg, PAM:Denza e Ward:Denza; (iii) as medidas relacionadas a similaridade existente entre os itens distintos não apresentam, em geral, um bom desempenho; (iv) com exceção das organizações que proporcionaram 100% de utilização dos grupos no que se refere a distribuição das transações, existe uma tendência desse valor diminuir conforme o valor de k aumenta (veja a tendência decrescente nos gráficos). A princípio, a escolha pelo intervalo de k entre 5 e 25, em passos de 5, se deu justamente pelo fato de se realizar uma avaliação inicial da distribuição das transações nos agrupamentos e checar posteriormente a necessidade de se expandir esse valor. Desse modo, visando-se averiguar a necessidade de expansão desse intervalo e considerando a observação do item (iv), estendeu-se o valor de k até 50, em passos de 5, apenas para as organizações do item (ii), ou seja, apenas para aquelas a serem consideradas. As Figuras 12 e 13 apresentam os resultados obtidos. Avaliando os gráficos tem-se que:

- as organizações, anteriormente identificadas, no conjunto Adult continuam a proporcionar 100% de utilização dos grupos em todo o intervalo;

- o comportamento das organizações, anteriormente identificadas, no conjunto Income permanece o mesmo somente até $k = 25$. Após esse limiar, para algumas das organizações, a proporção de utilização dos grupos começa a cair;

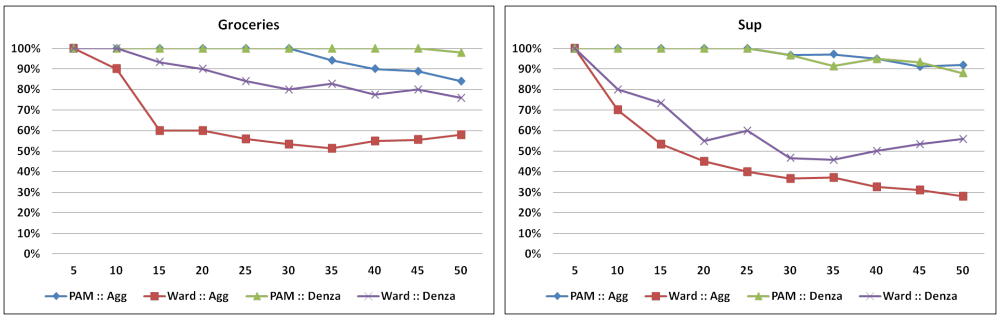
- o comportamento das organizações, anteriormente identificadas, no conjunto Groceries permanece o mesmo somente até $k = 30$. Após esse limiar, para algumas da organizações, a proporção de utilização dos grupos começa a cair;

- o comportamento das organizações, anteriormente identificadas, no conjunto Sup permanece o mesmo somente até $k = 25$. Após esse limiar, para todas as organizações, a proporção de utilização dos grupos começa a cair.



(a) (b)

Figura 12: Distribuição das transações nos agrupamentos dos CD-P para um novo intervalo de k.



(a) (b)

Figura 13: Distribuição das transações nos agrupamentos dos CD-NP para um novo intervalo de k.

Observando os resultados acima nota-se, como identificado anteriormente, que a proporção de utilização dos grupos tende a diminuir conforme o valor de k aumenta. Desse modo, a fim de garantir uma comparação mais justa entre as organizações a serem consideradas, no que se refere a distribuição das transações nos agrupamentos, optou-se por manter o intervalo de k entre 5 e 25 em passos de 5. Além disso, observa-se que valores de k não

muito grandes devem ser escolhidos a fim de garantir uma boa distribuição das transações. Caso contrário, se poucos grupos contiverem quase que a totalidade das transações, pequena será a diferença entre o conjunto de regras obtido no pré-processamento em relação ao obtido via processo tradicional. Com base no exposto, com o objetivo de analisar as organizações selecionadas (PAM:Agg, Ward:Agg, PAM:Denza e Ward:Denza) segundo os índices abaixo propostos (Seção 4.2) e, consequentemente, discutir a contribuição do agrupamento na etapa de pré-processamento, experimentos foram realizados, os quais são descritos mais adiante nas Seções 4.2.1 e 4.2.2.

4.2 Índices de Avaliação

A fim de responder as perguntas elencadas no início da Seção 4, referentes as particularidades do agrupamento na etapa de pré-processamento, vários índices foram propostos, os quais encontram-se descritos a seguir. Para cada particularidade um ou mais índices são propostos, para que em conjunto sejam utilizados para analisar a contribuição da abordagem de pré-processamento. É importante mencionar que todos os índices, com exceção do I_{NR-RcP}, variam entre 0 e 1. Além disso, cabe ressaltar que em RcNP as regras são únicas, ou seja, não existem regras repetidas no conjunto. Por outro lado, em RcP podem existir regras repetidas, uma vez que as regras são extraídas dentro de cada grupo. Assim, nas equações abaixo apresentadas, é importante observar que embora RcP seja um conjunto, o mesmo possui elementos repetidos, diferente do utilizado na teoria dos conjuntos. Desse modo, nas operações de conjunto abaixo utilizadas, os conjuntos resultantes podem conter, quando for o caso, algumas regras repetidas.

Questão 1 Em relação a "Questão 1", referente a sobreposição existente entre as regras em RcP e RcNP, quatro índices foram propostos, os quais são descritos a seguir:

I_{A-RcP} (Equação 10): mede a proporção de regras "antigas" em RcP, i.e., a proporção de regras em RcNP existentes em RcP. Considera-se uma regra como "antiga" aquela existente em RcNP, ou seja, contida no conjunto de regras obtido a partir de um conjunto de dados não particionado. Desse modo, quanto maior for o índice melhor, uma vez que o valor indica que não houve perda de conhecimento durante o processo.

$$I_{A-RcP} = \frac{|RcNP \cap RcP|}{|RcNP|} \tag{10}$$

I_{N-RcP} (Equação 11): mede a proporção de regras "'novas" em RcP, i.e., a proporção de regras em RcP não existentes em RcNP. Considera-se uma regra

como "nova" aquela não existente em RcNP, ou seja, não contida no conjunto de regras obtido a partir de um conjunto de dados não particionado. Embora seja importante que não haja perda de conhecimento (índice I_{A-RcP}), espera-se que a proporção de regras "novas" em RcP seja maior do que a de "antigas". Desse modo, quanto maior for o índice melhor, uma vez que o valor indica a quantidade de conhecimento, até então desconhecido, obtida durante o processo.

$$I_{N-RcP} = \frac{|RcP - RcNP|}{|RcP|} \tag{11}$$

$I_{R-A-RcP}$ (Equação 12): mede a proporção de regras "antigas" em RcP que se repetem. Considera-se que uma regra deve existir em apenas um dos grupos do agrupamento, uma vez que a mesma deverá estar contida em um sub-domínio (grupo) que expresse sua próprias associações. Desse modo, quanto menor for o índice melhor, uma vez que o valor indica que os conhecimentos, já conhecidos, estão contidos em sub-domínios que expressam suas próprias associações.

$$I_{R-A-RcP} = \frac{ComputaRepeticaoRcP(RcNP \cap RcP)}{|RcNP \cap RcP|}, em\ que \tag{12}$$

ComputaRepeticaoRcP: função que recebe por parâmetro um conjunto de regras sem repetições e retorna quantas das regras do conjunto se repetem em RcP

$I_{R-N-RcP}$ (Equação 13): mede a proporção de regras "novas" em RcP que se repetem. Idem ao índice $I_{R-A-RcP}$. Desse modo, assim como em $I_{R-A-RcP}$, quanto menor for o índice melhor, uma vez que o valor indica que os conhecimentos, até então desconhecidos, estão contidos em sub-domínios que expressam suas próprias associações.

$$I_{R-N-RcP} = \frac{EncontraRepeticao(RcP - RcNP)}{|RcP - RcNP|}, em\ que \tag{13}$$

EncontraRepeticao: função que recebe por parâmetro um conjunto de regras com repetições e retorna quantas das regras do conjunto se repetem

Questão 2 Em relação a "Questão 2", referente a sobreposição existente entre as regras em RcNP e RcP em relação a interessabilidade do conhecimento, quatro índices foram propostos, os quais são descritos a seguir:

$I_{N-I-RcP}$ (Equação 14): mede a proporção de regras "novas" entre as h-top regras

interessantes contidas em RcP. Dado um subconjunto de h-top regras interessantes, selecionadas de RcP, espera-se que a proporção de regras "novas" nesse subconjunto seja a maior possível. Assim como na PAR-COM, as h-top regras são as h regras que contêm os maiores valores referentes a uma medida objetiva, onde h é um número a ser escolhido. Desse modo, quanto maior for o índice melhor, uma vez que o valor indica que o custo do processo é compensado pela descoberta de conhecimentos interessantes antes não encontrados em RcNP.

$$I_{N-I-RcP} = \frac{ComputaQuantidade(h_{top}\ de\ RcP, RcP - RcNP)}{|h_{top}\ de\ RcP|}, em\ que \quad (14)$$

ComputaQuantidade: função que recebe por parâmetro um conjunto de h-top regras interessantes e um conjunto de regras e retorna quantas das regras do conjunto aparecem dentre as h-top

$I_{A-I-N-RcP}$ (Equação 15): mede a proporção de regras "antigas" não existentes em RcP entre as h-top regras interessantes contidas em RcNP. Dado um subconjunto de h-top regras interessantes, selecionadas de RcNP, espera-se que todas essas regras existam em RcP, uma vez que não é desejável que os padrões interessantes em RcNP desapareçam em RcP, o que caracterizaria a perda de conhecimento relevante. Assim, esse índice mede a proporção de regras "antigas" interessantes não existentes em RcP. Assim como na PAR-COM, as h-top regras são as h regras que contêm os maiores valores referentes a uma medida objetiva, onde h é um número a ser escolhido. Desse modo, quanto menor for o índice melhor, uma vez que o valor indica que os conhecimentos interessantes em RcNP não foram perdidos durante o processo.

$$I_{A-I-N-RcP} = \frac{ComputaQuantidade(h_{top}\ de\ RcNP, RcNP - RcP)}{|h_{top}\ de\ RcNP|}, em\ que \quad (15)$$

ComputaQuantidade: função que recebe por parâmetro um conjunto de h-top regras interessantes e um conjunto de regras e retorna quantas das regras do conjunto aparecem entre as h-top (idem Equação 14)

I_{C-I} (Equação 16): mede a proporção de regras em comum entre as h-top regras interessantes em RcP e as h-top regras interessantes em RcNP. Considere dois subconjuntos, S_1 e S_2, contendo, respectivamente, as h-top regras interessantes em RcP e as h-top regras interessantes em RcNP. Esse índice mede a inter-

secção existentes entre esses dois subconjuntos, a qual espera-se que seja a menor possível. Desse modo, quanto menor for o índice melhor. Quanto maior for a intersecção, menos relevante será o processo, uma vez que todo o conhecimento já conhecido como interessante em RcNP é também identificado como interessante em RcP, não fornecendo, portanto, ao processo nenhuma informação relevante adicional.

$$I_{C-I} = \frac{|h_{top} \text{ de } RcP \cap h_{top} \text{ de } RcNP|}{h}, em\ que \tag{16}$$

h é o número escolhido para realizar a seleção das regras em ambos os conjuntos, i.e., RcP e RcNP

$I_{NC-I-RcP}$ (Equação 17): mede a proporção de grupos do agrupamento gerador de RcP que contém as h-top regras interessantes contidas em RcP. Desse modo, quanto menor for o índice melhor. Isso porque, nesses casos, assim como na PAR-COM, apenas alguns dos grupos teriam de ser explorados pelo usuário, os quais conteriam os "novos" conhecimentos relevantes extraídos durante o processo.

$$I_{NC-I-RcP} = \frac{EncontraGrupos(h_{top} \text{ de } RcP)}{N}, em\ que \tag{17}$$

N: quantidade de grupos do agrupamento; Encontra Grupos: função que recebe por parâmetro um conjunto de h-top regras interessantes, encontra seus respectivos grupos e retorna a quantidade de grupos selecionados

Questão 3 Em relação a "Questão 3", referente ao comportamento do processo em relação ao número de regras obtidas em RcP, um índice foi proposto, o qual é descrito a seguir:

I_{NR-RcP} (Equação 18): mede a proporção de regras obtidas em RcP em relação as obtidas em RcNP. É importante que se verifique o comportamento do processo em relação ao número de regras obtidas em RcP. Não é desejável que haja um aumento muito grande no volume de regras, pois se por um lado novos padrões serão descobertos, mais difícil será para o usuário identificá-los. Desse modo, quanto menor for o índice melhor, uma vez que o valor indica que embora novos padrões tenham sido obtidos, que a quantidade encontrada não é grande o suficiente a ponto de sobrecarregar o usuário com um volume excessivo de

regras.

$$I_{NR-RcP} = \frac{|RcP|}{|RcNP|} \tag{18}$$

Por fim, a Tabela 26 resume os índices propostos. A fim de avaliar a abordagem de pré-processamento por meio dos índices acima descritos, experimentos foram realizados, os quais são apresentados e discutidos a seguir (Seções 4.2.1 e 4.2.2).

Tabela 26: Síntese dos índices propostos.

Índice	Descrição
I_{A-RcP}	Proporção de regras "antigas" em RcP
I_{N-RcP}	Proporção de regras "'novas" em RcP
$I_{R-A-RcP}$	Proporção de regras "antigas" em RcP que se repetem
$I_{R-N-RcP}$	Proporção de regras "novas" em RcP que se repetem
$I_{N-I-RcP}$	Proporção de regras "novas" entre as h-top regras interessantes contidas em RcP
$I_{A-I-N-RcP}$	Proporção de regras "antigas" não existentes em RcP entre as h-top regras interessantes contidas em RcNP
I_{C-I}	Proporção de regras em comum entre as h-top regras interessantes em RcP e as h-top regras interessantes em RcNP
$I_{NC-I-RcP}$	Proporção de grupos do agrupamento gerador de RcP que contém as h-top regras interessantes contidas em RcP
I_{NR-RcP}	Proporção de regras obtidas em RcP em relação as obtidas em RcNP

4.2.1 Experimentos

Visando avaliar a abordagem de agrupamento na etapa de pré-processamento, segundo os índices acima descritos, experimentos foram realizados. Uma vez que vários desses índices baseiam-se em comparações entre as regras obtidas por meio do processo tradicional (RcNP) e as obtidas via pré-processamento (RcP), os mesmos conjuntos de regras da Seção 3.1.1 (página 14) poderiam ser utilizados para representar os RcNP. Contudo, a fim de manter o mesmo padrão de execução em todos os experimentos aqui apresentados, os quatro conjuntos de dados considerados foram novamente minerados de modo que os suportes mínimos fossem obtidos como descrito na Seção 4.1. Em relação aos valores de confiança, os mesmos foram obtidos empiricamente após vários execuções do *Apriori*. Assim, por meio do processo tradicional, 5816 regras foram extraídas a partir do conjunto Adult amostral (vide Seção 4.1) usando sup-min de 10.89% e conf-min de 50%; 554 regras a partir do Income considerando sup-min=28% e conf-min=50%; 64 regras a partir do Groceries considerando sup-min=2.61% e conf-min=10%; 295 regras a partir do Sup considerando sup-min=0.55% e conf-min=100%. É importante mencionar que esses mesmos valores de confiança, referentes a cada conjunto de dados, foram utilizados na obtenção dos RcP. Em relação a escolha dos algoritmos, das medidas de similaridade e do intervalo de k, tomou-se como base os resultados apresentados na Seção 4.1. Além disso, como alguns dos índices levam em consideração as h-top regras interessantes de um

dado conjunto, uma medida objetiva deveria ser selecionada. Entretanto, assim como no pós-processamento, uma vez que espera-se que qualquer medida produza bons resultados, em vez de se escolher uma medida específica considerou-se, como na Seção 3.1.1, a média do ranqueamento obtida por meio das mesmas 18 medidas objetivas apresentadas na Tabela 2 (página 17). Nesse caso, o h foi definido para assumir, em todos os conjuntos (RcNP e RcP), 0.5% do total de regras contidas em RcNP (já que o número de regras é menor em relação a RcP (vide Seção 4.2.2)) de modo a tornar a comparação justa. Assim, cada conjunto de regras contém seus próprios valores que são proporcionais em todos eles. A Tabela 27 apresenta as configurações usadas nos experimentos. A partir dessas configurações, os experimentos foram executados, segundo o processo da Figura 9, e os índices computados para cada uma das diferentes organizações, cujos resultados encontram-se apresentados na próxima seção.

Tabela 27: Configurações usadas para avaliar a abordagem de agrupamento no pré-processamento segundo os índices propostos.

Conjuntos de Dados	Adult [conf-min 50%]; Income [conf-min 50%]; Groceries [conf-min 10%]; Sup [conf-min 100%]
Algoritmos	PAM; Ward Linkage
Medidas de Similaridade	Agg; Denza
k	5 a 25, passos de 5
h	0.5% do total de regras contidas em RcNP
Medidas Objetivas	Média dos ranqueamentos das medidas Added Value, Certainty Factor, Collective Strength, Confidence, Conviction, IS, ϕ-coefficient, Gini Index, J-Measure, Kappa, Klosgen, λ, Laplace, Lift, Mutual Information (assimétrica), Novelty, Support, Odds Ratio

4.2.2 Resultados e Discussões

Considerando as configurações apresentadas na Tabela 27 e os RcNP acima discutidos, os experimentos foram executados, segundo o processo da Figura 9, e os valores obtidos em cada um dos índices encontram-se apresentados nas Tabelas 28 a 35. Os resultados foram agrupados por algoritmo para cada conjunto de dados. As Tabelas 28 e 32 apresentam os resultados para o conjunto Adult, as Tabelas 29 e 33 para o Income, as Tabelas 30 e 34 para o Groceries e as Tabelas 31 e 35 para o Sup. Cada tabela apresenta os resultados de ambas as medidas consideradas, i.e., Agg e Denza. Para cada medida são apresentados os valores de cada um dos índices. Cada linha apresenta o desempenho da organização para um dado valor de k. A fim de facilitar a interpretação dos resultados, todos os índices tendendo para um desempenho a menor foram processados de modo a armazenar o complemento da informação. Assim, todos os índices, com exceção de I_{NR-RcP}, apresentam a mesma interpretação: quanto maior melhor. Em relação ao índice I_{NR-RcP}, o mesmo mantém

a mesma interpretação descrita anteriormente, quanto menor melhor, representando a proporção de regras obtidas em RcP em relação as obtidas em RcNP. Além disso, todos os índices podem ser mensurados em porcentagem se multiplicados por 100 (ex.: 0.779*100 = 77.9%).

De modo a auxiliar a interpretação das tabelas, considere a organização PAM:Agg da Tabela 28. Pode-se observar, para $k = 5$, que a organização apresenta um I_{A-RcP} de 0.779, o que significa que 77.9% (0.779*100) das regras em RcNP encontram-se em RcP, havendo, portanto, uma perda de apenas 22.1% (100%-77.9%) do conhecimento contido em RcNP.

Tabela 28: Resultados do PAM, segundo os índices propostos, no conjunto de dados Adult.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.779	0.794	0.478	0.915	0.547	0.833	0.738	0.400	7.243
10	0.796	0.843	0.303	0.883	0.589	0.826	0.785	0.556	12.578
15	0.819	0.845	0.313	0.885	0.602	0.847	0.790	0.600	13.767
20	0.820	0.843	0.311	0.884	0.601	0.847	0.788	0.600	13.752
25	0.820	0.843	0.311	0.884	0.601	0.847	0.788	0.600	13.768
	Denza								
5	0.584	0.836	0.418	0.925	0.781	0.361	0.821	0.200	7.096
10	0.556	0.881	0.279	0.866	0.864	0.267	0.892	0.600	13.439
15	0.605	0.901	0.341	0.847	0.804	0.387	0.892	0.467	20.247
20	0.557	0.910	0.275	0.840	0.843	0.267	0.907	0.550	27.301
25	0.620	0.914	0.259	0.839	0.826	0.384	0.921	0.600	34.216

Tabela 29: Resultados do PAM, segundo os índices propostos, no conjunto de dados Income.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.866	0.970	0.208	0.937	0.745	0.964	0.800	0.400	63.229
10	0.935	0.979	0.180	0.886	0.764	0.964	0.873	0.600	160.265
15	0.939	0.980	0.077	0.865	0.782	0.964	0.855	0.600	238.332
20	0.890	0.983	0.108	0.846	0.982	0.873	1.000	0.800	324.162
25	0.917	0.984	0.098	0.822	0.945	0.945	0.964	0.800	406.605
	Denza								
5	0.872	0.975	0.346	0.958	0.818	0.964	0.891	0.200	66.955
10	0.832	0.982	0.130	0.890	0.945	0.727	0.945	0.700	176.856
15	0.883	0.986	0.131	0.894	0.927	0.818	0.927	0.667	293.682
20	0.883	0.987	0.100	0.871	0.964	0.782	0.964	0.850	423.374
25	0.912	0.987	0.075	0.856	0.964	0.873	0.964	0.800	542.173

Tabela 30: Resultados do PAM, segundo os índices propostos, no conjunto de dados Groceries.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.922	0.705	0.729	0.995	0.500	1.000	1.000	0.800	4.031
10	0.984	0.804	0.492	0.952	0.333	1.000	1.000	0.900	8.859
15	1.000	0.829	0.359	0.944	0.500	1.000	1.000	0.933	12.031
20	1.000	0.847	0.281	0.928	0.500	1.000	1.000	0.950	15.172
25	1.000	0.850	0.281	0.924	0.667	1.000	1.000	0.960	16.094
	Denza								
5	1.000	0.661	0.188	0.912	0.333	1.000	0.833	0.400	6.172
10	1.000	0.781	0.188	0.890	0.500	1.000	1.000	0.600	11.063
15	1.000	0.825	0.031	0.881	0.500	1.000	0.833	0.800	19.578
20	1.000	0.863	0.078	0.857	0.667	1.000	1.000	0.850	24.344
25	1.000	0.860	0.016	0.862	0.833	1.000	1.000	0.920	29.734

Tabela 31: Resultados do PAM, segundo os índices propostos, no conjunto de dados Sup.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.827	1.000	1.000	1.000	1.000	1.000	1.000	0.800	1998.922
10	0.756	0.994	0.991	0.999	1.000	1.000	1.000	0.900	119.905
15	0.766	0.996	0.987	0.999	1.000	1.000	1.000	0.933	175.329
20	0.783	0.995	0.987	0.999	1.000	1.000	1.000	0.950	156.098
25	0.759	0.996	0.987	0.999	1.000	1.000	1.000	0.960	191.854
	Denza								
5	0.878	0.872	1.000	0.998	0.621	1.000	0.793	0.600	6.834
10	0.942	1.000	1.000	1.000	1.000	1.000	1.000	0.900	1990.492
15	0.936	1.000	0.993	1.000	1.000	1.000	1.000	0.867	2005.986
20	0.746	0.992	0.805	0.999	1.000	0.931	1.000	0.950	110.919
25	0.742	0.992	0.804	0.999	1.000	0.931	1.000	0.960	111.064

Tabela 32: Resultados do Ward, segundo os índices propostos, no conjunto de dados Adult.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.562	0.818	0.363	0.888	0.917	0.139	0.988	0.200	6.579
10	0.559	0.862	0.305	0.863	0.907	0.136	0.997	0.400	11.681
15	0.554	0.876	0.232	0.842	0.954	0.157	1.000	0.400	17.515
20	0.569	0.884	0.175	0.840	0.954	0.158	1.000	0.450	23.037
25	0.583	0.896	0.188	0.834	0.928	0.162	0.979	0.520	29.124
	Denza								
5	0.522	0.825	0.340	0.909	0.816	0.117	0.981	0.000	6.770
10	0.512	0.881	0.266	0.843	0.941	0.117	0.986	0.300	13.251
15	0.578	0.890	0.323	0.836	0.861	0.301	0.876	0.533	18.093
20	0.602	0.895	0.282	0.839	0.862	0.301	0.878	0.650	23.810
25	0.614	0.898	0.282	0.841	0.871	0.348	0.883	0.440	28.363

Tabela 33: Resultados do Ward, segundo os índices propostos, no conjunto de dados Income.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	0.729	0.972	0.327	0.959	0.909	0.600	0.982	0.400	54.540
10	0.892	0.978	0.178	0.890	0.818	0.818	0.855	0.700	147.175
15	0.903	0.982	0.172	0.875	0.945	0.836	0.982	0.600	235.693
20	0.924	0.983	0.119	0.862	0.909	0.891	0.927	0.600	337.733
25	0.910	0.984	0.069	0.855	0.964	0.764	0.964	0.760	433.735
	Denza								
5	0.767	0.967	0.292	0.964	0.873	0.618	0.909	0.400	51.157
10	0.881	0.979	0.168	0.915	0.891	0.782	0.891	0.700	153.977
15	0.906	0.980	0.064	0.891	0.782	0.836	0.891	0.667	246.726
20	0.912	0.982	0.046	0.866	0.800	0.836	0.891	0.800	353.552
25	0.928	0.983	0.029	0.867	1.000	0.836	1.000	0.840	444.238

Tabela 34: Resultados do Ward, segundo os índices propostos, no conjunto de dados Groceries.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	1.000	0.030	1.000	1.000	0.000	1.000	0.000	0.800	1.031
10	1.000	0.030	1.000	1.000	0.000	1.000	0.000	0.889	1.031
15	1.000	0.030	1.000	1.000	0.000	1.000	0.000	0.889	1.031
20	1.000	0.514	0.734	1.000	0.833	1.000	0.833	0.667	2.734
25	1.000	0.621	0.734	0.986	1.000	1.000	1.000	0.714	3.500
	Denza								
5	1.000	0.030	1.000	1.000	0.000	1.000	0.000	0.800	1.031
10	1.000	0.516	0.781	1.000	0.333	1.000	0.500	0.900	2.516
15	1.000	0.652	0.578	0.989	0.667	1.000	1.000	0.857	4.531
20	1.000	0.730	0.141	0.892	0.833	1.000	1.000	0.833	9.359
25	1.000	0.764	0.047	0.865	1.000	1.000	1.000	0.857	12.453

Tabela 35: Resultados do Ward, segundo os índices propostos, no conjunto de dados Sup.

	Agg								
k	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
5	1.000	0.010	1.000	1.000	0.000	1.000	0.103	0.800	1.010
10	1.000	0.010	1.000	1.000	0.000	1.000	0.103	0.857	1.010
15	0.868	0.105	1.000	1.000	0.000	0.966	0.310	0.875	0.969
20	0.946	0.549	1.000	1.000	0.103	1.000	0.276	0.889	2.095
25	0.946	0.551	1.000	1.000	0.172	1.000	0.345	0.900	2.108
	Denza								
5	0.969	0.662	1.000	1.000	0.241	1.000	0.379	0.800	2.868
10	0.969	0.857	1.000	1.000	0.655	1.000	0.655	0.875	6.790
15	0.969	0.860	1.000	0.998	0.655	1.000	0.655	0.909	6.915
20	0.969	0.860	1.000	0.998	0.655	1.000	0.655	0.909	6.908
25	0.851	1.000	1.000	1.000	1.000	0.966	1.000	0.933	2165.797

Com base nesses dados, de modo a identificar, inicialmente, as organizações mais adequadas segundo os índices propostos e, posteriormente, a contribuição do agrupamento no pré-processamento, uma análise baseada na média de cada um dos índices, considerando os diferentes tipos de dado, independente do conjunto, foi realizada. A Tabela 36 apresenta os resultados. Cada média foi obtida a partir dos resultados dos experimentos relacionados a configuração apresentada. O valor I_{A-RcP} de 0.858 da configuração CD-P:PAM:Agg, por exemplo, foi obtido por meio da média dos valores de I_{A-RcP} das configurações PAM:Adult:Agg e PAM:Income:Agg das Tabelas 28 e 29. Para facilitar o entendimento, esses valores encontram-se sublinhados nas respectivas tabelas. As médias mais altas, referentes a cada um dos índices, encontram-se marcadas com "*" em cada algoritmo. A única exceção se refere ao índice I_{NR-RcP}, em que se destacam as médias mais baixas. Para a configuração CD-P:PAM, por exemplo, a média mais representativa para I_{A-RcP} é a relacionada a medida Agg (0.858). Entretanto, uma vez que as médias são, em geral, próximas, uma marcação baseada na diferença existente entre as médias também foi considerada. Os valores marcados com "**" indicam que a diferença existente entre as médias de um dado índice encontra-se acima de 0.1 ($dif. \geq 0.1$). Para a configuração CD-P:PAM, por exemplo, a média mais representativa para I_{A-RcP} é a relacionada a Agg (0.858), exibindo uma diferença de 0.128 em relação a Denza (0.858-0.730). Desse modo, é possível visualizar, para cada configuração tipo-de-dado+algoritmo, a medida de similaridade que apresenta o melhor desempenho. Por fim, é importante mencionar que os resultados são determinísticos e, portanto, nenhum teste estatístico precisa realizado para checar se existe diferença significativa entre as médias. Pode-se observar que:

- em relação ao algoritmo PAM nos CD-P, a medida que se mostra mais adequada é a Agg, uma vez que a mesma se destaca em relação a Denza em 6 dos 9 índices. Além disso, em 3 dos 6 índices a Agg exibe uma diferença acima de 0.1 em relação a Denza. Nesses casos ($dif. \geq 0.1$), nota-se também que os valores em Agg tendem a ser mais representativos do que os obtidos em Denza – observe, por exemplo, que enquanto Agg em $I_{A-I-N-RcP}$ apresenta desempenho acima dos 89%, Denza fica abaixo dos 59%.

- em relação ao algoritmo Ward nos CD-P, embora a medida Denza se destaca em relação a Agg em 5 dos 9 índices, a medida Agg se mostra mais adequada mesmo se destacando em 4 dos 9 índices. Isso porque, enquanto a Agg exibe uma diferença acima de 0.1 em relação a Denza em 1 dos 4 índices, Denza não se destaca em relação a nenhum dos índices. Embora a diferença tenha ocorrido em apenas 1 dos índices, o mesmo é de grande importância, uma vez que ele mensura o quanto o espaço de exploração aumenta em relação a RcNP.

- em relação ao algoritmo PAM nos CD-NP, a medida que se mostra mais adequada é a Agg, uma vez que a mesma se destaca em relação a Denza em 8 dos 9 índices. Além disso, em 3 dos 8 índices a Agg exibe uma diferença acima de 0.1 em relação a Denza. Nesses casos ($dif. \geq 0.1$), nota-se também que os valores em Agg tendem a ser mais representativos do que os obtidos em Denza – observe, por exemplo, que enquanto Agg em $I_{R-A-RcP}$ apresenta desempenho acima dos 70%, Denza fica abaixo dos 52%.

- em relação ao algoritmo Ward nos CD-NP, ambas as medidas se destacam em 4 dos 9 índices, desconsiderando o empate ocorrido em $I_{A-I-N-RcP}$. Entretanto, em 3 dos 4 índices a Denza exibe uma diferença acima de 0.1 em relação a Agg. Nesses casos ($dif. \geq 0.1$), nota-se também que os valores em Denza tendem a ser mais representativos do que os obtidos em Agg – observe, por exemplo, que enquanto Denza em I_{N-RcP} apresenta desempenho acima dos 69%, Agg fica abaixo dos 25%. Portanto, a medida que se mostra mais adequada é a Denza. Por fim, observa-se que o único problema da medida em relação a Agg se refere ao índice I_{NR-RcP}.

Diante do exposto, pode-se notar que:

- em relação aos CD-P, a medida que se mostra mais adequada, independente do algoritmo, é a Agg. Assim, comparando os resultados das configurações PAM:Agg e Ward:Agg, observa-se que o PAM se destaca em relação ao Ward em 6 dos 9 índices (I_{A-RcP}, $I_{R-A-RcP}$, $I_{R-N-RcP}$, $I_{A-I-N-RcP}$, $I_{NC-I-RcP}$, I_{NR-RcP}), apresentando em 3 desses 6 (I_{A-RcP}, $I_{A-I-N-RcP}$, I_{NR-RcP}) uma diferença acima de 0.1. Portanto, a configuração que se mostra mais adequada, segundo os índices, é a PAM:Agg.

- em relação aos CD-NP, comparando os resultados das configurações PAM:Agg e Ward:Denza, observa-se que o PAM se destaca em relação ao Ward em 5 dos 9 índices (I_{N-RcP}, $I_{N-I-RcP}$, $I_{A-I-N-RcP}$, I_{C-I}, $I_{NC-I-RcP}$), desconsiderando o empate ocorrido em $I_{R-N-RcP}$, apresentando em 3 desses 5 (I_{N-RcP}, $I_{N-I-RcP}$, I_{C-I}) uma diferença acima de 0.1. Portanto, a configuração que se mostra mais adequada, segundo os índices, é a PAM:Agg.

Com base no apresentado, tem-se que a melhor configuração para abordagens de pré-processamento, segundo os índices utilizados, é a PAM:Agg. Tendo como base essa organização, torna-se possível discutir a contribuição do agrupamento na etapa de pré-processamento. A Tabela 37 apresenta os valores dos índices, na organização selecionada, em cada um dos tipos de dado. Esses valores são os mesmos das configurações CD-P:PAM:Agg e CD-NP:PAM:Agg da Tabela 36, porém em suas escalas originais, ou seja,

a maior ou a menor, já que as escalas a menor foram anteriormente convertidas a fim de facilitar a interpretação dos resultados. A indicação da escala, em cada índice, encontra-se entre "[]". Observe que os valores nas escalas a maior (↑) se mantém os mesmos e nas a menor (↓) os valores contêm o complemento dos valores anteriormente apresentados. Avaliando cada um dos índices nessa organização nota-se que:

- em relação a I_{A-RcP}, observa-se que há pouca perda de conhecimento durante o processo, por volta de 15%, uma vez que mais de 85% das regras em RcNP estão contidas em RcP. Assim, segundo esse critério, o processo parece ser adequado.

- em relação a I_{N-RcP}, observa-se que quase a totalidade das regras em RcP são "novas", por volta de 90%, indicando a obtenção de uma grande quantidade de conhecimento até então desconhecido. Assim, segundo esse critério, o processo parece adequado.

- em relação a $I_{R-A-RcP}$, observa-se que a repetição de regras "antigas" em RcP nos CD-NP é baixa, por volta dos 30%, o que não acontece nos CD-P, que fica em torno dos 77%. Assim, segundo esse critério, o processo não apresenta um padrão definido de desempenho.

- em relação a $I_{R-N-RcP}$, observa-se que a repetição de regras "novas" em RcP é baixa, por volta dos 12%, indicando que os conhecimentos, até então desconhecidos, estão contidos em sub-domínios que expressam suas próprias associações. Assim, segundo esse critério, o processo parece ser adequado.

- em relação a $I_{N-I-RcP}$, observa-se que grande parte das h-top regras interessantes em RcP são "novas", acima dos 71%, indicando que o custo do processo é compensado pela descoberta de conhecimentos interessantes antes não encontrados em RcNP. Assim, segundo esse critério, o processo parece adequado.

- em relação a $I_{A-I-N-RcP}$, observa-se que a perda de conhecimentos "antigos" e interessantes é baixa, por volta de 11%, uma vez que quase a totalidade das h-top regras interessantes em RcNP estão contidas em RcP, por volta de 89% (100%-11%). Assim, segundo esse critério, o processo parece adequado.

- em relação a I_{C-I}, observa-se que a intersecção existente entre as h-top regras interessantes em RcP e as h-top regras interessantes em RcNP é baixa, por volta de 17%, indicando que pouco do conhecimento já conhecido como interessante em RcNP continua se destacando em RcP. Assim, segundo esse critério, o processo parece adequado.

- em relação a $I_{NC-I-RcP}$, observa-se que a quantidade de grupos que contêm as h-top regras interessantes em RcP nos CD-NP é baixa, por volta de 10%, o que não acontece nos CD-P, que fica em torno dos 41%. Assim, segundo esse critério, o processo não apresenta um padrão definido de desempenho.

- em relação a I_{NR-RcP}, observa-se que a quantidade de regras obtidas em RcP é muito maior em relação as obtidas em RcNP, podendo sobrecarregar, se nenhum processo de apóio for utilizado, o usuário com um volume excessivo de regras. Assim, segundo esse critério, o processo não parece adequado.

Tabela 36: Média dos índices propostos nos diferentes tipos de dado nas diversas configurações consideradas (algoritmo+medida).

Tipo de Dado	Algoritmo	Medida	I_{A-RcP}	I_{N-RcP}	$I_{R-A-RcP}$	$I_{R-N-RcP}$	$I_{N-I-RcP}$	$I_{A-I-N-RcP}$	I_{C-I}	$I_{NC-I-RcP}$	I_{NR-RcP}
CD-P	PAM	Agg	0.858**	0.906	0.239*	0.881*	0.716	0.891**	0.838	0.596*	125.370**
		Denza	0.730	0.936*	0.235	0.878	0.874**	0.583	0.912*	0.563	160.534
	Ward	Agg	0.718	0.923	0.213*	0.871	0.920*	0.466	0.967*	0.503	129.681**
		Denza	0.722*	0.928*	0.209	0.877*	0.870	0.509*	0.919	0.533*	133.994
CD-NP	PAM	Agg	0.880	0.901*	0.709**	0.974*	0.750*	1.000*	1.000*	0.909**	269.830**
		Denza	0.924*	0.885	0.510	0.940	0.745	0.986	0.946	0.785	431.619
	Ward	Agg	0.976*	0.245	0.947**	0.999*	0.211	0.997*	0.297	0.828	1.652**
		Denza	0.973	0.693**	0.755	0.974	0.604**	0.997*	0.684**	0.867*	221.917

Tabela 37: Média dos índices propostos nos diferentes tipos de dado na organização PAM:Agg.

Tipo de Dado	I_{A-RcP} [↑]	I_{N-RcP} [↑]	$I_{R-A-RcP}$ [↓]	$I_{R-N-RcP}$ [↓]	$I_{N-I-RcP}$ [↑]	$I_{A-I-N-RcP}$ [↓]	I_{C-I} [↓]	$I_{NC-I-RcP}$ [↓]	I_{NR-RcP} [↓]
CD-P	0.858	0.906	0.761	0.119	0.716	0.109	0.162	0.404	125.370
CD-NP	0.880	0.901	0.291	0.026	0.750	0.000	0.000	0.091	269.830

Diante do exposto, o agrupamento na etapa de pré-processamento se mostra interessante, tendo apresentado bons valores em grande parte dos índices propostos (em 6 dos 9). Dos 3 índices que apresentaram valores ruins, destacam-se como preocupantes os índices $I_{NC-I-RcP}$ e I_{NR-RcP}, referentes ao espalhamento das h-top regras interessantes contidas em RcP pelos grupos dos agrupamentos e a quantidade de regras obtidas pelo processo. *Em decorrência desses fatores fica evidente a necessidade de abordagens de apóio a exploração dos resultados, uma vez que um grande de volume de regras é gerado e que os conhecimentos interessantes encontram-se espalhados ao longo dos grupos.* Assim, nas próximas seções, avaliam-se os aspectos "Redução via Interesse" e "Descritivo/Representativo" a fim de verificar se o agrupamento no pré-processamento se mostra realmente viável ou não.

4.3 Redução via Interesse

Como mencionado anteriormente, o agrupamento se torna útil ao usuário, assim como no pós-processamento, se de todos os grupos apresentados destacaram-se aqueles que contêm os padrões potencialmente mais interessantes, reduzindo assim o espaço de exploração. Essa vertente se torna ainda mais importante no pré-processamento, devido a grande quantidade de regras extraídas durante o processo. Uma vez que os resultados obtidos no pré-processamento serão comparados com o do pós-processamento na Seção 5, essa seção adota os mesmos procedimentos de avaliação da Seção 3.1.

A fim de explorar esse aspecto no pré-processamento, propõe-se aqui uma adaptação da metodologia PAR-COM (Seção 3.1, página 11). A Figura 14 apresenta a adaptação proposta, na qual as letras correspondem aos mesmos passos da PAR-COM (Figura 2, página 13). Observa-se que a diferença encontra-se no Passo C, não presente nessa nova proposta. Isso porque, como descrito e apresentado na Seção 1 (Figura 9), as regras já encontram-se agrupadas, uma vez que os grupos resultantes são vistos como integrantes de um único agrupamento. Portanto, a partir de um conjunto de regras já agrupado, nesse caso, RcP, a metodologia funciona como descrito a seguir:

Passo A: o valor de uma medida objetiva é computado para todas as regras em RcP.

Passo B: as h-top regras são selecionadas considerando os valores computados.

Passo C: não se aplica ao pré-processamento.

Passo D: uma busca é realizada para encontrar os grupos que contêm uma ou mais das h-top regras selecionadas no Passo B. Esses grupos são os que contêm o conhecimento potencialmente interessante do domínio (PIK: *P*otentially *I*nteresting *K*nowledge). Quanto mais h-top regras um grupo contiver mais interessante ele será.

Passo E: apenas os m primeiros grupos mais interessantes são apresentados ao usuário, o qual é direcionado para um espaço de exploração que contém o PIK do domínio, onde m é um valor a ser escolhido.

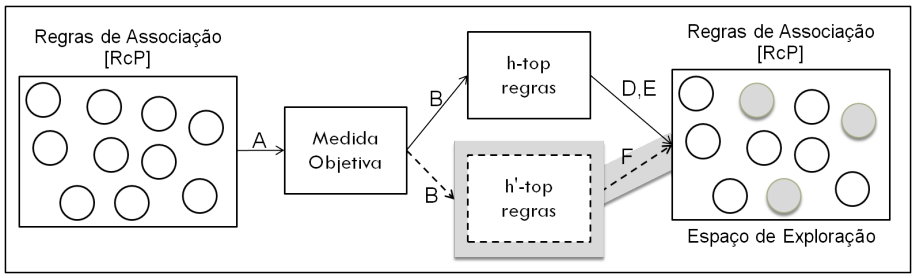


Figura 14: Adaptação da metodologia PAR-COM [Carvalho et al., 2011, Carvalho et al., 2012] para o pré-processamento.

O Passo F, destacado em cinza na Figura 14, é usado, como no pós-processamento, para validar a metodologia. Assim, da mesma maneira, considera-se todas as h'-top regras interessantes a serem selecionadas também no Passo B. Para esse propósito, uma busca é realizada para verificar se os m grupos selecionados contêm uma ou mais das h'-top regras. Desse modo, é possível verificar se o agrupamento no pré-processamento, assim como no pós-processamento, consegue reduzir o espaço de busca selecionando os grupos potencialmente interessantes.

4.3.1 Experimentos

Visando avaliar a abordagem de agrupamento na etapa de pré-processamento, segundo o aspecto "Redução via Interesse", experimentos foram realizados via PAR-COM adaptada. Para aplicar a metodologia foi necessário escolher uma medida objetiva (Passo A) e os valores de h (Passo B) e m (Passo E). Assim como no pós-processamento, o h foi definido para assumir 0.5% do total de regras do conjunto de modo a tornar a comparação entre os mesmos mais justa. O m foi definido como no máximo 3, ou seja, se apenas um grupo contiver todas as h-top regras somente ele será selecionado; se apenas dois grupos cobrirem todas as h-top regras somente os dois serão selecionados; e assim sucessivamente até no máximo 3 grupos. Além disso, em relação a medida objetiva, ao invés de se escolher uma medida específica, considerou-se também a média de ranqueamento obtida por meio das mesmas 18 medidas objetivas anteriormente utilizadas. Em relação aos algoritmos, medidas de similaridade e k, tendo como base os resultados previamente discutidos, a única organização considerada para obtenção dos RcPs foi a PAM:Agg com k entre 5 e 25 em passos de 5. Diante do apresentado, a Tabela 38 apresenta as configurações utilizadas

para aplicação da PAR-COM adaptada. Nota-se que os conjuntos de dados se mantêm os mesmos.

Tabela 38: Configurações usadas para aplicação da PAR-COM adaptada.

Conjuntos de Dados	Adult [conf-min 50%]; Income [conf-min 50%]; Groceries [conf-min 10%]; Sup [conf-min 100%]
Algoritmos	PAM
Medidas de Similaridade	Agg
k	5 a 25, passos de 5
h	0.5% do total de regras
m	≤ 3
Medidas Objetivas	Média dos ranqueamentos das medidas Added Value, Certainty Factor, Collective Strength, Confidence, Conviction, IS, ϕ-coefficient, Gini Index, J-Measure, Kappa, Klosgen, λ, Laplace, Lift, Mutual Information (assimétrica), Novelty, Support, Odds Ratio

É importante que se verifique também, no caso do pré-processamento, a proporção de regras "novas" entre as h-top, uma vez que uma baixa proporção indica a existência de uma grande quantidade de regras "antigas" e, consequentemente, que o processo não se mostra interessante.

4.3.2 Resultados e Discussão

Considerando as configurações apresentadas na Tabela 38, a PAR-COM adaptada foi aplicada e os resultados encontram-se apresentados nas Tabelas 39 e 40. A Tabela 39 apresenta os resultados para os conjuntos de dados Adult e Income e a Tabela 40 para o Groceries e Sup. Assim como no pós-processamento, cada coluna apresenta o resultado referente a um dado valor de k. Em relação as linhas tem-se que: (i) h-top e h'-top apresentam a porcentagem de h-top e h'-top regras contidas nos m primeiros clusters interessantes; (ii) R a porcentagem de redução do espaço de exploração; (iii) m a quantidade de clusters necessários para se recuperar as h-top regras ($m \leq 3$); (iv) diferentemente do pós-processamento, $\%RNovas$ a porcentagem de regras "novas" entre as h-top regras. A interpretação das tabelas é realizada da mesma maneira que no pós-processamento (Seção 3.1.2, página 17).

Tabela 39: Resultados da organização PAM:Agg, na PAR-COM adaptada, nos conjuntos de dados Adult e Income.

k	5	10	15	20	25
Adult					
h-top	100%	82%	81%	80%	80%
h'-top	100%	72%	62%	62%	62%
R	42%	63%	66%	66%	66%
m	3	3	3	3	3
$\%RNovas$	59%	80%	81%	81%	81%
Income					
h-top	86%	66%	65%	86%	78%
h'-top	65%	43%	78%	87%	66%
R	33%	76%	80%	79%	86%
m	3	3	3	3	3
$\%RNovas$	89%	96%	97%	99%	99%

Tabela 40: Resultados da organização PAM:Agg, na PAR-COM adaptada, nos conjuntos de dados Groceries e Sup.

k	5	10	15	20	25
Groceries					
h-top	100%	100%	100%	100%	100%
h'-top	100%	100%	100%	80%	67%
R	68%	78%	93%	93%	93%
m	1	1	1	1	1
$\%RNovas$	0%	0%	50%	60%	67%
Sup					
h-top	100%	100%	100%	100%	100%
h'-top	100%	100%	98%	100%	100%
R	1%	48%	65%	63%	66%
m	1	1	1	1	2
$\%RNovas$	100%	100%	100%	100%	100%

Assim como na Seção 3.1.2, computou-se também a média de cada critério (h, h', R, $RNovas$), considerando os diferentes tipos de dado, independente do conjunto. A Tabela 41 apresenta os resultados. É importante lembrar que cada média é obtida a partir dos resultados dos experimentos relacionados a configuração apresentada. Além disso, as Figuras 15 e 16 apresentam graficamente o comportamento da PAR-COM adaptada em cada um dos conjuntos. Diante do exposto, pode-se observar que (Tabelas 39, 40 e 41 e Figuras 15 e 16):

- em relação a h e h' o desempenho da PAR-COM adaptada nos CD-NP é melhor do que nos CD-P. Além disso, na maioria dos casos, com apenas 1 cluster recuperam-se essas h e h' regras, indicando a existência de grupos altamente interessantes. Contudo, ambos apresentam um bom desempenho.

- a redução do espaço de exploração fica, na média, em torno dos 65% em ambos os tipos de dado.

- em relação a proporção de regras interessantes e "novas", os CD-P apresentam um comportamento mais padronizado e, portanto, na média, melhor do que os CD-NP. Contudo, ambos apresentam um bom desempenho.

- os CD-P apresentam um comportamento mais padronizado em relação a todos os critérios do que os CD-NP, i.e., ambos os conjuntos de dados exibem comportamento semelhante nos critérios considerados.

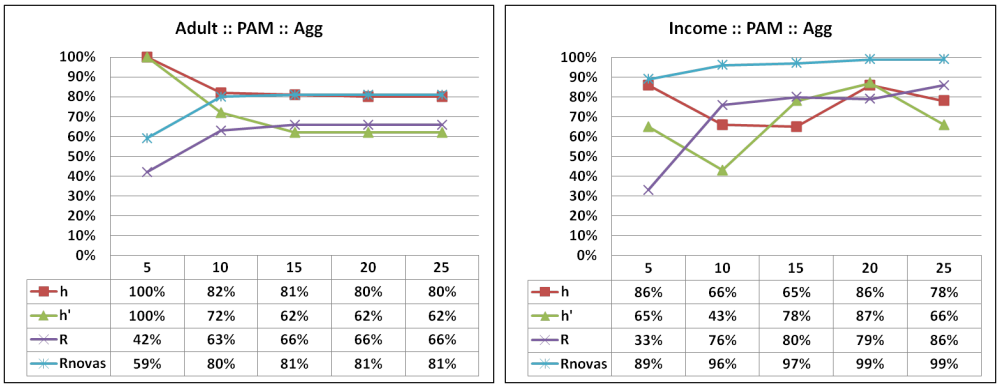
Portanto, considerando o apresentado, pode-se dizer que em ambos os tipos de dado, CD-P e CD-NP, a PAR-COM adaptada auxilia o usuário tanto na identificação de conhecimentos interessantes (h, h') quanto na redução do espaço de exploração (R). Desse modo, por meio da PAR-COM adaptada, é possível afirmar que o agrupamento no pré-processamento é útil ao usuário.

Tabela 41: Média de h, h', R e $RNovas$ nos diferentes tipos de dado na organização PAM:Agg.

Tipo de Dado	h	h'	R	RNovas
CD-P	80%	70%	66%	86%
CD-NP	100%	95%	67%	68%

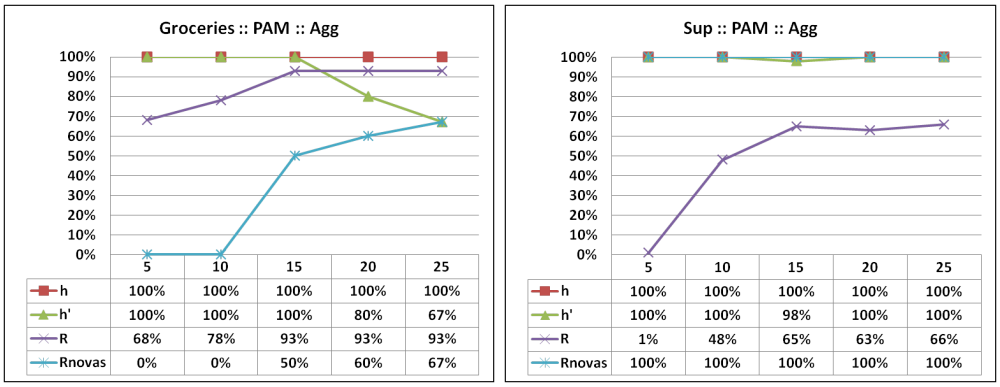
4.4 Descritivo/Representativo

Como mencionado anteriormente, o agrupamento se torna útil ao usuário, assim como no pós-processamento, se os grupos contiverem bons descritores que forneçam uma visão



(a) (b)

Figura 15: Comportamento da PAM-COM adaptada nos conjuntos de dados Adult e Income.



(a) (b)

Figura 16: Comportamento da PAM-COM adaptada nos conjuntos de dados Groceries e Sup.

dos tópicos a serem explorados. Desse modo, a fim de investigar esse aspecto no pré-processamento, essa seção analisa os mesmos métodos de rotulação empregados no pós-processamento de modo a identificar os mais adequados a serem utilizados em agrupamentos de regras de associação. Além disso, a avaliação desses métodos é realizada da mesma maneira, ou seja, por meio das medidas P e RF. É importante lembrar que o ideal é que os descritores de cada grupo representem o mais precisamente possível o conhecimento do seu próprio grupo (P (Seção 3.2.2)) e seja o mais diferente possível dos descritores dos outros grupos (RF (Seção 3.2.2)).

4.4.1 Experimentos

Visando avaliar a abordagem de agrupamento na etapa de pré-processamento, segundo o aspecto "Descritivo/Representativo", experimentos foram realizados. Para tanto, os mesmos métodos de rotulação empregados no pós-processamento foram utilizados, assim

como as mesmas medidas de avaliação (P e RF). Em relação aos algoritmos, medidas de similaridade e k, tendo como base os resultados previamente discutidos, a única organização considerada foi a PAM:Agg com k entre 5 e 25 em passos de 5. Diante do apresentado, a Tabela 42 apresenta as configurações usadas. Nota-se que os conjuntos de dados se mantêm os mesmos.

Tabela 42: Configurações usadas para avaliação dos métodos de rotulação no pré-processamento.

Conjuntos de Dados	Adult [conf-min 50%]; Income [conf-min 50%]; Groceries [conf-min 10%]; Sup [conf-min 100%]
Algoritmos	PAM
Medidas de Similaridade	Agg
k	5 a 25, passos de 5

Considerando as configurações da Tabela 42, os quatro métodos de rotulação (LM-M; LM-T; LM-S; LM-PU) foram aplicados à organização PAM:Agg. Em relação aos métodos de rotulação, assim como no pós-processamento, LM-M e LM-T selecionam apenas uma regra para compor os descritores, LM-S uma ou mais regras, em caso de empate, e LM-PU os 5 itens que apresentam o melhor equilíbrio entre frequência e predição. Assim, na média, todos os métodos de rotulação geram a mesma quantidade de descritores por grupo. Ao final, como mencionado, o desempenho de cada método de rotulação foi avaliado via P e RF, cujos resultados encontram-se apresentados na próxima seção. É importante lembrar que o ideal é identificar métodos que tenham altos valores para ambas as medidas. Por fim, é importante mencionar que, para que os descritores dos métodos LM-M e LM-T sejam computados, apenas as transações contidas em cada grupo são consideradas. No caso do LM-T, por exemplo, uma vez que o método encontra para cada grupo a regra que cobre o maior número de transações, apenas as transações contidas no grupo são levadas em consideração, uma vez que as regras são obtidas a partir de subconjuntos de transações diferentes. No caso do pós-processamento todas as transações são examinadas em todos os grupos, já que todas as regras são originadas do mesmo conjunto de transações.

4.4.2 Resultados e Discussão

O desempenho dos métodos de rotulação, avaliados via P e RF, encontram-se apresentados nas Tabelas 43 e 44. A Tabela 43 apresenta os resultados para os conjuntos de dados Adult e Income e a Tabela 44 para o Groceries e Sup. Cada tabela possui 8 colunas principais, cada 4 relacionada a um conjunto de dados. Cada uma dessas 4 colunas corresponde a um método de rotulação. Para cada método os valores de P e RF são apresentados. Cada linha apresenta o desempenho dos métodos para um dado valor de k. A interpretação das tabelas é realizada da mesma maneira que no pós-processamento (Seção 3.2.4, página 31).

Tabela 43: Resultados da organização PAM:Agg, nos LM's, nos conjuntos de dados Adult e Income.

	Adult								Income							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.920	0.360	0.670	0.600	0.720	0.000	0.910	0.560	0.830	0.640	0.680	0.860	0.630	0.000	0.830	0.860
10	0.920	0.530	0.660	0.700	0.720	0.000	0.920	0.500	0.750	0.450	0.660	0.830	0.610	0.000	0.860	0.290
15	0.930	0.270	0.660	0.670	0.710	0.000	0.910	0.540	0.750	0.420	0.660	0.560	0.640	0.000	0.850	0.150
20	0.930	0.270	0.660	0.670	0.710	0.000	0.910	0.540	0.710	0.230	0.660	0.380	0.650	0.000	0.840	0.190
25	0.930	0.270	0.660	0.670	0.710	0.000	0.910	0.540	0.690	0.190	0.650	0.250	0.640	0.000	0.830	0.170

Tabela 44: Resultados da organização PAM:Agg, nos LM's, nos conjuntos de dados Groceries e Sup.

	Groceries								Sup							
k	LM-M		LM-T		LM-S		LM-PU		LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF	P	RF
5	0.620	0.780	0.970	0.750	0.980	0.470	1.000	0.810	0.450	1.000	0.730	0.780	0.780	0.810	0.900	0.960
10	0.590	0.540	0.970	0.580	0.980	0.370	1.000	0.590	0.460	0.960	0.810	0.710	0.870	0.620	0.910	0.940
15	0.650	0.740	0.980	0.740	0.990	0.160	1.000	0.620	0.470	0.790	0.800	0.680	0.870	0.580	0.880	0.940
20	0.750	0.650	0.980	0.810	0.990	0.210	1.000	0.630	0.490	0.840	0.830	0.790	0.910	0.360	0.890	0.920
25	0.680	0.670	0.990	0.810	0.990	0.210	1.000	0.540	0.500	0.760	0.850	0.790	0.910	0.430	0.910	0.890

Assim como na Seção 3.2.4, computou-se também a média de cada medida (P, RF), considerando os diferentes tipos de dado, independente do conjunto. A Tabela 45 apresenta os resultados. É importante lembrar que cada média é obtida a partir dos resultados dos experimentos relacionados a configuração apresentada. Além disso, as médias mais altas, referentes a cada uma das medidas (P, RF), também encontram-se marcadas com "*" em cada configuração. A cada "*" destaca-se na tabela o par P/RF referente a marcação. A medida com "*" no método sublinhado indica o valor que fez com que o método fosse selecionado. Desse modo, é possível visualizar, para cada tipo de dado, o método que apresenta o melhor desempenho. Por fim, é importante mencionar que os resultados são determinísticos e, portanto, nenhum teste estatístico precisa ser realizado para checar se existe diferença significativa entre as médias. Pode-se observar que:

- em relação aos CD-P, o método que apresenta o melhor resultado em P é o LM-PU e em RF LM-T. Entretanto, uma vez que o valor de RF em LM-PU é baixo e o valor de P em LM-T não tão baixo, LM-T se torna mais indicado quando se busca um equilíbrio entre P e RF, uma vez que ele melhora RF enquanto não diminui drasticamente o valor de P.

- em relação aos CD-NP, o método que apresenta o melhor resultado tanto em P quanto em RF é o LM-PU, sendo, portanto, o mais indicado nesse caso. Porém, nota-se que o LM-T também apresenta bons resultados.

Tabela 45: Média de P e RF nos diferentes tipos de dado na organização PAM:Agg em cada LM.

Tipo de Dado	LM-M		LM-T		LM-S		LM-PU	
	P	RF	P	RF	P	RF	P	RF
CD-P	0.836	0.363	<u>0.662</u>	<u>0.619</u>*	0.674	0.000	<u>0.877</u>*	<u>0.434</u>
CD-NP	0.566	0.773	0.891	0.744	0.927	0.422	<u>0.949</u>*	<u>0.784</u>*

Considerando as discussões acima apresentadas, tem-se que: (i) nos CD-P o método que se mostra mais adequado para agrupamentos de regras de associação é o LM-T; (ii) nos CD-NP o método que se mostra mais adequado para agrupamentos de regras de associação é o LM-PU, embora o método LM-T também tenha apresentado um bom desempenho. As Figuras 17 e 18 apresentam o comportamento dos métodos LM-T e LM-PU, na organização PAM:Agg, tanto nos CD-P quanto nos CD-NP. É importante notar que embora P se mantenha estável em todos os valores de k o mesmo não ocorre com RF, cuja tendência é diminuir conforme o número de clusters aumenta. Entretanto, **nota-se que os descritores obtidos pelos métodos LM-T e LM-PU fornecem ao usuário uma visão dos tópicos a serem explorados com uma razoável precisão (P) e distinção (RF) nos CD-P e com uma boa precisão (P) e distinção (RF) nos**

CD-NP. Portanto, por meio desses dois métodos (LM-T, LM-PU), é possível afirmar que o agrupamento no pré-processamento é útil ao usuário.

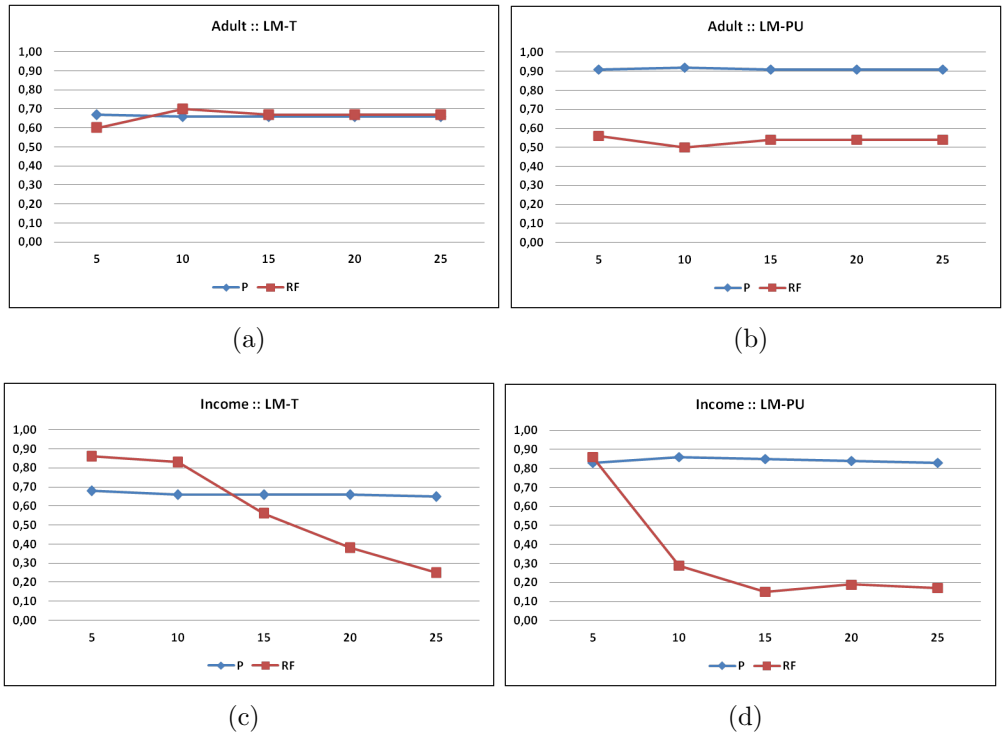


Figura 17: Comportamento dos métodos LM-T e LM-PU, na organização PAM:Agg, nos CD-P.

Por fim, é importante mencionar que, diferentemente do pós-processamento, nenhum dos trabalhos descritos na Seção 2.3.2 (página 7) cobrem o aspecto relacionado aos métodos de rotulação. Portanto, o estudo apresentado representa uma importante contribuição para o agrupamento de regras no pré-processamento. Além disso, assim como no pós-processamento, acredita-se, com base no que foi exposto, que outros métodos de *MT* e *RI* poderiam ser adaptados, uma vez que o LM-T e o LM-PU apresentaram um bom desempenho.

4.5 Considerações Finais

A Tabela 46 apresenta as conclusões obtidas nas Seções 4.2, 4.3 e 4.4 referentes as particularidades da abordagem e aos dois aspectos abordados: "Redução via Interesse" e "Descritivo/Representativo". Em relação as particularidades da abordagem, a organização que se destaca na etapa de pré-processamento, segundo os índices propostos, é a PAM:Agg. Tendo como base essa organização, uma vez que visa-se a integração dos dois aspectos, de modo a atender, simultaneamente, os requisitos (R-i) e (R-ii), tem-se que: (i) independente do tipo de dado utilizado (CD-P; CD-NP) a organização que se

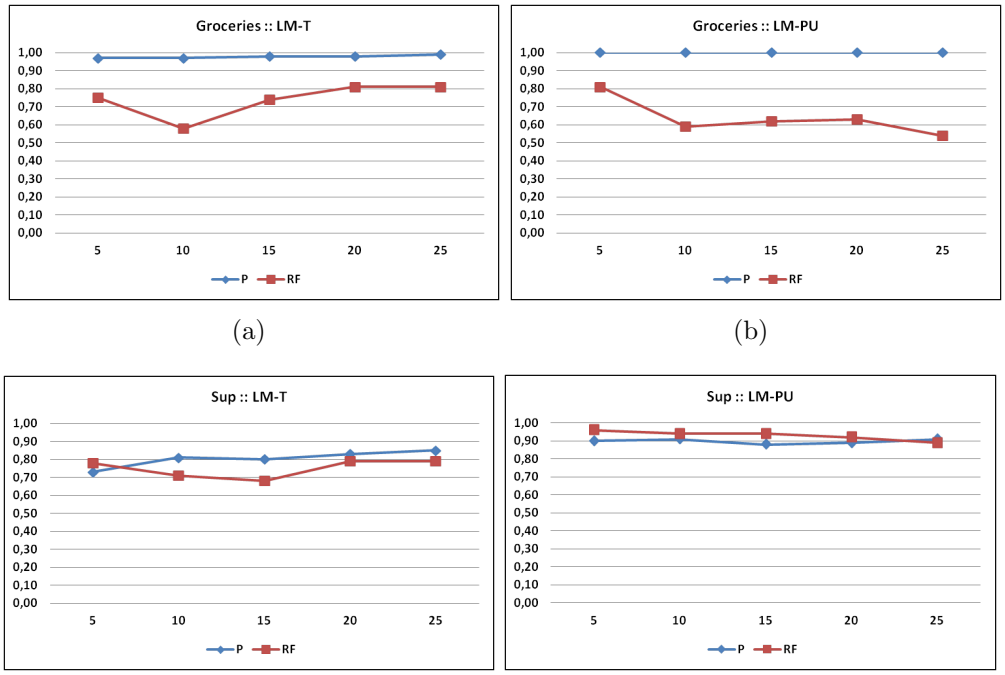


Figura 18: Comportamento dos métodos LM-T e LM-PU, na organização PAM:Agg, nos CD-NP.

destaca nos dois aspectos é a PAM:Agg, sendo, portanto, a mais indicada para agrupamentos de regras de associação na etapa de pré-processamento; (ii) em relação aos métodos de rotulação, tem-se que o LM-T é mais indicado para CD-P e o LM-PU para CD-NP. Além disso, diante do exposto anteriormente, tem-se que **o agrupamento de regras no pré-processamento possibilita que: (i) via PAR-COM adaptada, o espaço de exploração seja reduzido direcionando o usuário para o que é potencialmente interessante; (ii) via LM-T e LM-PU, bons descritores sejam obtidos de modo a fornecer uma visão dos tópicos a serem explorados. Desse modo, considerando o apresentado, pode-se afirmar que o agrupamento no pré-processamento apóia o usuário.**

Por fim, destacam-se as contribuições obtidas, por meio desse estudo, em relação ao agrupamento de regras de associação na etapa de pré-processamento:

- proposta de índices de avaliação referentes as particularidades do pré-processamento;
- identificação da viabilidade de se utilizar agrupamento na etapa de pré-processamento, uma vez que bons valores foram obtidos em grande parte dos índices de avaliação propostos (em 6 dos 9). Embora em 2 dos índices ($I_{NC-I-RcP}$ e I_{NR-RcP}), referentes ao espalhamento das h-top regras interessantes contidas em RcP pelos grupos dos agrupamentos e a quantidade de regras obtidas pelo processo, a abordagem

Tabela 46: Organizações e métodos de rotulação adequados ao pré-processamento referente as particularidades da abordagem e aos aspectos "Redução via Interesse" e "Descritivo/Representativo".

Organização indicada via Índices de Avaliação	
CD-P	PAM:Agg
CD-NP	PAM:Agg
"Redução via Interesse"	
CD-P	PAM:Agg
CD-NP	PAM:Agg
"Descritivo/Representativo"	
CD-P	LM-T
CD-NP	LM-PU

apresente valores preocupantes, a mesma se torna viável devido a utilização da PAR-COM adaptada, a qual visa suprir essas deficiências;

- proposta da metodologia PAR-COM adaptada;
- identificação da organização mais adequada para se reduzir o espaço de exploração, identificar o conhecimento potencialmente interessante e obter bons descritores;
- identificação dos métodos de rotulação mais adequados ao contexto apresentado;
- identificação da possibilidade de se integrar a PAR-COM adaptada com os métodos de rotulação adequados a cada tipo de dado de modo a permitir a identificação dos tópicos potencialmente mais interessantes do domínio;
- identificação da necessidade de valores de k não muito grandes a fim de garantir uma boa distribuição das transações nos grupos dos agrupamentos.

5 Avaliação Comparativa entre o Pós-processamento e o Pré-processamento

Essa seção apresenta uma comparação entre a utilização de agrupamento nas etapas de pós-processamento e pré-processamento a fim de responder a pergunta inicialmente levantada: o que vale mais a pena? Para tanto, a comparação foi divida em duas partes. Na primeira (Seção 5.1), a comparação é realizada objetivamente considerando os aspectos "Redução via Interesse" e "Descritivo/Representativo". Na segunda (Seção 5.2), a comparação é realizada subjetivamente por meio da participação de um especialista.

5.1 Comparação Objetiva: "Redução via Interesse" e "Descritivo/Representativo"

Para que a comparação entre o pós-processamento e o pré-processamento fosse realizada objetivamente, a mesma foi dividida em duas partes, cada qual visando um dos aspectos previamente abordados, ou seja, "Redução via Interesse" e "Descritivo/Representativo". As Tabelas 47 e 48 apresentam os resultados anteriormente obtidos nas Seções 3.1.2 (página 17), 3.2.4 (página 31), 4.3.2 (página 64) e 4.4.2 (página 68). Cada tabela exibe a etapa do processo em que o agrupamento foi realizado juntamente com a organização que obteve o melhor desempenho. A Tabela 47 se refere aos resultados, oriundos das Tabelas 12 (página 26) e 41 (página 66), obtidos via PAR-COM/PAR-COM adaptada em cada etapa, organização e tipo de dado. A Tabela 48 se refere aos resultados, oriundos das Tabelas 23 (página 39) e 45 (página 70), obtidos via P e RF nos métodos de rotulação que se mostraram mais adequados em cada etapa, organização e tipo de dado. Em cada uma dessas tabelas marcações foram realizadas. Na Tabela 47 as médias mais altas, referentes a cada um dos critérios (h, h', R), encontram-se marcadas com "*" em cada tipo de dado. Para os tipos de dado CD-P, por exemplo, a média mais representativa para h é a relacionada a etapa de pós-processamento (95%); já em relação a h' a etapa de pré-processamento (70%). Desse modo, é possível visualizar, para cada tipo de dado, a etapa mais adequada para aplicar-se o agrupamento segundo o aspecto "Redução via Interesse". Do mesmo modo, na Tabela 48 as médias mais altas, referentes a cada uma das medidas (P, RF), encontram-se marcadas com "*" em cada tipo de dado. Para os tipos de dado CD-P, por exemplo, a média mais representativa para P, em LM-T, é a relacionada a etapa de pós-processamento (0.929); já em relação a RF a etapa de pré-processamento (0.619). Desse modo, é possível visualizar, para cada tipo de dado, a etapa mais adequada para aplicar-se o agrupamento segundo o aspecto "Descritivo/Representativo". Por fim, é importante mencionar que os resultados são determinísticos e, portanto, nenhum teste es-

tatístico precisa ser realizado para checar se existe diferença significativa entre as médias. Pode-se observar que:

"Redução via Interesse" (Tabela 47)

- em relação aos CD-P, a etapa que apresenta o melhor resultado em h é a de pós, em h' a de pré e em R a de pós. Entretanto, considerando os três critérios simultaneamente, uma vez os valores de h' encontram-se muito próximos (67% x 70%), tem-se que a etapa de pós-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.
- em relação aos CD-NP, a etapa que apresenta o melhor resultado tanto em h quanto em h' é a de pré e em R a de pós. Entretanto, considerando os três critérios simultaneamente, uma vez os valores de R encontram-se muito próximos (71% x 67%), tem-se que a etapa de pré-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.

"Descritivo/Representativo" (Tabela 48)

- em relação aos CD-P, a etapa que apresenta o melhor resultado, em LM-T, em P é a de pós e em RF a de pré. Entretanto, considerando as duas medidas simultaneamente, uma vez os valores de RF encontram-se próximos (0.535 x 0.619 (diferença de 0.084)), tem-se que a etapa de pós-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.
- em relação aos CD-NP, a etapa que apresenta o melhor resultado, em LM-PU, em ambas as medidas é a de pré. Portanto, tem-se que a etapa de pré-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.

Tabela 47: Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Redução via Interesse".

Etapa	Organização	Tipo de Dado	h	h'	R
Pós	Ward:J-RT	CD-P	95%*	67%	87%*
		CD-NP	68%	28%	71%*
Pré	PAM:Agg	CD-P	80%	70%*	66%
		CD-NP	100%*	95%*	67%

Considerando o apresentado, indica-se a utilização de agrupamento na etapa de pós-processamento para os CD-P e na etapa de pré-processamento para os CD-NP, tanto no que se refere ao aspecto "Redução via Interesse" quanto ao "Descritivo/Representativo". É importante observar que os resultados das Tabelas 47 e 48 foram anteriormente computados considerando k entre 5-50 no pós-processamento e entre 5-25 no pré-processamento.

Tabela 48: Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Descritivo/Representativo".

Etapa	Organização	Tipo de Dado	LM-T		LM-PU	
			P	RF	P	RF
Pós	Ward:J-RT	CD-P	0.929*	0.535	-	-
		CD-NP	-	-	0.929	0.645
Pré	PAM:Agg	CD-P	0.662	0.619*	-	-
		CD-NP	-	-	0.949*	0.784*

Desse modo, a fim de garantir uma comparação mais justa entre as etapas e checar se as conclusões acima se mantêm, computou-se novamente as médias do pós-processamento considerando apenas k entre 5-25. As Tabelas 49 e 50 apresentam as novas médias, as quais foram computadas como as obtidas para k entre 5-50, porém considerando apenas os resultados obtidos até k=25. Note que no caso do pré-processamento os valores se mantêm iguais aos das Tabelas 47 e 48. O mesmo tipo de marcação das Tabelas 47 e 48 foi realizado. Assim, pode-se observar que:

"Redução via Interesse" (Tabela 49)

- em relação aos CD-P, a etapa que apresenta o melhor resultado em todos os critérios é a de pós. Portanto, tem-se que a etapa de pós-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.
- em relação aos CD-NP, a etapa que apresenta o melhor resultado em todos os critérios é a de pré. Portanto, tem-se que a etapa de pré-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.

"Descritivo/Representativo" (Tabela 50)

- em relação aos CD-P, a etapa que apresenta o melhor resultado, em LM-T, em ambas as medidas é a de pós. Portanto, tem-se que a etapa de pós-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.
- em relação aos CD-NP, a etapa que apresenta o melhor resultado, em LM-PU, em ambas as medidas é a de pré. Portanto, tem-se que a etapa de pré-processamento é a que se ajusta mais adequadamente a utilização de agrupamento.

Diante do exposto, observa-se o mesmo padrão anteriormente obtido, ou seja, **indica-se a utilização de agrupamento na etapa de pós-processamento para os CD-P e a etapa de pré-processamento para os CD-NP, tanto no que se refere ao aspecto "Redução via Interesse" quanto ao "Descritivo/Representativo"**. Além

Tabela 49: Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Redução via Interesse" e intervalo de k padronizado.

Etapa	Organização	Tipo de Dado	h	h'	R
Pós	Ward:J-RT	CD-P	96%*	77%*	81%*
		CD-NP	80%	36%	60%
Pré	PAM:Agg	CD-P	80%	70%	66%
		CD-NP	100%*	95%*	67%*

Tabela 50: Comparação objetiva das etapas de pós-processamento e pré-processamento sob o aspecto "Descritivo/Representativo" e intervalo de k padronizado.

Etapa	Organização	Tipo de Dado	LM-T		LM-PU	
			P	RF	P	RF
Pós	Ward:J-RT	CD-P	0.882*	0.689*	-	-
		CD-NP	-	-	0.902	0.675
Pré	PAM:Agg	CD-P	0.662	0.619	-	-
		CD-NP	-	-	0.949*	0.784*

disso, **observa-se um bom desempenho dos critérios h, h' e R e das medidas P e RF nos tipos de dado em suas respectivas etapas, i.e., pós-processamento nos CD-P e pré-processamento nos CD-NP**. Portanto, **pode-se afirmar que o agrupamento apóia o usuário em ambas as etapas considerando o tipo adequado de dado**.

5.2 Comparação Subjetiva: Análise Especialista

Para que a comparação entre o pós-processamento e o pré-processamento fosse realizada subjetivamente, selecionou-se para cada conjunto de dados dois agrupamentos: um obtido no pós e o outro no pré. A Tabela 51 apresenta as seleções realizadas em cada conjunto. Em relação ao pós, selecionou-se agrupamentos obtidos por meio da organização Ward:J-RT, já que a mesma é a mais indicada para essa etapa (Seção 3.3). Por outro lado, em relação ao pré, selecionou-se agrupamentos obtidos por meio da organização PAM:Agg, já que a mesma é a mais indicada para essa etapa (Seção 4.5). Em relação aos métodos de rotulação, selecionou-se para os CD-P o LM-T e para os CD-NP o LM-PU, por serem os mais adequados para cada um dos tipos de dado (Seções 3.3 e 4.5). Em todos os casos, considerou-se um valor de k=15. Essa escolha foi realizada avaliando os resultados anteriormente apresentados e discutidos. Além disso, esse valor representa o valor mediano do intervalo de k até então considerado (5, 10, **15**, 20, 25). A Tabela 51 apresenta os valores obtidos nos critérios h, h' e R em cada organização selecionada, assim como o desempenho do método de rotulação segundo P e RF. Indica-se também, juntamente com cada valor, a tabela onde a informação apresentada é encontrada.

Tabela 51: Agrupamentos selecionados para comparação subjetiva das abordagens de pós-processamento e pré-processamento.

Etapa	Tipo de Dado	Conjunto de Dado	Organização	Método de Rotulação	h	h'	R	P	RF
Pós	CD-P	Adult	Ward:J-RT [k=15]	LM-T	100% [Tab. 7]	91% [Tab. 7]	92% [Tab. 7]	0.956 [Tab. 18]	0.733 [Tab. 18]
		Income			89% [Tab. 8]	74% [Tab. 8]	80% [Tab. 8]	0.869 [Tab. 19]	0.733 [Tab. 19]
	CD-NP	Groceries		LM-PU	82% [Tab. 9]	18% [Tab. 9]	44% [Tab. 9]	0.958 [Tab. 20]	0.655 [Tab. 20]
		Sup			74% [Tab. 10]	45% [Tab. 10]	84% [Tab. 10]	0.863 [Tab. 21]	0.711 [Tab. 21]
Pré	CD-P	Adult	PAM:Agg [k=15]	LM-T	81% [Tab. 39]	62% [Tab. 39]	66% [Tab. 39]	0.660 [Tab. 43]	0.670 [Tab. 43]
		Income			65% [Tab. 39]	78% [Tab. 39]	80% [Tab. 39]	0.660 [Tab. 43]	0.560 [Tab. 43]
	CD-NP	Groceries		LM-PU	100% [Tab. 40]	100% [Tab. 40]	93% [Tab. 40]	1.000 [Tab. 44]	0.620 [Tab. 44]
		Sup			100% [Tab. 40]	98% [Tab. 40]	65% [Tab. 40]	0.880 [Tab. 44]	0.940 [Tab. 44]

Como observa-se na Tabela 51, o agrupamento apresenta um melhor desempenho no pós-processamento nos CD-P e no pré-processamento nos CD-NP, considerando h, h', R, P e RF, como já concluído na Seção 5.1. Desse modo, a fim de checar se as conclusões acima se mantêm, segundo análise feita por um especialista, a avaliação foi dividida em duas partes, cada qual visando um dos aspectos previamente abordados, ou seja, "Redução via Interesse" e "Descritivo/Representativo". O aspecto primeiramente abordado na análise foi o "Descritivo/Representativo", uma vez que o mesmo fornece uma visão geral do domínio, a partir do qual o processo de avaliação pode ser refinado. É importante mencionar que os experimentos abaixo descritos foram avaliados por um especialista em regras de associação, para o qual foram fornecidas informações sobre cada um dos conjuntos de dados utilizados.

A fim de avaliar os resultados segundo o aspecto "Descritivo/Representativo" apresentou-se ao especialista, para cada conjunto de dados, os descritores de seus agrupamentos, i.e., um do pós e um do pré. Os descritores encontram-se nas Tabelas 52 (Adult), 53 (Income), 54 (Groceries) e 55 (Sup). Cada tabela apresenta os descritores do agrupamento do pós, apresentados como "Agrupamento 1", e os descritores do agrupamento do pré, apresentados como "Agrupamento 2". O especialista não foi informado sobre qual dos agrupamentos pertencia a cada etapa. Observe que cada agrupamento é composto por um conjunto de grupos identificados pelo seu número (G#). Além disso, o único agrupamento que não contêm os 15 grupos é o obtido no pré no conjunto Adult – 5 grupos não são considerados segundo o critério de distribuição das transações (Seção 4.1). Para cada conjunto de dados, solicitou-se ao especialista que o mesmo indicasse:

(a) se os descritores, em cada agrupamento, estavam ou não adequados, i.e., se os descritores distribuíam adequadamente os tópicos do domínio e, consequente, o conhecimento do domínio. Ele poderia responder *sim*, *não* ou *mais ou menos*. Além disso, ele deveria justificar a resposta.

(b) dentro de cada conjunto de dados qual dos dois agrupamentos melhor distribuía os tópicos do domínio. Ele poderia responder "indiferente"caso não conseguisse concluir.

Tabela 52: Descritores dos agrupamentos selecionados para o conjunto de dados Adult.

Agrupamento 1				
G1 sex=Female capital-loss=None	**G2** marital-status=Never-married capital-loss=None	**G3** income=large native-country=United-States	**G4** education=Some-college capital-loss=None	**G5** relationship=Not-in-family capital-loss=None
G6 hours-per-week=Over-time capital-loss=None	**G7** age=Senior capital-loss=None	**G8** education=HS-grad capital-loss=None	**G9** marital-status=Married-civ-spouse capital-loss=None	**G10** marital-status=Married-civ-spouse age=Middle-aged
G11 marital-status=Married-civ-spouse hours-per-week=Full-time	**G12** capital-gain=None capital-loss=None	**G13** income=small capital-loss=None	**G14** education=HS-grad marital-status=Married-civ-spouse capital-loss=None	**G15** marital-status=Married-civ-spouse income=small capital-loss=None
Agrupamento 2				
G1 capital-loss=None sex=Female	**G2** capital-gain=None capital-loss=None	**G3** capital-gain=None capital-loss=None	**G4** marital-status=Divorced capital-loss=None	**G5** race=White hours-per-week=Over-time
G6 capital-gain=None age=Young	**G7** capital-gain=None capital-loss=None	**G8** capital-loss=None age=Senior	**G9** capital-gain=None capital-loss=None	**G10** age=Young marital-status=Never-married

(a)
Resposta "Agrupamento 1": *mais ou menos.* É possível identificar alguns segmentos como pessoas casadas, estudantes de determinado nível, etc. Há vários descritores com o mesmo item. Por exemplo, "capital-loss=None" aparece em G1, G2, G4, G5, G6, G7, G8, G9, G12, G13, G14, G15.
Resposta "Agrupamento 2": *não.* Há vários descritores com o mesmo item. Por exemplo, "capital-loss=None" aparece em G1, G2, G3, G4, G7, G8, G9. Há diversos grupos com os mesmos descritores (G2, G3, G7, G9).
(b)
Conclusão: *Agrupamento 1.*

Tabela 53: Descritores dos agrupamentos selecionados para o conjunto de dados Income.

Agrupamento 1				
G1 type_of_home=house income=0-40.000	**G2** income=40.000+ language_in_home=english	**G3** householder_status=rent language_in_home=English	**G4** number_of_children=1+ language_in_home=english	**G5** age=14-34 number_in_household=1
G6 marital_status=married language_in_home=english	**G7** age=35+ language_in_home=english	**G8** sex=male language_in_home=english	**G9** income=0-40.000 language_in_home=english	**G10** income=0-40.000 years_in_bay_area=10+
G11 income=0-40.000 ethnic_classification=white	**G12** sex=female income=0-40.000	**G13** years_in_bay_area=10+ language_in_home=english	**G14** education=no_college_graduate language_in_home=english	**G15** ethnic_classification=white language_in_home=english
Agrupamento 2				
G1 type_of_home=house marital_status=married	**G2** language_in_home=English dual_incomes=yes	**G3** dual_incomes=not_married age=14-34	**G4** householder_status=live_with_parents/family education=no_college_graduate	**G5** sex=male marital_status=married
G6 income=0-40.000 language_in_home=english	**G7** language_in_home=english sex=male	**G8** dual_incomes=no marital_status=married	**G9** language_in_home=English householder_status=own	**G10** language_in_home=English ethnic_classification=white
G11 marital_status=married sex=female	**G12** language_in_home=English householder_status=own	**G13** dual_incomes=yes marital_status=married	**G14** householder_status=rent type_of_home=apartment	**G15** householder_status=rent type_of_home=apartment

(a)
Resposta "Agrupamento 1": *mais ou menos*. É possível identificar alguns segmentos como sexo, idade e etnia. Há vários descritores com o mesmo item. Por exemplo, "language_in_home=english" aparece em G2, G3, G4, G6, G7, G8, G9, G13, G14, G15. Alguns grupos contêm descritores não informativos como G6 ("marital_status=married + language_in_home=english"), G8 ("sex=male + language_in_home=english").
Resposta "Agrupamento 2": *mais ou menos*. Os grupos G14 e G15 contêm os mesmos descritores. Alguns grupos contêm descritores não informativos como G5 ("sex=male + marital_status=married"), G11 ("marital_status=married + sex=female").
(b)
Conclusão: *Agrupamento 1*.

Tabela 54: Descritores dos agrupamentos selecionados para o conjunto de dados Groceries.

Agrupamento 1				
G1	**G2**	**G3**	**G4**	**G5**
brown_bread	frozen_vegetables	pork	curd	domestic_eggs
whole_milk	chicken	beef	whole_milk	whole_milk
bottled_beer	margarine	frankfurter	tropical_fruit	other_vegetables
napkins	whipped_sour_cream	whole_milk	other_vegetables	whipped_sour_cream
chocolate	rolls_buns	other_vegetables	yogurt	margarine
G6	**G7**	**G8**	**G9**	**G10**
butter	newspapers	fruit_vegetable_juice	pip_fruit	whipped_sour_cream
whole_milk	rolls_buns	other_vegetables	other_vegetables	yogurt
other_vegetables	yogurt	whole_milk	root_vegetables	other_vegetables
whipped_sour_cream	other_vegetables	yogurt	whole_milk	root_vegetables
tropical_fruit	root_vegetables	tropical_fruit	yogurt	whole_milk
G11	**G12**	**G13**	**G14**	**G15**
citrus_fruit	pastry	bottled_water	sausage	tropical_fruit
tropical_fruit	other_vegetables	soda	soda	root_vegetables
pip_fruit	rolls_buns	other_vegetables	shopping_bags	yogurt
other_vegetables	whole_milk	tropical_fruit	other_vegetables	whole_milk
root_vegetables	soda	rolls_buns	rolls_buns	other_vegetables
Agrupamento 2				
G1	**G2**	**G3**	**G4**	**G5**
tropical_fruit	whole_milk	Yogurt	other_vegetables	rolls_buns
pip_fruit	rolls_buns	whole_milk	whole_milk	sausage
citrus_fruit	ham	bottled_water	root_vegetables	shopping_bags
grapes	pastry	curd	yogurt	soda
other_vegetables	bottled_water	rolls_buns	whipped_sour_cream	frankfurter
G6	**G7**	**G8**	**G9**	**G10**
root_vegetables	Newspapers	Chocolate	Pastry	bottled_water
beef	chewing_gum	Waffles	Beverages	bottled_beer
herbs	chicken	specialty_bar	specialty_bar	soda
packaged_fruit_vegetables	oil	candy	hamburger_meat	fruit_vegetable_juice
onions	sugar	specialty_chocolate	specialty_chocolate	white_wine
G11	**G12**	**G13**	**G14**	**G15**
Soda	canned_beer	Coffee	Liquor	shopping_bags
shopping_bags	red_blush_wine	condensed_milk	bottled_beer	canned_beer
canned_beer	newspapers	chicken	red_blush_wine	seasonal_products
fruit_vegetable_juice	shopping_bags	berries	soda	sausage
ice_cream	rolls_buns	cream_cheese_	liquor_appetizer	cat_food

(a)
Resposta "Agrupamento 1": *mais ou menos*. É possível identificar algumas tendências como pessoas que dão preferência a frutas e legumes (G11). Pode-se notar que os vegetais aparecem em praticamente todos os descritores.
Resposta "Agrupamento 2": *sim*. É possível identificar grupos referentes a bebidas (G10, G14) e doces (G8). Há vários descritores com o mesmo item. Por exemplo, "rolls_buns" aparece em G2, G3, G5, G12.
(b)
Conclusão: *Agrupamento 2*.

Tabela 55: Descritores dos agrupamentos selecionados para o conjunto de dados Sup.

Agrupamento 1				
G1	**G2**	**G3**	**G4**	**G5**
GUARANA_ANTARTICA	SABLUX	AGUA_SANITCANDURA	ACUCAR_UNIAO	OLEO_SOJA_SOYA
LEITE_PARMALAT	ABSSLIVRE	LEITE_LACTUS	DETERGLIMPOL	ACUCAR_DA_BARRA
BISCNABISCO	GUARDGOURMET	CDCLOSEUP	LUSTRA_MOVPOLIFLOR	FEIJAO_TORRESAN
CHOCLACTA	DETERGYPE	MOLHO_TOMATE_CIRIO	SAL_CISNE	FUBA_YOKI
CDCOLGATE	FANTA	PAPEL_HIGPERSONAL	CREME_DE_LEITE_NESTLE	ACUCAR_CRISTAL_DA_BARRA
G6	**G7**	**G8**	**G9**	**G10**
FEIJAO_BROTO_LEGAL	MARGDORIANA	ATUM_COQUEIRO	PAPEL_HIGNEVE	ESPONJA_BOMBRIL
VINAGRE_CASTELO	DESINFPINHO	SUCO_MAGUARY	AGUA_SANITVAREK	CDSORRISO
POP_CORN_YOKI	FILTRO_PAPEL_MELITTA	SCLIXO_PLASLIXO	BISCTOSTINES	PAPEL_HIGPERSONAL
LEITE_MOCA	TORRADA_BAUDUCCO	VEJA_MUSO	AMACCOMFORT	ACHOCNESCAU
CREME_DE_LEITE_NESTLE	LEITE_MOCA	ALCOOL_MULTIALCOOL	LEITE_MOCA	LEITE_MOCA
G11	**G12**	**G13**	**G14**	**G15**
MACRENATA	FERMROYAL	GELATINA_ROYAL	SAL_CISNE	ARROZ_PRATO_FINO
CALDO_MAGGI	MAIONESE_HELLMANN_S	FARTRIGO_RENATA	BOMBRIL	COCA_COLA
BISCNESTLE	GELATINA_ROYAL	BISCNESTLE	ACHOCNESCAU	MAIONESE_HELLMANN_S
MAIONESE_HELLMANN_S	FARTRIGO_RENATA	SPO_OMO	COCA_COLA	SPO_OMO
LEITE_MOCA	LEITE_MOCA	CREME_DE_LEITE_NESTLE	BISCNESTLE	BISCNESTLE
Agrupamento 2				
G1	**G2**	**G3**	**G4**	**G5**
PAPEL_HIGPERSONAL	CAFE_SERRA_DA_GRAMA	LEITE_MOCA	ADOCANTE_FINN	BISCNESTLE
DETERGLIMPOL	OLEO_SOJA_LIZA	CREME_DE_LEITE_NESTLE	MARGQUALY	PAPEL_HIGNEVE
COCA_COLA	LEITE_PARMALAT	BISCNESTLE	SAPOLIO	MARGQUALY
SPO_OMO	QUEIJO_RALADO_FAIXA_AZUL	FARTRIGO_RENATA	FEIJAO_GRAO_DO_CAMPO	DANONINHO
ACUCAR_UNIAO	AGUA_SANITVAREK	GELATINA_ROYAL	FARTRIGO_DONA_BENTA	BATATA_RUFFLES
G6	**G7**	**G8**	**G9**	**G10**
SOBRECOXA_FRANGO_PERDIGAO	CHOCLACTA	BOLO_CAFE_DA_MANHA	SPO_TIXAN	PAO_PULLMAN_BABY_FORMA
BOMBOM_LACTA	AMACBABY_SOFT	PAO_WICKBOLD_FORMA	CAFE_RIBEIRAO_BONITO	CDPREVENT
MANTEIGA_BATAVO	AMENDOIM_C_R	NESCAU_PRONTINHO	OVOS_BRANCOS_KAMIMURA	CERVBRAHMA
IOGVIGOR	ALGODAO_JOHNSON	SUCO_SINUELO	SPO_CAMPEIRO	COCA_COLA
SALSSADIA	COALHADA_VIGOR	AVEIA_QUAKER	CAFE_CABOCLO	CATCHUP_PIC
G11	**G12**	**G13**	**G14**	**G15**
GUARANA_ANTARTICA	GUARANA_KUAT	SALSA_SATO	POMAROLA_KNORRCICA	LEITE_DE_COCO_SOCOCO
SUCO_DEL_VALLE	COCO_RALADO_MAIS_COCO_SOCOCO	CONFETI	POMODORO_KNORRCICA	LEITE_SAO_CARLOS
CDCOLGATE	QUEIJO_RALADO_TEIXEIRA_4_FORMAGGI	MACMEZZANI	MOLHO_AJINOSHOYU	IOGBATAVO
IOGMOLICO	BISCMABEL	ALFACE_CRESPA	CHEIRO_VERDE	ADES_PESSEGO
OVOS_BRANCOS_KAMIMURA	FANTA	GATORADE	BEBLACTEA_CHAMY	PAO_WICKBOLD_LIGHT

(a)
Resposta "Agrupamento 1": *mais ou menos*. Não há muita sobreposição entre os descritores dos grupos. Não foi possível identificar uma linha de produtos como produtos de limpeza.
Resposta "Agrupamento 2": *sim*. É possível identificar algumas linhas de produtos como produtos alimentícios (G3, G14, G15). Há menos sobreposição entre os descritores dos grupos em relação ao "Agrupamento 1".
(b)
Conclusão: *Agrupamento 2*.

As respostas fornecidas pelo especialista encontram-se ao final de cada uma das tabelas. Nota-se que:

- para os CD-P o agrupamento que melhor distribui os tópicos do domínio é o "Agrupamento 1", i.e., os obtidos na etapa de pós-processamento, confirmando a conclusão anteriormente obtida.

- para os CD-NP o agrupamento que melhor distribui os tópicos do domínio é o "Agrupamento 2", i.e., os obtidos na etapa de pré-processamento, confirmando a conclusão anteriormente obtida.

- observando apenas os agrupamentos mais adequados a cada conjunto tem-se que:

  - no conjunto Adult, o especialista conseguiu identificar, por meio dos descritores do "Agrupamento 1", alguns segmentos do domínio. Contudo, o mesmo observou a repetição de itens. Embora haja 15 itens distintos entre os descritores, em que apenas 4 se repetem (RF=0.733; "capital-loss=None": 12x, "marital-status=Married-civ-spouse": 5x, "education=HS-grad": 2x, "income=small": 2x), a repetição de alguns itens ocorre várias vezes, os quais apresentam um alto valor de suporte ("capital-loss=None": 95.33%, "marital-status=Married-civ-spouse": 45.82%).

  - no conjunto Income, o especialista conseguiu identificar, por meio dos descritores do "Agrupamento 1", alguns segmentos do domínio. Contudo, o mesmo observou a repetição de itens. Embora haja 15 itens distintos entre os descritores, em que apenas 4 se repetem (RF=0.733; "language_in_home=english": 10x, "income=0-40.000": 5x, "ethnic_classification=white": 2x, "years_in_bay_area=10+": 2x), a repetição de alguns ocorre várias vezes, os quais apresentam um alto valor de suporte ("language_in_home=english": 91.29%, "income=0-40.000": 62.25%).

  - no conjunto Groceries, o especialista conseguiu identificar, por meio dos descritores do "Agrupamento 2", alguns segmentos do domínio. Contudo, o mesmo observou a repetição de itens. Embora haja 48 itens distintos entre os descritores, 18 deles se repetem (RF=0.625; "rolls_buns": 4x, "shopping_bags": 4x, "soda": 4x, "bottled_water": 3x, "canned_beer": 3x, "whole_milk": 3x, "bottled_beer": 2x, "fruit_vegetable_juice": 2x, "red_blush_wine": 2x, "specialty_bar": 2x, "newspapers": 2x, "pastry": 2x, "chicken": 2x, "root_vegetables": 2x, "other_vegetables": 2x, "sausage": 2x, "specialty_chocolate": 2x, "yogurt": 2x).

– no conjunto Sup, o especialista conseguiu identificar, por meio dos descritores do "Agrupamento 2", alguns segmentos do domínio. Nesse caso, a repetição de itens foi sutil. Dos 71 itens distintos, apenas 4 se repetem (RF=0.940; "BISC-NESTLE": 2x, "COCA_COLA": 2x, "MARGQUALY": 2x, "OVOS_BRANCOS_KAMIMURA": 2x).

Diante do exposto, pode-se notar, segundo o aspecto "Descritivo/Representativo", que: (i) a utilização de agrupamento na etapa de pós-processamento é mais adequada para os CD-P e na etapa de pré-processamento para os CD-NP, como concluído na análise objetiva; (ii) embora todos os agrupamentos tenham apresentado problema em relação a repetição de itens, é possível identificar tendências e segmentos do domínio por meio dos descritores sendo, portanto, úteis ao usuário; (iii) existe uma tendência dos CD-NP serem mais adequados a aplicação de agrupamento, uma vez que em ambos os conjuntos analisados, nos agrupamentos do pré-processamento, o especialista avaliou os descritores dos grupos como adequados (conclusão: *sim*) e, portanto, que uma boa distribuição do conhecimento foi realizada. Contudo, observa-se um bom desempenho dos descritores nos conjuntos considerados em suas respectivas etapas, i.e., pós-processamento nos CD-P e pré-processamento nos CD-NP.

Visando avaliar os resultados segundo o aspecto "Redução via Interesse" aplicou-se a cada agrupamento da Tabela 51 a metodologia PAR-COM no pós e a PAR-COM adaptada no pré. Os parâmetros utilizados foram: m=1 (seleção apenas do grupo mais interessante); h=0.5% do total de regras do agrupamento; média dos ranqueamentos das medidas 18 medidas objetivas já utilizadas. A Tabela 56 apresenta os resultados obtidos: (i) na coluna "Grupo Selecionado" encontra-se o número do grupo selecionado, i.e., o identificado como o mais interessante do agrupamento – os mesmos encontram-se destacados em cinza nas Tabelas 52 (Adult), 53 (Income), 54 (Groceries) e 55 (Sup); (ii) na coluna "# Regras no Grupo" encontra-se o número de regras do agrupamento contidas no grupo selecionado (item (i)); (iii) na coluna "h [0.5%]" encontra-se a quantidade de regras referente a 0.5% do total das existentes no agrupamento; (iv) na coluna "h no Grupo" encontra-se a porcentagem das h regras do agrupamento (item (iii)) contidas no grupo selecionado (item (i)), além do valor absoluto entre parênteses; (v) na coluna "h' no Grupo" encontra-se a porcentagem das h' regras do agrupamento contidas no grupo selecionado (item (i)); (vi) na coluna "R no Grupo" encontra-se a redução do espaço de exploração obtida por meio do grupo selecionado (item (i)). O objetivo era que o especialista avaliasse as regras contidas nos grupos selecionados a fim de responder algumas questões. Contudo, o número de regras a serem avaliadas era muito alto, nesse caso, 40232 (soma da coluna "# Regras no Grupo"). Desse modo, optou-se pelo seguinte procedimento: considerar um subconjunto

das regras do grupo de tal modo que o mesmo contivesse as h-top regras do grupo mais n regras escolhidas aleatoriamente, em que n contivesse a mesma proporção de h-top regras do grupo. Por exemplo, o conjunto Income, na organização Ward:J-RT, contém 13 h-top regras no grupo selecionado. Portanto, o subconjunto a ser considerado contém essas 13 regras mais 13 regras escolhidas aleatoriamente, totalizando 26 regras a serem avaliadas. A quantidade de regras a serem avaliadas pelo especialista em cada organização encontra-se na Tabela 56, coluna "# Regras a Avaliar", totalizando 1276 regras (soma da coluna). Esse procedimento foi adotado levando em consideração a hipótese, adotada na PAR-COM/PAR-COM adaptada, de que uma regra interessante dentro de um grupo auxilia na descoberta de outros conhecimentos potencialmente interessantes. Desse modo, qualquer regra selecionada aleatoriamente representa uma regra potencialmente interessante.

Tabela 56: Resultados da PAR-COM/PAR-COM adaptada nos agrupamentos selecionados e apresentados na Tabela 51.

Etapa	Tipo de Dado	Conjunto de Dado	Organização	Método de Rotulação	Grupo Selecionado	# Regras no Grupo	h [0.5%]	h no Grupo	h' no Grupo	R no Grupo	# Regras a Avaliar
Pós	CD-P	Adult	Ward:J-RT [k=15]	LM-T	9	544	33	100% (33)	91%	92%	66
		Income			9	175	19	68% (13)	47%	95%	26
	CD-NP	Groceries		LM-PU	15	66	11	55% (6)	18%	97%	12
		Sup			13	500	38	39% (15)	32%	93%	30
Pré	CD-P	Adult	PAM:Agg [k=15]	LM-T	9	10233	401	30% (119)	24%	87%	238
		Income			10	10309	661	29% (189)	73%	92%	378
	CD-NP	Groceries		LM-PU	1	51	4	100% (4)	100%	93%	8
		Sup			1	18354	259	100% (259)	98%	65%	518

Considerando os subconjuntos de regras construídos a partir dos grupos selecionados, apresentou-se ao especialista, para cada conjunto de dados, dois subconjuntos, um do pós e um do pré. Os subconjuntos não são exibidos aqui devido ao elevado número de regras (1276). Os subconjuntos referentes aos agrupamentos do pós foram apresentados como "Subconjunto 1" e os do pré como "Subconjunto 2". O especialista não foi informado sobre qual dos subconjuntos pertencia a cada etapa. É importante mencionar que juntamente com as regras disponibilizou-se também o descritor do grupo a partir do qual as regras foram selecionadas. Para cada conjunto de dados, solicitou-se ao especialista que o mesmo indicasse:

1. as regras, em cada subconjunto, que *não* apresentassem conhecimento *óbvio* (comum; não interessante). Optou-se por utilizar a classificação "óbvio" e "não óbvio" em vez de "interessante" e "não interessante" pelo fato de não se ter nenhum objetivo pré-definido para avaliar as regras segundo o aspecto de interesse. Além disso, o conceito de interesse é mais subjetivo do que o de "óbvio". Desse modo, considera-se que um conhecimento "não óbvio" contém alguma informação relevante.

2. se os descritores, em cada subconjunto, estavam ou não adequados, i.e., se os descritores realmente representavam o conhecimento contido nas regras. Ele poderia responder *sim*, *não* ou *mais ou menos*.

3. dentro de cada conjunto de dados qual dos dois subconjuntos era mais interessante em termos de conhecimento gerado. Ele poderia responder "indiferente"caso não conseguisse concluir.

As Tabelas 57 (Adult), 58 (Income), 59 (Groceries) e 60 (Sup) apresentam os resultados da avaliação: (i) na coluna "Não Óbvias" encontra-se a quantidade de regras classificadas como não óbvias; (ii) na coluna "Não Óbvias h" encontra-se quantas das regras classificadas como não óbvias (item (i)) são h-top regras; (iii) na coluna "Não Óbvias $\overline{h}$" encontra-se quantas das regras classificadas como não óbvias (item (i)) são regras selecionadas aleatoriamente; (iv) na coluna "Descritores" encontra-se a avaliação sobre a adequação dos descritores ao subconjunto; (v) na coluna "Subconjunto Selecionado" encontra-se indicado, por um "X", o subconjunto que o especialista achou mais interessante. Por exemplo, na Tabela 57, referente ao conjunto Adult, das 238 regras apresentadas ao especialista no "Subconjunto 2" (vide Tabela 56), 9 foram classificadas como não óbvias, sendo 2 das 9 h-top regras e 7 das 9 selecionadas aleatoriamente. Nota-se que:

- No conjunto Adult, a diferença entre a proporção de regras classificadas como não óbvias, nos dois subconjuntos, foi pequena (0.78%). Contudo, o pré-processamento

apresentou um desempenho um pouco melhor. Ambos os subconjuntos apresentaram bons descritores tendo, contudo, o do pós obtido uma melhor avaliação. Por fim, o especialista indicou como subconjunto mais interessante o do pré.

- No conjunto Income, o subconjunto referente ao pré-processamento apresentou um melhor desempenho em relação a proporção de regras classificadas como não óbvias. Nenhum dos subconjuntos apresentou bons descritores. Por fim, o especialista indicou como subconjunto mais interessante o do pré.

- No conjunto Groceries, o subconjunto referente ao pré-processamento apresentou um melhor desempenho em relação a proporção de regras classificadas como não óbvias. Ambos os subconjuntos apresentaram bons descritores tendo, contudo, o do pós obtido uma melhor avaliação. Por fim, o especialista indicou como subconjunto mais interessante o do pré.

- No conjunto Sup, a diferença entre a proporção de regras classificadas como não óbvias, nos dois subconjuntos, foi pequena (0.23%). Contudo, o pré-processamento apresentou um desempenho um pouco melhor. Ambos os subconjuntos apresentaram bons descritores tendo, contudo, o do pós obtido uma melhor avaliação. Por fim, o especialista indicou como subconjunto mais interessante o do pré.

- nos CD-P os subconjuntos do pré sempre contêm h-top regras classificadas como não óbvias, diferente do pós, que só contêm as selecionadas aleatoriamente. Isso indica que as h-top regras do pré são provavelmente mais interessantes do que as do pós. Contudo, é possível verificar em todos os casos, CD-P e CD-NP, que uma regra interessante dentro de um grupo auxilia na descoberta de outros conhecimentos potencialmente interessantes – hipótese considerada na metodologia PAR-COM/PAR-COM adaptada.

Tabela 57: Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Adult.

Subconjunto 1				
Não Óbvias	Não Óbvias h	Nao Óbvias $\overline{h}$	Descritores	Subconjunto Selecionado
2 [3%]	0 [0%]	2 [6%]	*sim*	
Subconjunto 2				
9 [3.78%]	2 [1.68%]	7 [5.88%]	*mais ou menos*	X

Tabela 58: Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Income.

Subconjunto 1				
Não Óbvias	Não Óbvias h	Nao Óbvias $\overline{h}$	Descritores	Subconjunto Selecionado
1 [3.85%]	0 [0%]	1 [7.69%]	*não*	
Subconjunto 2				
26 [6.88%]	17 [8.99%]	9 [4.76%]	*não*	X

Tabela 59: Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Groceries.

Subconjunto 1				
Não Óbvias	Não Óbvias h	Nao Óbvias $\overline{h}$	Descritores	Subconjunto Selecionado
0 [0%]	0 [0%]	0 [0%]	*sim*	
Subconjunto 2				
3 [37.5%]	2 [50%]	1 [25%]	*mais ou menos*	X

Tabela 60: Resultados da avaliação especialista, segundo o aspecto "Redução via Interesse", para o conjunto de dados Sup.

Subconjunto 1				
Não Óbvias	Não Óbvias h	Nao Óbvias $\overline{h}$	Descritores	Subconjunto Selecionado
3 [10%]	1 [6.67%]	2 [13.33%]	*sim*	
Subconjunto 2				
53 [10.23%]	27 [10.42%]	26 [10.04%]	*mais ou menos*	X

Diante do exposto, pode-se notar, segundo o aspecto "Redução via Interesse", que: (i) a utilização de agrupamento na etapa de pré-processamento é mais adequada para os CD-NP, como concluído na análise objetiva; (ii) em relação aos CD-P, existe uma tendência do pré-processamento ser também a etapa mais adequada ao agrupamento, não tendo, contudo, a avaliação apresentado um padrão bem definido; (iii) os descritores dos CD-NP, de todos os subconjuntos considerados, obtiveram uma boa avaliação, i.e., representaram adequadamente os conceitos contidos nas regras; (iv) existe uma tendência dos CD-NP serem mais adequados a aplicação de agrupamento, uma vez que em ambos os conjuntos analisados, nos subconjuntos do pré-processamento, os mesmos obtiveram um bom desempenho em relação aos conhecimentos não óbvios e aos descritores.

Por fim, considerando toda a análise subjetiva realizada, **indica-se (i) a utilização de agrupamento na etapa de pré-processamento para os CD-NP, tanto no que se refere ao aspecto "Redução via Interesse" quanto ao "Descritivo/Representativo" e (ii) a utilização de agrupamento na etapa de pós-processamento para os CD-P no que se refere ao aspecto "Descritivo/Representativo"**.

5.3 Considerações Finais

A Tabela 61 apresenta as conclusões obtidas nas Seções 5.1 e 5.2 referente as duas comparações: objetiva e subjetiva. Observa-se que: **(i) a utilização de agrupamento na etapa de pré-processamento é indicada para os CD-NP em ambas as análises e aspectos considerados; (ii) a utilização de agrupamento na etapa de pós-processamento é indicada para os CD-P em quase todos os aspectos das análises realizadas; (iii) considerando (i) e (ii), existe uma tendência dos CD-NP serem mais adequados a aplicação de agrupamento do que os CD-P; (iv) o agrupamento, em ambas as etapas, apóia o usuário**.

Tabela 61: Síntese dos resultados das comparações objetiva e subjetiva realizadas entre as abordagens de pós-processamento e pré-processamento.

Tipo de Análise	"Redução via Interesse"	"Descritivo/Representativo"
Objetiva	pós: CD-P pré: CD-NP	pós: CD-P pré: CD-NP
Subjetiva	- pré: CD-NP	pós: CD-P pré: CD-NP

6 Conclusão

Devido a grande quantidade de padrões que se obtém em um processo de mineração de regras de associação, diversas linhas de pesquisa vem sendo adotadas para solucionar o problema, propondo, para tanto, abordagens distintas. Entre as abordagens existentes encontra-se a de agrupamento. O agrupamento pode ser aplicado de maneira distinta em duas das etapas do processo de mineração de dados, pré-processamento e pós-processamento, a fim de facilitar a busca pelos padrões relevantes a aplicação. Uma vez que ambas possibilidades existem, surge a pergunta: o que vale mais a pena?

A fim de responder essa questão, este relatório técnico dividiu a investigação em três partes: (i) a contribuição do agrupamento no pós-processamento (Seção 3), (ii) a contribuição do agrupamento no pré-processamento (Seção 4) e (iii) a comparação entre o pós-processamento e o pré-processamento considerando os resultados obtidos em (i) e (ii) (Seção 5). Em cada parte do estudo, metodologias foram propostas e diversas contribuições obtidas, a saber:

- proposta da metodologia PAR-COM;
- proposta da metodologia PAR-COM adaptada;
- proposta das medidas P e RF para avaliar métodos de rotulação;
- proposta dos índices para avaliar as particularidades do agrupamento no pré-processamento;
- identificação da organização mais adequada para se reduzir o espaço de exploração, identificar o conhecimento potencialmente interessante e obter bons descritores, tanto no pós-processamento como no pré-processamento;
- identificação dos métodos de rotulação mais adequados a cada uma das etapas investigadas.

Como resultado de todo o estudo, obteve-se que:

- a organização Ward:J-RT é a mais indicada para agrupamentos de regras de associação na etapa de pós-processamento;
- a organização PAM:Agg é a mais indicada para agrupamentos de regras de associação na etapa de pré-processamento;

- o método de rotulação LM-T é mais indicado para CD-P e o LM-PU para CD-NP, independente da etapa em que o agrupamento é aplicado;
- a utilização de agrupamento na etapa de pré-processamento é indicada para os CD-NP;
- a utilização de agrupamento na etapa de pós-processamento é indicada para os CD-P;
- existe uma tendência dos CD-NP serem mais adequados a aplicação de agrupamento do que os CD-P;
- o agrupamento, em ambas as etapas, apóia o usuário.

Diante do exposto, **respondendo a pergunta inicialmente realizada (o que vale mais a pena?), tem-se que utilizar agrupamento é interessante em ambas as etapas considerando que o mesmo deve ser utilizado no pós-processamento para os CD-P e no pré-processamento para os CD-NP**.

Por fim, é importante mencionar que todos os processos e metodologias apresentadas foram implementadas no RulEE-AREC [Carvalho et al., 2012,Padua and Carvalho, 2011][8]. O RulEE-AREC (*A*ssociation *R*ules *E*xploration through *C*lustering) compõe um dos módulos de um ambiente de exploração de regras denominado RulEE (*Rul*e *E*xploration *E*nvironment) [Paula, 2003], o qual visa disponibilizar o conhecimento obtido no processo de mineração de dados para avaliação pelos usuários finais. O módulo incorpora ao RulEE uma nova abordagem de exploração, por meio dos processos e das metodologias propostas, facilitando a exploração das regras e a execução dos experimentos por meio de uma interface amigável.

[8]Trabalho de Iniciação Científica desenvolvido com apoio FAPESP 2011/00951-0

Referências

[Aggarwal et al., 2002] Aggarwal, C. C., Procopiuc, C., and Yu, P. S. (2002). Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):51–62.

[Aggelis, 2004] Aggelis, V. (2004). Association rules model of e-banking services. *Data Mining V – Information and Communication Technologies*, 5:46–55.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *VLDB'94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499.

[Baesens et al., 2000] Baesens, B., Viaene, S., and Vanthienen, J. (2000). Post-processing of association rules. In *KDD'00: Proceedings of the Special Workshop on Post-processing, The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2–8.

[Carvalho et al., 2011] Carvalho, V. O., Santos, F. F., and Rezende, S. O. (2011). Post-processing association rules with clustering and objective measures. In *ICEIS'11: Proceedings of the 13th International Conference on Enterprise Information Systems*, volume 1, pages 54–63.

[Carvalho et al., 2012] Carvalho, V. O., Santos, F. F., Rezende, S. O., and Padua, R. (2012). PAR-COM: A new methodology for post-processing association rules. *Lecture Notes in Business Information Processing*, 102:66–80.

[Changguo et al., 2009] Changguo, Y., Nianzhong, W., Tailei, W., Qin, Z., and Xiaorong, Z. (2009). The research on the application of association rules mining algorithm in network intrusion detection. In Hu, Z. and Liu, Q., editors, *ETCS'09: Proceedings of the 1st International Workshop on Education Technology and Computer Science*, volume 2, pages 849–852.

[Cutting et al., 1992] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR'92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, New York, NY, USA. ACM.

[D'Enza et al., 2005] D'Enza, A. I., Palumbo, F., and Greenacre, M. (2005). Exploratory data analysis leading towards the most interesting binary association rules. In *ASMDA'05: Proceedings of the 11th Symposium on Applied Stochastic Models and Data Analysis*, pages 256–265.

[Frank and Asuncion, 2010] Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences. On-line. `http://archive.ics.uci.edu/ml`. [16/07/2012].

[Fung et al., 2003] Fung, B. C., Wang, K., and Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proceedings of the 3rd SIAM International Conference on Data Mining*, pages 59–70.

[Glover et al., 2002] Glover, E. J., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002). Inferring hierarchical descriptions. In *CIKM'02: Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 507–514. ACM.

[Hahsler et al., 2012] Hahsler, M., Grün, B., Hornik, K., and Buchta, C. (2012). Introduction to arules – a computational environment for mining association rules and frequent item sets. On-line. `http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf`. [25/04/2012].

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Series in Statistics.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall.

[Jorge, 2004] Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In Berry, M. W., Dayal, U., Kamath, C., and Skillicorn, D., editors, *SIAM'04: Proceedings of the 4th SIAM International Conference on Data Mining*. 10p.

[Kashyap et al., 2005] Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266.

[Koh and Pears, 2008] Koh, Y. S. and Pears, R. (2008). Rare association rule mining via transaction clustering. In Roddick, J. F., Li, J., Christen, P., and Kennedy, P. J., editors, *AusDM'08: 7th Australasian Data Mining Conference*, volume 87 of *CRPIT*, pages 87–94, Glenelg, South Australia. ACS.

[Larsen and Aone, 1999] Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *KDD'99: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22, New York, NY, USA. ACM.

[Lopes et al., 2007] Lopes, A. A., Pinho, R., Paulovich, F. V., and Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326.

[Manning et al., 2009] Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press. 544p.

[Moura and Rezende, 2010] Moura, M. F. and Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications*, pages 336–371.

[Natarajan and Shekar, 2005] Natarajan, R. and Shekar, B. (2005). Interestingness of association rules in data mining: issues relevant to e-commerce. *SĀDHANĀ – Academy Proceedings in Engineering Sciences (The Indian Academy of Sciences)*, 30(Parts 2&3):291–310.

[P. Hansen, 1997] P. Hansen, B. J. (1997). *Cluster analysis and mathematical programming*, pages 191–215.

[Padua and Carvalho, 2011] Padua, R. and Carvalho, V. O. (2011). Rulee-arec: Módulo de exploração de regras de associação via agrupamento. In *Anais do XXIII Congresso de Iniciação Científica da Unesp (CIC)*. 5p. Disponível em: `http://prope.unesp.br/xxiii_cic/fase1.php` [26/07/2012].

[Paula, 2003] Paula, M. (2003). Ambiente para exploração de regras. Master's thesis, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos – Brasil.

[Plasse et al., 2007] Plasse, M., Niang, N., Saporta, G., Villeminot, A., and Leblond, L. (2007). 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*, 52(1):596–613.

[Popescul and Ungar, 2000] Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters, unpublished manuscript. On-line. `http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf`. [16/07/2012].

[Rajasekar and Weng, 2009] Rajasekar, U. and Weng, Q. (2009). Application of association rule mining for exploring the relationship between urban land surface temperature

and biophysical/social parameters. *Photogrammetric Engineering & Remote Sensing*, 75(3):385–396.

[Reynolds et al., 2006] Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504.

[Sahar, 2002] Sahar, S. (2002). Exploring interestingness through clustering: a framework. In *ICDM'02: Proceedings of the IEEE International Conference on Data Mining*, pages 677–680.

[Semenova et al., 2001] Semenova, T., Hegland, M., Graco, W., and Williams, G. (2001). Effectiveness of mining association rules for identifying trends in large health databases. In Kurfess, F. J. and Hilario, M., editors, *ICDM'01: Workshop on Integrating Data Mining and Knowledge Management, The IEEE International Conference on Data Mining*. 12p.

[Tan et al., 2004] Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313.

[Tan et al., 2006] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley Companion Book Site.

[Toivonen et al., 1995] Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. Workshop Notes of the ECML'95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, 47–52, MLnet.

[Treeratpituk and Callan, 2006] Treeratpituk, P. and Callan, J. (2006). Automatically labeling hierarchical clusters. In *Proceedings of the 6th International Conference on Digital Government Research*, pages 167–176, New York, NY, USA. ACM.

[Wang and Karypis, 2004] Wang, J. and Karypis, G. (2004). Summary: efficiently summarizing transactions for clustering. In *Proceedings of the 4th IEEE International Conference on Data Mining*, pages 241–248. IEEE Computer Society.

[Wang et al., 1999] Wang, K., Xu, C., and Liu, B. (1999). Clustering transactions using large items. In *CIKM'99: Proceedings of the 8th International Conference on Information and Knowledge Management*, pages 483–490.

[Xu and Wunsch, 2008] Xu, R. and Wunsch, D. C. (2008). *Clustering*. A John Wiley & Sons.

[Yang, 2005] Yang, L. (2005). Pruning and visualizing generalized association rules in parallel coordinates. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):60–70.

[Yang and Padmanabhan, 2005] Yang, Y. and Padmanabhan, B. (2005). Ghic: A hierarchical pattern-based clustering algorithm for grouping web transactions. *IEEE Transactions on Knowledge and Data Engineering*, 17:1300–1304.

[Yun et al., 2001] Yun, C.-H., Chuang, K.-T., and Chen, M.-S. (2001). An efficient clustering algorithm for market basket data based on small large ratios. In *Proceedings of the 25th International Computer Software and Applications Conference on Invigorating Software Development*, pages 505–510. IEEE Computer Society.

[Zhao et al., 2009] Zhao, Y., Zhang, C., and Cao, L. (2009). *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. IGI Global.

[Zheng et al., 2001] Zheng, Z., Kohavi, R., and Mason, L. (2001). Real world performance of association rule algorithms. In Provost, F. and Srikant, R., editors, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 401–406. ACM Press.