RENATA ASSIS DE MATOS

COMPARAÇÃO DE METODOLOGIAS DE ANÁLISE DE AGRUPAMENTOS NA PRESENÇA DE VARIÁVEIS CATEGÓRICAS E CONTÍNUAS

Belo Horizonte – MG

Abril / 2007

RENATA ASSIS DE MATOS

COMPARAÇÃO DE METODOLOGIAS DE ANÁLISE DE AGRUPAMENTOS NA PRESENÇA DE VARIÁVEIS CATEGÓRICAS E CONTÍNUAS

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Curso de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Mestre em Estatística

Orientadora: Profa. Dra. Sueli Aparecida Mingoti

Universidade Federal Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Belo Horizonte – MG

Abril / 2007

Agradecimentos

A Deus que me guiou durante a caminhada.

À professora doutora Sueli Aparecida Mingoti pela orientação e dedicação prestadas no desenvolvimento desta dissertação. Por todo o conhecimento que contribuiu para o meu aprimoramento intelectual, profissional e pessoal.

Aos meus pais, Marilene e Ricardo, e irmãos, André e Artur, por tudo que representam em minha vida. A Vannucci pelo apoio e incentivo constantes. Todos de forma especial e carinhosa me deram força e perseverança.

Aos bons amigos que encontrei durante o curso pela presença, companheirismo e paciência. Pessoas especiais que não preciso citar seus nomes, pois sabem quem são. Espero que a distância não nos separe tanto.

A CAPES e ao Departamento de Estatística da Universidade Federal de Minas Gerais pelo apoio financeiro concedido.

"Dificuldades no início é prenúncio de grande sucesso,
desde que haja perseverança"

Resumo

A análise de agrupamentos é um procedimento de estatística multivariada que engloba técnicas que objetivam organizar objetos em grupos de acordo com a proximidade existente entre eles. Os objetos de um mesmo grupo são tão similares quanto possível (coesão interna) e ao mesmo tempo tão dissimilares quanto possível dos objetos dos demais grupos (isolamento externo). Os métodos são compostos de dois elementos: a medida de proximidade e o algoritmo de agrupamento. Apesar da sua vasta aplicabilidade, desde o início do seu desenvolvimento o foco principal tem sido nas situações em que somente variáveis contínuas caracterizam os objetos. Atualmente é grande a necessidade de se considerar também a informação de variáveis categóricas. No entanto, os estudos encontrados na literatura envolvendo esse tipo de variável não comparam os métodos de forma adequada, fazendo com que a diversidade de possibilidades dificulte a escolha da melhor técnica.

Nesta dissertação é feito um estudo comparativo de cinco algoritmos de análise de agrupamentos somente na presença de variáveis categóricas e de três metodologias que são aplicáveis para casos de variáveis categóricas e contínuas. Dentre esses, a extensão do método ROCK para o caso de mistura de variáveis é uma proposta desta dissertação. Avaliam-se também outras questões tais como o efeito do grau de separação e sobreposição, do número de grupos, de variáveis e de categorias, a correlação entre as variáveis contínuas e a atribuição de pesos da medida de proximidade combinada, usada quando há os dois tipos de variáveis caracterizando os objetos em análise. A comparação é feita a partir de um esquema de simulação e de estudos de aplicação prática e a medida de desempenho utilizada é a taxa de alocação correta.

Pelos resultados obtidos, conclui-se que o aumento do número de grupos, independente da estrutura desses, prejudica o desempenho dos algoritmos. A influência do número de variáveis e de categorias depende da disposição dos grupos. Observou-se também que a correlação existente entre as variáveis contínuas não influenciou as taxas de alocação correta dos métodos e que esses têm melhores resultados quando é dado maior peso às variáveis contínuas na medida de proximidade combinada. Quanto à eficiência, o ROCK foi o algoritmo que se destacou nos estudos de simulação realizados.

Palavras-chave: Análise de agrupamentos, variáveis categóricas, variáveis contínuas, Ligação Média, *ROCK*, k-Modas, k-Protótipos, *Fuzzy* c-Modas, k-Populações

Abstract

Cluster Analysis is the name given to a group of several types of algorithms used to organize objects into groups taking into account the proximity that exists between them. Objects in the same group are as similar as possible to each other (internal cohesion) and are as dissimilar as possible to the objects in the different groups (external isolation). Cluster procedures are based upon two components: the proximity measure and the algorithm. Despite of the your wide applicability of these methods, the majority of the studies published in the literature focus on continuous variables. More recently attention has been given to new algorithms that can incorporate the information of categorical variables. However, the recent papers do not compare these new methods in a proper way and the existence of different possibilities difficult the choice of the better method.

In this dissertation a comparative study is performed. Five algorithms which are applicable for categorical variables and three which are applicable for both types of variables are examined. Among these last three algorithms, the extension of ROCK, which allows to cluster objects by using both types of variables is a new proposal of this dissertation. Besides that, it is also evaluated the influence of cluster overlapping, the number of groups, variables and categories, the correlation between the continuous variable and the choice of the weights of the combined proximity measure, that is used when the objects are clustered using the two types of variables.

Based on the results of this dissertation it can be concluded that when the number of groups increase, independent of their structure, the performance of the clustering algorithms decreased. The effect of the increase of the number of variables and categories depends on the internal structure of the clusters. It was also noticed that the correlation between the continuous variables does not cause any effect on the percentage of correct classification and that the clustering methods have better results when in the combined proximity measure more weight is given to the continuous variables. In terms of efficiency, the ROCK algorithm had better performance all simulation studies of this dissertation.

Keywords: Cluster analysis, categorical variables, continuous variables, Average Linkage, ROCK, k-Modes, k-Prototypes, Fuzzy c-Modes, k-Populations

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A análise de agrupamentos, também conhecida como análise de *cluster* ou de conglomerados, é uma técnica estatística multivariada que tenta sintetizar ou simplificar a estrutura de variabilidade dos dados. O objetivo dessa técnica é organizar elementos, sejam esses observações ou variáveis, em grupos, com base nas informações quanto à sua proximidade. O resultado obtido a partir da aplicação dessa técnica é um conjunto de grupos com coesão interna e isolamento externo (EVERITT, 1993), ou seja, elementos dentro de um mesmo grupo são tão similares quanto possível e são, ao mesmo tempo, tão dissimilares quanto possível dos elementos presentes nos demais grupos.

Os métodos de análise de agrupamentos vêm sendo desenvolvidos desde, principalmente, Sokal e Sneath (1963), e o fator primordial que corrobora positivamente para que muitos pesquisadores estejam interessados no tema é a vasta aplicação desses procedimentos. São diversas as áreas do conhecimento que os utilizam para compreender e explicar fenômenos. Hartigan (1975) apresenta aplicações na classificação de animais e plantas, de doenças e de campos de trabalho como exemplos de utilização na Biologia, Medicina e Arqueologia/Antropologia, respectivamente. Everitt (1993) traz ainda aplicações nas áreas de Serviço Social, Pesquisa de Mercado e Educação, dentre outras.

Os exemplos de aplicação encontrados na literatura demonstram também essa diversidade. Observa-se a classificação de grãos de soja, de cogumelos, de animais, de votos no congresso americano, de indivíduos que solicitam crédito, de pacientes com hepatite ou com distúrbios da glândula tireóide, dentre outras. Com base em determinadas características, algumas vezes exclusivamente categóricas, outras tanto categóricas quanto contínuas, tem-se por objetivo reunir esses elementos em grupos tais como o tipo de doença (grãos de soja), de partido político (votos no congresso), de concessão de crédito (indivíduos) e de desfecho médico (pacientes), dentre outros.

O avanço computacional e o desenvolvimento tecnológico proporcionam a construção de extensos bancos de dados, contendo todo e qualquer tipo de variável. A aplicação de diversas técnicas estatísticas em uma grande massa de informações objetivando melhorar a análise a partir da redução das dimensões de observações e variáveis (mineração de dados ou *data mining*) vem se tornando, portanto, necessária às grandes empresas e

a aplicabilidade da análise de agrupamentos é cada vez maior. Em geral, as técnicas estatísticas envolvidas são primeiramente aplicadas a uma amostra das informações a serem tratadas e, então, os resultados obtidos são usados para agrupar as observações restantes. Há portanto uma crescente preocupação em, principalmente, fazer com que os métodos tratem de forma adequada as informações existentes e sejam também de menor complexidade.

As metodologias existentes são determinadas, basicamente, pela medida de proximidade e pelo algoritmo empregados. As medidas de proximidade são quantidades comparativas entre as observações, os grupos de observações ou entre ambos e são definidas com base na natureza das variáveis que caracterizam os objetos em análise. De forma geral, os algoritmos descrevem como o procedimento de agrupamento deve ser realizado e alguns deles possuem características especiais que precisam ser analisadas com cautela como, por exemplo, a inicialização do algoritmo e a determinação de parâmetros. A medida de proximidade e o algoritmo são, portanto, os objetos de estudo dos pesquisadores.

Muitos dos métodos desenvolvidos focam, especialmente, em dados caracterizados por variáveis contínuas (MINGOTI; LIMA, 2006). Quando há ocorrência de variáveis categóricas, algumas aproximações são usuais: transformá-las em contínuas, atribuindo valores numéricos às suas categorias, ou em binárias, fazendo com que cada uma das suas categorias se torne uma variável que represente presença ou ausência desse determinado atributo, transformar as contínuas em categóricas criando classes de valores ou ainda aplicar aos dados medidas específicas que tratam as observações conjuntamente (JOBSON, 1991-92; MINGOTI, 2005). Segundo Everitt (1993) há ainda uma última aproximação que seria analisar os dados separadamente de acordo com o tipo de variável e, por fim, sintetizar os resultados dos diferentes estudos. Anderberg (1973) também apresenta algumas dessas soluções. As principais desvantagens dessas tentativas de tratar observações de diferentes variáveis são a perda de informação, o aumento indesejado do número de variáveis, a ineficiência quando grande volume de informações precisa ser agrupado, a impossibilidade de se consolidar resultados discordantes e a dificuldade na determinação dos pesos da medida de proximidade conjunta.

Tentando contornar tais desvantagens e também suprir as atuais necessidades, metodologias foram desenvolvidas. Entre elas destacam-se k-Modas e k-Protótipos (HUANG, 1997, 1998), *STIRR*: *Sieving Through Iterated Relational Reinforcement* (GIBSON *et al.*, 2000), *CACTUS*: *Clustering Categorical Data Using Summaries* (GANTI *et al.*, 1999), *Fuzzy* c-Modas (HUANG; NG, 1999), *ROCK*: *Robust Clustering Using Links* (GUHA *et*

al., 2000), *A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment* (CHIU *et al.*, 2001), *LIMBO* (ANDRITSOS, 2004), *QROCK*: *Quick ROCK* (DUTTA *et al.*, 2005), *M-BILCOM* (ANDREOPOULOS *et al.*, 2005), *BILCOM* (ANDREOPOULOS *et al.*, 2006), dentre outras. Percebe-se que há na literatura muitas publicações, mas também que proporcionalmente há ainda poucas referências que abordam variáveis categóricas e contínuas conjuntamente.

Além disso, a diversidade de possibilidades pode levar à dificuldade sobre qual técnica deve ser empregada e, conseqüentemente, levar a resultados diferentes, devido ao fato de os agrupamentos finais serem fortemente dependentes da metodologia usada. A escolha inadequada da técnica pode comprometer os resultados obtidos. Dessa forma, estudos comparativos contribuiriam na identificação dos métodos mais satisfatórios para uma determinada situação. O desempenho dos algoritmos propostos, combinados com as medidas adequadas, é de fato uma informação relevante que precisa ser cuidadosamente estudada.

Os trabalhos encontrados na literatura são limitados na medida em que as suas conclusões com respeito à eficiência dos métodos analisados são obtidas com base nos mesmos bancos de dados e uma comparação adequada só é viável quando se conhece a estrutura dos dados, o que somente é conseguido por meio de simulação. Esse tipo de estudo é de extrema importância, portanto, pois possibilita que sejam feitas recomendações úteis na prática. Dois artigos até trazem pequenos estudos de simulação. Em Huang (1997), as observações de três variáveis, sendo duas contínuas e uma categórica, foram geradas em apenas duas situações específicas. Em Chiu *et al.* (2001), a explicação é tão sucinta que não é possível compreender o que de fato foi feito para gerar o banco de dados contendo os dois tipos de variáveis.

1.1 Objetivos

Para contribuir para o desenvolvimento de técnicas mais eficientes e potencializar ainda mais a utilização dos métodos de análise de agrupamentos foram objetivos gerais desta dissertação:

- pesquisar os principais métodos de análise de agrupamentos na presença de variáveis categóricas;

- verificar a existência de extensões dessas metodologias para a ocorrência de variáveis

categóricas e contínuas;

- propor extensões do *ROCK* dado que atualmente essa técnica se aplica quando somente variáveis categóricas estão presentes;

- formalizar a extensão da Ligação Média para o agrupamento de observações dos dois tipos de variáveis. Essa abordagem é pouco encontrada na literatura, apesar de em essência o método poder ser usado nesse contexto;

- testar tais métodos em contextos mais gerais, via simulação, possibilitando uma comparação de suas eficiências;

- aplicar essas metodologias a estudos de caso encontrados na literatura;

- estudar o efeito:

    . do número de grupos, de variáveis e de categorias;

    . da correlação entre as variáveis contínuas;

- analisar a atribuição dos pesos da medida de proximidade combinada.

Oito métodos foram comparados por meio dos estudos de simulação e aplicação, sendo cinco para observações de variáveis categóricas e três para observações desse tipo e também de variáveis contínuas. Os métodos Ligação Média, k-Modas, *Fuzzy* c-Modas, k-Populações e *ROCK* compõem o primeiro conjunto de técnicas e os métodos Ligação Média, k-Protótipos e *ROCK* estendido (proposto nesta dissertação) compõem o segundo conjunto. Nesse último conjunto as técnicas *Fuzzy* c-Modas e o k-Populações não puderam ser consideradas porque até o momento não possuíam extensões para tratar observações de ambos os tipos de atributos.

1.2 Contribuições

Fazem parte das contribuições deste trabalho o desenvolvimento e o aprimoramento de técnicas não encontrados na literatura pesquisada. São estas:

- comparação via simulação dos métodos de agrupamentos de observações de variáveis categóricas e também de variáveis contínuas;

- extensão do método *ROCK*, possibilitando o seu uso quando observações de qualquer tipo de variável (categóricas e contínuas) precisam ser agrupadas;

- estudo da atribuição dos pesos da medida de proximidade combinada. Essa medida é usada pelos métodos aplicáveis quando há ocorrência tanto de variáveis categóricas quanto contínuas;

- alteração da inicialização do método k-Populações. Optou-se pela inicialização aleatória das sementes e com isso o algoritmo originalmente proposto precisou ser adaptado;

- modificação na forma de alocação das observações aos grupos para os métodos *Fuzzy* c-Modas e k-Populações. O critério comumente usando na literatura foi aprimorado, mas a sua implementação não altera significativamente a forma original dos algoritmos;

Cabe observar que o agrupamento de variáveis, assunto abordado por diversas fontes bibliográficas pesquisadas, não é o foco dessa dissertação. Sugere-se portanto que as referências apresentadas ao final do trabalho sejam consultadas caso seja esse o interesse do leitor.

1.3 Organização da Dissertação

Esse trabalho está organizado da seguinte forma: Inicialmente (capítulo 2), é feita uma revisão da literatura encontrada, em seguida (capítulo 3), descrevem-se as técnicas a serem comparadas na dissertação, em sua versão original (seção 3.1), as extensões do método *ROCK* (seção 3.2) e algumas observações referentes à implementação computacional dos algoritmos (seção 3.3).

Os estudos de simulação e aplicação são apresentados no capítulo 4. Na primeira seção desse capítulo (seção 4.1), definem-se as situações simuladas e alguns aspectos necessários à simulação, apresentam-se os resultados e também os comentários com respeito ao desempenho das metodologias. Na seção seguinte (seção 4.2), descrevem-se os bancos de dados das aplicações e comentam-se os resultados da análise de agrupamentos realizada.

No capítulo 5 encontram-se os principais resultados dos estudos realizados, as conclusões, as limitações da dissertação e também o minha lista de assuntos não abordados e quem fazem parte dos trabalhos que darão continuidade ao já realizado. Depois desse capítulo, finalmente, encontram-se as referências bibliográficas e os anexos.

2 Revisão da Literatura

Pode-se encontrar vasta bibliografia em que o tema análise de agrupamentos é abordado. Há livros específicos tais como Anderberg (1973), Hartigan (1975), Kaufman e Rousseeuw (1990) e Everitt (1993), e outros que abrangem todas as técnicas de análise multivariada de dados, o que inclui a de agrupamentos. São exemplos desse último tipo Jobson (1991-92), Johnson (1998), Johnson e Wichern (2002), Timm (2002) e Mingoti (2005), dentre outros. Muitos deles apresentam medidas de proximidade para observações de variáveis categóricas, alguns trazem sugestões quando há a ocorrência também de variáveis contínuas, mas apenas o último cita um método especificamente voltado para o agrupamento de variáveis categóricas. Assim, a dissertação baseou-se em grande parte em artigos, a serem comentados a seguir, por metodologia. Na primeira seção do próximo capítulo a proposta original de tais métodos será detalhadamente comentada.

2.1 k-Médias, k-Modas e k-Protótipos

Ralambondrainy (1995) desenvolveu a primeira abordagem do algoritmo das k-Médias para tratar variáveis categóricas. Nela, variáveis binárias representando presença e ausência dos atributos são consideradas, com as principais desvantagens: (1) em *data mining*, onde as variáveis possuem muitas categorias, a recodificação delas para binárias acarretaria na criação de muitas outras variáveis, aumentando consideravelmente o custo computacional, tanto para a implementação quanto para a alocação do volume de informação e (2) ao se calcularem as médias, valores no intervalo [0,1] não corresponderiam a nenhuma característica específica.

Huang (1997, 1998) propõe basicamente três modificações no algoritmo das k-Médias: a discordância simples como medida de dissimilaridade, modas ao invés de médias e um método baseado em freqüências para atualizar essas modas. Os resultados experimentais obtidos com o k-Modas são satisfatórios. Entretanto, é possível apontar duas desvantagens na implementação dessa metodologia: a inicialização do algoritmo e a não-unicidade da moda do grupo.

A inicialização do algoritmo é um fator de especial importância porque existe uma

dependência entre as sementes iniciais e os grupos obtidos ao final do processamento. Por isso, Sun *et al.* (2002) publicaram um estudo experimental da aplicação do algoritmo iterativo de refinamento dos pontos iniciais de *Brandley* e *Fayyad* (BRADLEY; FAYYAD, 1998) ao algoritmo das k-Modas. A técnica é complexa e demanda grande tempo de processamento, mas os autores afirmam que a acurácia e a confiabilidade dos resultados tornam-se melhores.

San *et al.* (2004) apresentam uma alternativa para o problema da não-unicidade da moda do grupo e a implementam em um algoritmo chamado k-Representativos. Nesse artigo os autores aplicam a noção de *fuzziness* para definir o "centro" ou o "representativo" dos grupos. A partir dessa nova formulação, fez-se necessário propor um novo método para se calcular a dissimilaridade entre as observações e os grupos formados.

Essas modificações do k-Modas não foram implementadas durante o desenvolvimento desta dissertação e assim não fizeram parte dos estudos de simulação e de aplicação elaborados. Para mais detalhes sobre essas metodologias sugere-se que os artigos de referência sejam consultados.

Huang (1997, 1998) propõe ainda o k-Protótipos, que faz a integração dos métodos k-Médias e k-Modas e usa uma medida de proximidade combinada para agrupar observações tanto de variáveis categóricas quanto de contínuas. As desvantagens presentes no k-Modas e que se referem à inicialização do algoritmo e à não-unicidade da moda também se apresentam aqui. Acredita-se que as modificações desenvolvidas por Sun *et al.* (2002) e San *et al.* (2004) também podem ser aplicadas ao k-Protótipos. Cabe ressaltar no entanto que nenhum estudo foi encontrado na literatura com essa abordagem.

2.2 *Fuzzy c*-Modas

Proposto em Huang e Ng (1999), o algoritmo *Fuzzy* c-Modas é uma generalização do algoritmo k-Modas. A principal contribuição dessa metodologia é a obtenção de graus de pertinência das observações aos grupos. Tais valores podem ser usados para estudar os pontos do interior e da fronteira dos grupos formados. Analogamente ao k-Modas, utilizam-se a discordância simples como medida de dissimilaridade e a moda ao invés de média para representar o centro do grupo. A diferença está em um novo procedimento que foi desenvolvido para gerar uma matriz de partição *fuzzy* a partir de um banco de dados caracterizados por variáveis categóricas.

Kim *et al.* (2004) e Kim *et al.* (2005) possuem abordagens do algoritmo *fuzzy* bastante

próximas. O primeiro artigo desenvolve os *fuzzy* centróides, que mantêm a informação a respeito da composição da moda do grupo de uma iteração para outra. Segundo os autores, isso dificulta a existência de erros de classificação e evita que o algoritmo encontre um mínimo local. No segundo artigo, o algoritmo k-Populações é proposto. A principal diferença entre esses dois artigos é o fato do segundo (2005) incorporar fatores de normalização na obtenção da medida de proximidade e na atualização dos *fuzzy* centróides. O objetivo dos autores era resolver o problema da não-unicidade da moda dos grupos.

Os artigos de Oh *et al.* (2001) e Umayahara *et al.* (2005) propõem novas formulações do algoritmo *Fuzzy*, cujos resultados se assemelham bastante àqueles da análise de correspondência (GREENACRE, 1984). As conclusões a que os autores do primeiro artigo chegam são análogas às obtidas por meio da análise de correspondência. Eles afirmam que essa nova metodologia tem a vantagem de não ser complexa, uma vez que utiliza cálculos algébricos simples. O método do segundo artigo é aplicável a um tipo de banco de dados denominado *multisets*. Um exemplo dessa aplicação é um banco em que as linhas são livros, as colunas, palavras-chave e as entradas, o número de vezes que as palavras ocorrem em cada um dos livros. Nesse artigo encontram-se também novas métricas denominadas "vetores dos graus de pertinência ponderados" e os "vetores unitários", ambos necessários à nova formulação.

Essas modificações do método *Fuzzy* que se aproximam da análise de correspondência fogem do escopo desta dissertação. De novo, recomenda-se que os artigos que apresentam essas metodologias sejam consultados, caso o leitor tenha interesse em estudá-las.

2.3 *ROCK* e *QROCK*

Diferentemente de todos os algoritmos encontrados, o ROCK (GUHA *et al.*, 2000) incorpora uma nova concepção à medida de proximidade entre as observações, os grupos de observações ou entre ambos. *ROCK* é um algoritmo que emprega *links* ao agrupar os elementos em análise. Além da definição de *links*, também são apresentados nesse artigo as definições de "vizinho", da "função objetivo" e da "medida de qualidade" (*goodness measure*), além dos passos para a implementação da metodologia. Em Dutta *et al.* (2005), é provado que, sob certas condições, os grupos obtidos através do *ROCK* são nada mais que os vértices conectados de um grafo. Segundo os autores, o novo método, chamado *QROCK*, contribui significativamente para a redução do tempo de processamento do *ROCK*. Propõe-se ainda uma nova medida de similaridade entre observações ou grupos

de observações quando esses são caracterizados apenas por variáveis categóricas.

As conclusões a que os autores chegaram em Guha *et al.* (2000) demonstram que o *ROCK* tem boas propriedades e os resultados obtidos a partir da sua aplicação são de melhor qualidade. A arbitrariedade na escolha do parâmetro envolvido no método (θ) e o desempenho do algoritmo no que se refere à complexidade e ao tempo de processamento são as dificuldades encontradas na sua implementação. Pantuzzo (2002) sugere um procedimento empírico para a escolha de θ. Tal procedimento baseia-se nos valores que a medida de proximidade entre as observações pode assumir. Além desses valores, outros pontos do domínio de θ foram testados em bancos de dados reais. Nos casos estudados por Pantuzzo, o *ROCK* tem desempenho bastante satisfatório. Entretanto, o método é muito dependente da escolha do parâmetro θ.

2.4 Outros Métodos

Não serão abordadas nesta dissertação os métodos *STIRR*, *CACTUS*, *LIMBO*, *M-BILCOM* e *BILCOM*. Mais uma vez sugere-se aos interessados que a referência bibliográfica presente no final da dissertação seja consultada.

Em resumo, *STIRR*, apresentado em Gibson *et al.* (2000), baseia-se em um método iterativo que utiliza atribuição e propagação de pesos. Os autores afirmam que o *STIRR* pode ser visto como uma generalização das técnicas de particionamento espectral aplicadas ao problema de agrupamento de hipergrafos e utiliza o que eles chamam de sistemas dinâmicos não-lineares. A idéia central do *CACTUS* (GANTI *et al.*, 1999) está relacionada ao fato de que um resumo das informações contidas no banco de dados é suficiente para se descobrirem todos os grupos bem-definidos.

O *LIMBO* (ANDRITSOS, 2004) utiliza a Teoria da Informação e constrói uma quantidade denominada *Information Bottleneck*, que definirá uma nova medida de proximidade entre as observações. Cabe notar que esse método pode ser aplicado para agrupar observações tanto de variáveis categóricas quanto contínuas. Assim como os métodos *M-BILCOM* (ANDREOPOULOS *et al.*, 2005) e *BILCOM* (ANDREOPOULOS *et al.*, 2006). O primeiro é aplicado no contexto em que não há certeza se as variáveis categóricas possuem os valores corretos. Por isso utilizam-se valores de confiança que variam entre 0.0 e 1.0 e que indicam o grau de certeza de correção desses atributos. Finalmente, a segunda metodologia incorpora um processo pseudo-Bayesiano em que a priori é os grupos formados pela análise de agrupamentos considerando apenas as variáveis categóricas.

3 *Metodologias de Análise de Agrupamentos*

3.1 Descrição das Metodologias

Nesta seção, descrevem-se as técnicas de análise de agrupamentos que foram consideradas para a dissertação. Elas foram escolhidas devido à sua empregabilidade, à maneira como as variáveis categóricas são tratadas e ao fato de alguns desses métodos basearem-se em algoritmos bastante eficientes no tratamento de variáveis contínuas.

Devido à existência de diversos métodos, faz-se necessário padronizar a notação. Supõe-se que se deseja organizar n objetos em k grupos. Esse objetos são caracterizados por m variáveis categóricas e por $(p - m)$ variáveis contínuas. No total existem então p variáveis, que serão denotadas por $A_1, \ldots, A_m, B_1, \ldots, B_{p-m}$.

O conjunto de todas as possíveis categorias que uma variável categórica A_j pode assumir será denotado por $DOM\,(A_j)$, com $j \in [1, m]$. Esse conjunto é chamado domínio de A_j e a sua cardinalidade, $|DOM\,(A_j)\,| = n_{A_j}$, representa o número de possíveis categorias da *j-ésima* variável. Quando uma determinada categoria t de A_j precisar ser especificada, isso será feito por meio de $c_{t,j}$, com $t \in DOM\,(A_j)$.

Assim, podemos representar o conjunto de observações por uma matriz de dimensão $(n \times p)$, $X = \{X_1, X_2, \ldots, X_n\}'$, de tal forma que para $i \in [1, n]$,

$$X_i = \{x_{i,A_1}, \ldots, x_{i,A_m}, x_{i,B_1}, \ldots, x_{i,B_{p-m}}\}$$

Assim, X_q e X_r representam a *q-ésima* e a *r-ésima* observações de X.

3.1.1 Algoritmo Hierárquico Aglomerativo

Os algoritmos hierárquicos diferem dos não-hierárquicos basicamente em dois aspectos: nos primeiros, uma vez que observações ou grupos de observações foram combinados, eles não se separarão. Por outro lado, o segundo tipo permite que esses elementos sejam

deslocados de um grupo para outro. Outra distinção fundamental entre os procedimentos é a escolha do número de grupos. Para o primeiro tipo de algoritmo, uma forma usual de se estimar o número de grupos da partição é analisar um gráfico chamado dendograma enquanto que para o segundo, esse número deve ser previamente fixado. Além do dendograma existem outros critérios para se decidir quantos grupos são necessários (EVERITT, 1993; JOHNSON, 1998; MINGOTI, 2005).

Os procedimentos hierárquicos podem ainda ser classificados em outros dois subgrupos: algoritmos aglomerativos e algoritmos divisivos. Como os próprios nomes sugerem, um método do tipo aglomerativo começa o agrupamento com a quantidade de grupos igual ao número de objetos a serem agrupados ($k = n$) e termina com todos eles em um único grupo ($k = 1$). O divisivo, ao contrário, começa com todos os objetos unidos e que são separados até que cada objeto esteja em um único grupo (Figura 3.1).

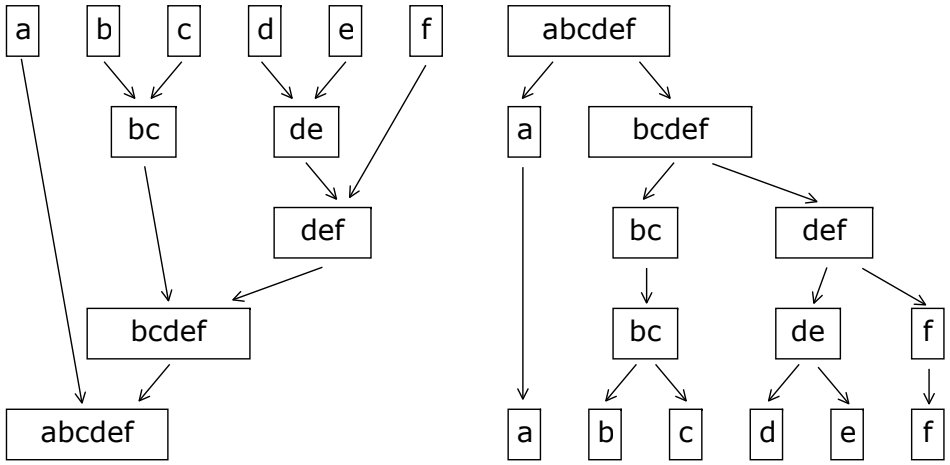

Figura 3.1: Esquema ilustrativo de algoritmos hierárquicos aglomerativo (esquerda) e divisivo (direita)

.

Os algoritmos divisivos têm um custo computacional elevado (REIS, 1997), pois, segundo Kaufman e Rousseeuw (1990), o primeiro passo é dividir todas as observações em dois grupos e para que isso seja feito é necessário considerar todas as possíveis divisões. Essa separação pode não ser possível de ser feita devido ao grande número de combinações, especialmente para grandes bancos de dados. Por isso tais métodos não são tão

empregados quanto os aglomerativos. Everitt (1993) descreve os dois tipos de métodos divisivos, o *monothetic* e o *polythetic*, por meio de exemplos. No *monothetic* a divisão é feita com base em uma única variável enquanto que no *polythetic*, todos os atributos são usados conjuntamente.

Os algoritmos hierárquicos aglomerativos seguem os seguintes passos:

1. definir, se necessário, a quantidade de grupos k a serem obtidos;
2. encontrar o par de observações ou grupos de observações mais similares, ou seja, mais próximos;
3. combinar esse par em um único grupo e recalcular, sob algum critério, as novas medidas de proximidade;
4. repetir os passos 2 e 3 até que o número de grupos fixado seja alcançado ou até que todos os objetos estejam em um mesmo grupo ($k = 1$);
5. identificar os grupos unidos a cada iteração do algoritmo e também o valor da medida de proximidade no momento em que eles foram agrupados, caso o critério de parada tenha sido o segundo.

O critério especificado no passo 2 é o que define os algoritmos hierárquicos aglomerativos tradicionais que são os métodos "ligação simples" (*single linkage*), "ligação completa" (*complete linkage*), "ligação média" (*average linkage*), "centróide" (*centroid method*) e "*Ward*", entre outros. De forma geral, a cada passo do algoritmo, a matriz que contém as proximidades é atualizada de modo a conter, naquele passo, as medidas de proximidade entre os grupos formados e as demais observações ou os grupos de observações. Essa atualização se dá conforme o método escolhido.

Sejam dois grupos G_1 e G_2, com n_{G_1} e n_{G_2} elementos respectivamente e $\mathcal{L}$ uma lista de todos os pares de elementos entre esses grupos. Nos métodos de ligação simples (ou do vizinho mais próximo) e completa (ou do vizinho mais distante), a distância entre G_1 e G_2 é definida como o mínimo e o máximo, respectivamente, das distâncias entre os elementos de $\mathcal{L}$. No método da ligação média, a distância entre os grupos é a média das distâncias entre todos os pares pertencentes a $\mathcal{L}$. Nos métodos centróide e *Ward*, cada grupo passa a ser caracterizado pelo seu vetor de médias e a distância entre G_1 e G_2 é dada pela distância entre tais vetores. Maiores detalhes podem ser encontrados em Everitt (1993).

Objetivando-se agrupar observações de variáveis categóricas ou de ambos os tipos de variáveis, não se recomenda o uso dos métodos centróide e *Ward*, pois se baseiam

em médias e tais quantidades não fazem sentido nesse contexto. Assim, como algoritmo representativo da classe dos hierárquicos aglomerativos, escolheu-se o da ligação média, que será melhor descrito a seguir.

3.1.1.1 Método da Ligação Média

Nesse método, a proximidade entre dois grupos de observações é dada pela média das medidas de proximidade entre todas as combinações de observações desses grupos, ou seja, pela média de todas as $n_{G_1} \times n_{G_2}$ medidas existentes entre todos os pares de elementos pertencentes aos grupos G_1 e G_2, isto é,

$$d\left(G_1, G_2\right) = \frac{\sum_{X_q \in G_1} \sum_{X_r \in G_2} d\left(X_q, X_r\right)}{n_{G_1} \times n_{G_2}} \tag{3.1}$$

Para ilustrar essa metodologia, consideremos o exemplo a seguir:

Exemplo 1 *Seja D a matriz de dissimilaridade de $n = 6$ elementos, isto é,*

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 & & & & & \\ 0,3 & 0 & & & & \\ 12,2 & 9,2 & 0 & & & \\ 12,7 & 9,9 & 0,4 & 0 & & \\ 2,3 & 1,9 & 15,2 & 17,1 & 0 & \\ 9,8 & 7,4 & 4,3 & 3,4 & 13,8 & 0 \end{pmatrix}$$

As duas primeiras observações devem ser agrupadas por possuírem a menor medida de dissimilaridade, ou seja, por estarem mais próximas. Assim, a nova matriz será dada por:

$$\mathbf{D} = \begin{array}{c} \{1;2\} \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 & & & & \\ 10,7 & 0 & & & \\ 11,3 & 0,4 & 0 & & \\ 2,1 & 15,2 & 17,1 & 0 & \\ 8,6 & 4,3 & 3,4 & 13,8 & 0 \end{pmatrix}$$

em que o valor $10,7$ é obtido pela Equação (3.1), ou seja, $(12,2 + 9,2)/2$. As demais medidas da primeira coluna, referente ao grupo que contém as observações 1 e 2, são

obtidas de forma análoga. Conclui-se que as próximas observações a serem agrupadas são aquelas que possuem como medida de dissimilaridade o valor $0,4$*, ou seja, forma-se o grupo* $\{3,4\}$.

Definido como o procedimento de agrupamento deve ser realizado, basta agora estabelecer as medidas de proximidade entre observações de variáveis categóricas, exclusivamente, ou também de variáveis contínuas.

3.1.1.2 Medidas de Proximidade

Antes de introduzirmos as diferentes medidas de proximidade no contexto de interesse, vale a pena ressaltar que quando o estudo contém $(p-m)$ variáveis contínuas, a medida de dissimilaridade mais usada na literatura, devido à sua interpretação geométrica, é a distância euclidiana, definida para as observações X_q e X_r como:

$$d\left(X_q, X_r\right) = \left[\sum_{j=1}^{p-m} \left(x_{q,j} - x_{r,j}\right)^2\right]^{\frac{1}{2}} \tag{3.2}$$

A distância euclidiana é um caso particular da métrica de *Minkowski*, também conhecida como Norma L_p:

$$d\left(X_q, X_r\right) = \left(\sum_{j=1}^{p-m} \left|x_{q,j} - x_{r,j}\right|^{\lambda}\right)^{\frac{1}{\lambda}} \tag{3.3}$$

A métrica chamada *city-block*, *Manhattan* ou Normal L_1 é obtida fazendo-se $\lambda = 1$ na Equação (3.3). Há também a distância euclidiana ponderada e a de *Mahalanobis*, usadas quando as variáveis contínuas possuem escalas diferentes ou quando essas são correlacionadas. Ao empregar essas últimas medidas de dissimilaridade os efeitos da escala e da correlação são anulados e isso pode excluir alguma distinção natural entre as observações, mascarando os grupos existentes (ANDERBERG, 1973; JOBSON, 1991-92; REIS, 1997).

Para observações de variáveis categóricas binárias, muitas são as medidas de proximidade propostas. Entre elas, as mais empregadas são o coeficiente de concordância simples ou de *Sokal* e o de *Jaccard* (DUDA; HART, 1973; EVERITT, 1993). Suponhamos que duas observações sejam caracterizadas por m variáveis que assumem o valor 1 se uma determinada característica está presente e 0 caso contrário. A Tabela 3.1 resume as informações da matriz de observações X:

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccc} A_1 & A_2 & A_3 & \dots & A_m \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 1 \end{pmatrix} \end{array}$$

Tabela 3.1: Resumo das informações

Categoria	1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	a+c	b+d	m=a+b+c+d

As quantidades a e d representam o número de variáveis em que as observações comparadas coincidem quanto à presença ou à ausência da característica de interesse, respectivamente. Os valores b e c, por outro lado, representam o número de variáveis em que os elementos analisados pertencem a categorias diferentes. O coeficiente de concordância simples, dado por $\frac{a+d}{m}$, calcula a proporção de variáveis em que os objetos comparados têm o mesmo código. Nesse caso, são atribuídos pesos iguais tanto à presença (par 1-1) quando à ausência (par 0-0) de um determinado atributo. Ou seja, ambos os casos são considerados concordâncias.

Há situações, entretanto, em que o par 0-0 não representa uma concordância. O exemplo a seguir foi adaptado de Anderberg (1973) e é apresentado no sentido de elucidar a afirmação anterior: suponhamos que animais precisam ser agrupados segundo duas variáveis, a saber, "possuir penas" e "possuir bico". O par 0-0 indica, dessa forma, a ausência de penas e de bico e contribuiria para que cães e gatos fossem considerados similares caso esse par fosse, nesse caso, uma tipo de concordância. Nesse caso, o coeficiente de *Jaccard*, dado por $\frac{a}{m-d}$, deve ser empregado, uma vez que desconsidera o par 0-0 no cálculo da medida de proximidade. Na seção 3.1.4.4 uma nova medida de similaridade é apresentada e ela se baseia no coeficiente de *Jaccard*, que sofre pequenas alterações. A título de ilustração, a Tabela 3.2 traz algumas medidas de similaridade encontradas na literatura (JOHNSON; WICHERN, 2002):

Tabela 3.2: Coeficientes e situações em que se aplicam

Coeficiente	Situação em que se aplica
$\frac{2(a+d)}{2(a+d)+(b+c)}$	Concordância dupla: duplo peso para pares concordantes
$\frac{(a+d)}{(a+d)+2(b+c)}$	*Tanimoto-Rogers*: duplo peso para pares discordantes
$\frac{a}{m}$	*Russel-Rao*: proporção de pares 1-1
$\frac{2a}{2a+b+c}$	*Czekanowski-Sorensen-Dice*: duplo peso para pares 1-1 e nenhum para 0-0
$\frac{a}{b+c}$	*Kulczynski*: razão do par 1-1 com os pares discordantes
$\frac{a}{a+2(b+c)}$	Duplo peso para pares discordantes e nenhum para 0-0
$\frac{a+d}{b+c}$	Razão dos pares concordantes com os discordantes

Para variáveis com mais de duas categorias, a estratégia usualmente adotada é a utilização de variáveis indicadoras (*dummies*). Transforma-se cada categoria das variáveis em uma nova variável de tal forma que 1 indicaria a presença daquela categoria e 0 a ausência. Assim, todos os coeficientes de similaridade descritos nos parágrafos anteriores podem ser usados. Cabe observar, no entanto, que o total da tabela-resumo não será o número de variáveis m, mas o número de categorias existentes em todas as m variáveis, $\sum_{j=1}^{m} n_{A_j}$ (JOBSON, 1991-92). Essa abordagem implica no aumento do número de variáveis que caracterizam os objetos e pode ser inviável caso sejam muitas as categorias originais. O exemplo a seguir ilustra nesse caso o cálculo dos coeficiente de concordância simples e o de *Jaccard*:

Exemplo 2 *Suponhamos* 3 *observações* X_1, X_2 *e* X_3 *caracterizadas por 4 variáveis categóricas, cujos domínios são:* $DOM(A_1) = \{a, b\}$, $DOM(A_2) = \{c, d, e\}$, $DOM(A_3) = \{f, g\}$ *e* $DOM(A_4) = \{h, i\}$. *São estas:* $X_1 = \{a, c, f, h\}$, $X_2 = \{a, d, g, i\}$ *e* $X_3 = \{b, e, f, h\}$.

$$\mathbf{X} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & A_3 & A_4 \end{array} \\ \begin{pmatrix} a & c & f & h \\ a & d & g & i \\ b & e & f & h \end{pmatrix} \end{array} ; \mathbf{X_{bin}} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

O coeficiente de concordância simples será:

$$\begin{aligned} s(X_1, X_2) &= \frac{3}{9} = \frac{1}{3} \approx 0,33 \\ s(X_1, X_3) &= \frac{5}{9} \approx 0,56 \end{aligned}$$

e o de Jaccard:

$$\begin{aligned} s(X_1, X_2) &= \frac{1}{7} \approx 0,14 \\ s(X_1, X_3) &= \frac{2}{6} = 0,33 \end{aligned}$$

Utilizando qualquer um dos coeficientes percebe-se que as observações X_1 e X_3 são mais similares que as observações X_1 e X_2. Como o coeficiente de Jaccard desconsidera os pares 0-0, as medidas de similaridade diminuem uma vez que esses pares ocorrem entre as observações analisadas.

3.1.1.2.1 Medidas de Proximidade Combinada .

Duas são as alternativas existentes na literatura para combinar as medidas de proximidade entre observações de variáveis categóricas e contínuas. Inicialmente, é necessário calcular tais quantidades para cada tipo de variável e, em seguida, combiná-las usando pesos adequados. Suponhamos que entre as p variáveis existam m categóricas e, conseqüentemente, $(p-m)$ contínuas. A primeira alternativa (MINGOTI, 2005) é dada por:

$$d_{comb}(X_q, X_r) = w_{cat} \times d_{cat}(X_q, X_r) + w_{cont} \times d_{cont}(X_q, X_r), \tag{3.4}$$

em que os sub-índices *cat* e *cont* referem-se aos tipos de variáveis. A medida de proximidade definida combina então as medidas de proximidade entre as variáveis categóricas (d_{cat}) e contínuas (d_{cont}) que caracterizam as observações X_q e X_r de tal forma que as quantidades w ponderam essas medidas.

É importante que as medidas de proximidade d_{cat} e d_{cont} tenham a mesma direção e que estejam definidas no mesmo intervalo de variação. A determinação dos pesos w_{cat} e w_{cont} é a principal dificuldade na obtenção dessa quantidade. É usual utilizar a proporção de variáveis de cada tipo como pesos.

Na segunda alternativa, proposta em Gower (1971), a medida de proximidade precisa estar definida no intervalo $[0, 1]$. Assim, a medida combinada é a seguinte:

$$d_{comb}(X_q, X_r) = \frac{\sum_{j=1}^{p} \mathbf{1}_j(X_q, X_r) \times d(X_q, X_r)}{\sum_{j=1}^{p} \mathbf{1}_j(X_q, X_r)} \tag{3.5}$$

em que $\mathbf{1}_j(X_q, X_r)$ é uma variável indicadora que assume o valor 1 se as observações X_q e X_r puderem ser comparadas com respeito à variável j e assume o valor 0 caso contrário.

Para as duas alternativas de combinação, caso uma dissimilaridade precise ser trans-

forma para ter então a mesma direção e estar definida no mesmo intervalo de variação de uma similaridade, segundo Mingoti (2005), basta utilizar as seguintes equações. Seja $d(X_q, X_r)$ uma medida de dissimilaridade entre duas observações X_q e X_r. A medida de similaridade será dada por:

$$s(X_q, X_r) = 1 - d^*(X_q, X_r) \tag{3.6}$$

$$d^*(X_q, X_r) = \frac{d(X_q, X_r) - min(D)}{max(D) - min(D)}, \tag{3.7}$$

em que D é a matriz de dissimilaridades entre todos os elementos (matriz do primeiro passo), $min(D)$ e $max(D)$ correspondem, nessa ordem, ao mínimo e ao máximo das entradas da matriz D.

A mesma transformação (Equações (3.6) e (3.7)) será usada nesta dissertação na obtenção da medida de proximidade combinada dos métodos k-Protótipos e ROCK estendido, a serem apresentados neste capítulo. Ao aplicá-la, não se faz necessário preocupar-se com a padronização das variáveis contínuas que possuem diferentes escalas. Essa é uma vantagem bastante interessante uma vez que a padronização, apesar de às vezes ser necessária, pode diluir as diferenças existentes entre os grupos (EVERITT, 1993).

A variável indicadora presente na Equação (3.5) proporciona uma vantagem à combinação proposta por *Gower*. Usando essa alternativa, todas as observações podem ser analisadas, mesmo aquelas que possuem valores faltantes (*missings*). Em geral, observações com essa característica são descartadas da análise. Assim, no coeficiente de *Gower* consideram-se, para cada par de observações, aquelas variáveis em que há informação.

Nota-se outra utilidade do coeficiente de *Gower* na aplicação a variáveis binárias. Nessa situação, a variável indicadora poderá assumir o valor zero quando ocorrer o par 0-0 entre as observações comparadas. Assim, será possível excluir esse par do cálculo da medida de similaridade (EVERITT, 1993).

3.1.2 *k*-Modas e *k*-Protótipos

O algoritmo k-Modas (*k-Modes*) é uma modificação do k-Médias, que é um eficiente procedimento aplicado ao caso em que somente variáveis contínuas caracterizam os objetos em análise. Essa metodologia é apresentada em Huang (1998) e utiliza o coeficiente de disconcordância simples como medida de proximidade, modas ao invés de médias e um método baseado em freqüências para atualizar essas modas. A seguir, cada um das

mudanças implementadas são comentadas.

3.1.2.1 *k*-Modas

3.1.2.1.1 Medida de Proximidade .

Sejam duas observações X_q e X_r. O coeficiente de discordância simples entre X_q e X_r (Equação (3.8)) corresponde ao total de discordâncias que ocorrem ao se analisar essas duas observações com respeito às m variáveis categóricas. Ou seja, quanto menor o valor dessa medida, mais similares são X_q e X_r.

$$d\left(X_q, X_r\right) = \sum_{j=1}^{m} \delta\left(x_{q,j}, x_{r,j}\right) \tag{3.8}$$

em que

$$\delta\left(x_{q,j}, x_{r,j}\right) = \begin{cases} 0, & \text{se } x_{q,j} = x_{r,j} \\ 1, & \text{se } x_{q,j} \neq x_{r,j} \end{cases} \tag{3.9}$$

Essa quantidade pode ser transformada no coeficiente de concordância simples, definido na seção 3.1.1.2. Basta, para isso, definir a dissimilaridade entre X_q e X_r como:

$$d^{-}\left(X_q, X_r\right) = \frac{1}{m} \times \sum_{j=1}^{m} \delta\left(x_{q,j}, x_{r,j}\right) = \frac{1}{m} \times \ d\left(X_q, X_r\right) \tag{3.10}$$

Assim, o coeficiente de concordância simples será:

$$s\left(X_q, X_r\right) = 1 - d^{-}\left(X_q, X_r\right) \tag{3.11}$$

O exemplo numérico anterior é repetido para facilitar o entendimento da medida de dissimilaridade utilizada no método k-Modas.

Exemplo 3 *Suponhamos* 3 *observações* X_1, X_2 *e* X_3 *caracterizadas por 4 variáveis categóricas, cujos domínios são:* $DOM\left(A_1\right) = \{a, b\}$, $DOM\left(A_2\right) = \{c, d, e\}$, $DOM\left(A_3\right) = \{f, g\}$ *e* $DOM\left(A_4\right) = \{h, i\}$. *São estas:* $X_1 = \{a, c, f, h\}$, $X_2 = \{a, d, g, i\}$ *e* $X_3 = \{b, e, f, h\}$. *Baseado em variáveis binárias (ver matriz* X *na página 16), o coeficiente de discordância será, então :*

$$\begin{aligned} d\left(X_1, X_2\right) &= 6 \\ d\left(X_1, X_3\right) &= 4 \end{aligned}$$

3.1.2.1.2 Moda do Grupo .

Seja G um conjunto de n_G observações caracterizadas por m variáveis categóricas. A moda de $G = \{X_1, \ldots, X_{n_G}\}'$, é o vetor $Q_g = \{q_{g,1}, q_{g,2}, \ldots, q_{g,m}\}$ que minimiza

$$M(G, Q_g) = \sum_{i=1}^{n_G} d(X_i, Q_g) \tag{3.12}$$

Teorema 1 *A função $M(G, Q_g)$ é minimizada se e somente se $f_{q_{g,j}|G} \geq f_{c_{t,j}|G}$ para $q_{g,j} \neq c_{t,j}$ $\forall t|j$ $\forall j$, em que $f_{c_{t,j}|G} = f(A_j = c_{t,j}|G)$ é a freqüência relativa da categoria t da variável A_j em G, isto é, $f_{c_{t,j}|G} = n_{c_{t,j}}/n_G$ em que $n_{c_{t,j}}$ é o número de objetos em G que possuem a categoria t no atributo A_j. Analogamente, $f_{q_{g,j}|G} = f(A_j = q_{g,j}|G)$ é a freqüência relativa da categoria da variável j que irá compor a moda do grupo G.*

Pela definição e pelo Teorema 1 (HUANG, 1998) o vetor Q_g é definido pelas categorias de maior freqüência $q_{g,j}$, $j \in [1, m]$, podendo não ser um elemento de G e ainda podendo não ser único. Caso exista mais de uma categoria para compor a moda do grupo, Huang e Ng (1999) sugerem que a escolha seja arbitrária e que a categoria que primeiro minimiza a função $M(G, Q_g)$ seja aquela que fará parte de Q_g.

3.1.2.1.3 Algoritmo *k*-Modas .

A função objetivo (ou custo) que deve ser minimizada pelo algoritmo das k-Modas é dada por:

$$\begin{aligned} P(W, \mathcal{Q}) &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} \times d(X_i, Q_l) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,l} \times \delta(x_{i,j}, q_{l,j}) \\ \text{sujeito a} \quad & w_{i,l} \in \{0, 1\} \ , \ i \in [1, n] \ \text{ e } l \in [1, k] \\ & \sum_{l=1}^{k} w_{i,l} = 1 \ , \ i \in [1, n] \end{aligned} \tag{3.13}$$

em que $\mathcal{Q} = \{Q_1, Q_2, \ \ldots \ , Q_l\}$, com $Q_l = [q_{l,1}, q_{l,2}, \ldots, q_{l,m}]$, $w_{i,l}$ é o elemento da matriz de partição W que indica se a *i-ésima* observação pertence ($w_{i,l} = 1$) ou não ($w_{i,l} = 0$) ao grupo l, k é o número de grupos e n é o número de observações.

Na prática, os seguintes passos podem ser seguidos:

1. definir a quantidade de grupos k a serem obtidos e selecionar k sementes iniciais;

2. alocar cada observação a um dos k grupos, de acordo com a medida de proximidade adotada. Após cada realocação, atualizar as modas dos novos grupos;

3. comparar novamente as observações às modas atualizadas, depois que todos os objetos tiverem sido alocados. Caso existam observações que estejam mais próximas a outras modas que não àquelas do próprio grupo, realocá-las e, então, atualizar novamente as modas dos grupos formados;

4. repetir o passo anterior até que não seja possível fazer nenhuma outra alocação.

Huang (1998) sugere ainda duas formas de se inicializar o algoritmo. Na primeira, escolhe-se aleatoriamente as k sementes e na segunda, implementam-se os seguintes passos:

1. calcular as freqüências para todas as categorias de todas as variáveis e armazená-las em uma matriz em que cada coluna corresponde a uma variável e as linhas representam as categorias, organizadas em ordem decrescente;

2. atribuir a cada uma das k sementes as categorias mais freqüentes, combinando-as para todas as variáveis;

3. substituir a *k-ésima* semente pela observação mais similar a essa, até que todas sejam redefinidas.

O artigo de Huang (1998) traz o seguinte exemplo para o passo 2:

Exemplo 4 *Suponhamos que as categorias de 4 variáveis já estejam organizadas em ordem decrescente na matriz M a seguir:*

$$\mathbf{M} = \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & A_3 & A_4 \end{array} \\ \begin{pmatrix} c_{(1),1} & c_{(1),2} & c_{(1),3} & c_{(1),4} \\ c_{(2),1} & c_{(2),2} & c_{(2),3} & c_{(2),4} \\ c_{(3),1} & & c_{(3),3} & c_{(3),4} \\ c_{(4),1} & & c_{(4),3} & \\ & & c_{(5),3} & \end{pmatrix} \end{array}$$

Ou seja $f_{c_{(1),j}} \geq f_{c_{(2),j}} \geq \ldots \geq f_{c_{\left(\left|DOM(A_j)\right|\right),j}}$, *em que* $f_{c_{(t),j}}$ *é a freqüência relativa da categoria* t *da variável* A_j, *com* $j \in [1, m]$ *e* $t \in [1, DOM(A_j)]$. *Suponhamos ainda que*

se precisa definir 3 sementes iniciais. A sugestão do autor é esta:

$$\begin{aligned} Q_1 &= \{q_{1,1} = c_{(1),1}, q_{1,2} = c_{(2),2}, q_{1,3} = c_{(3),3}, q_{1,4} = c_{(1),4}\} \\ Q_2 &= \{q_{2,1} = c_{(2),1}, q_{2,2} = c_{(1),2}, q_{2,3} = c_{(4),3}, q_{2,4} = c_{(2),4}\} \\ Q_3 &= \{q_{3,1} = c_{(3),1}, q_{3,2} = c_{(2),2}, q_{3,3} = c_{(1),3}, q_{3,4} = c_{(3),4}\} \end{aligned}$$

É possível perceber que não parece haver uma regra geral para a composição das k sementes, podendo ocorrer arbitrariedade. Essa inicialização mostra-se ineficiente quando se tem um grande número de atributos caracterizando as observações a serem agrupadas e inviável de ser implementada em uma simulação. No entanto, os resultados obtidos no artigo demonstram que a partir do método das k-Modas resultados satisfatórios são obtidos. Além disso, o autor conclui que a metodologia é eficiente ao ser aplicada a bancos de dados muito grandes e complexos.

3.1.2.2 *k*-Protótipos

A metodologia do k-Protótipos foi apresentada em Huang (1997, 1998) e tem por objetivo realizar a análise de agrupamentos em observações tanto de variáveis categóricas quanto de contínuas. Para isso, esse método faz a integração dos métodos k-Médias e k-Modas e utiliza uma medida de proximidade combinada para comparar as observações analisadas.

3.1.2.2.1 Medida de Proximidade .

Sejam duas observações X_q e X_r da matriz X definida no início do capítulo. A medida de proximidade combinada é dada por:

$$d_{comb}\left(X_q, X_r\right) = \xi \times \sum_{j=1}^{m} \delta\left(x_{q,j}, x_{r,j}\right) + \sum_{j=m+1}^{p} \varphi\left(x_{q,j}, x_{r,j}\right), \tag{3.14}$$

em que o primeiro termo refere-se à medida de proximidade para variáveis categóricas, podendo ser a discordância simples (Equação (3.8)) e o segundo termo, à medida para variáveis contínuas. No artigo sugere-se o quadrado da distância euclidiana, ou seja, o quadrado da Equação (3.2). Assim, a medida será uma dissimilaridade combinada. O peso $\xi \in \mathcal{R}$ é usado para evitar o favorecimento de um ou outro tipo de variável. Informações mais detalhadas sobre a influência dessa quantidade no procedimento podem ser vistas em Huang (1998), mas, resumidamente, tem-se que grandes valores de ξ indicam que o procedimento é dominado por variáveis categóricas e, por outro lado, pequenos valores

indicam que as contínuas são mais importantes para a análise.

3.1.2.2.2 Função Objetivo Modificada .

Devido a essa modificação na medida de proximidade, a função objetivo explicitada para o algoritmo k-Modas pode ser reescrita como:

$$\begin{aligned} P(W, \mathcal{Q}) &= \sum_{l=1}^{k} \left(\xi \times \sum_{i=1}^{n} w_{i,l} \times \sum_{j=1}^{m} \delta(x_{q,j}, x_{r,j}) + \sum_{i=1}^{n} w_{i,l} \times \sum_{j=m+1}^{p} \varphi(x_{q,j}, x_{r,j}) \right) \\ &= \sum_{l=1}^{k} \left(P_l^{cat} + P_l^{cont} \right) \end{aligned} \tag{3.15}$$

em que

$$P_l^{cat} = \xi \times \sum_{i=1}^{n} w_{i,l} \times \sum_{j=1}^{m} \delta(x_{q,j}, x_{r,j})$$

$$P_l^{cont} = \sum_{i=1}^{n} w_{i,l} \times \sum_{j=m+1}^{p} \varphi(x_{q,j}, x_{r,j})$$

3.1.2.2.3 Algoritmo *k*-Protótipos .

O algoritmo k-Protótipos baseia-se no procedimento das k-Médias e se for aplicado a observações de variáveis contínuas, exclusivamente, é idêntico a esse último. Ele utiliza um método que atualiza de forma dinâmica os k protótipos, tentando maximizar a similaridade das observações em um mesmo grupo. Na função objetivo dada pela Equação (3.15), P_l^{cat} e P_l^{cont} são quantidades não-negativas e, portanto, minimizar a função é o mesmo que minimizar cada uma dessas parcelas. Para aquela referente à parte categórica utiliza-se o critério apresentado no k-Modas e para a outra parte, o que se usa nas k-Médias. A metodologia pode ser descrita desta forma:

1. selecionar k protótipos iniciais;
2. alocar cada observação ao protótipo mais próximo, levando em consideração a medida de proximidade combinada. Após cada realocação, atualizar os protótipos dos novos grupos;
3. testar novamente a pertinência das observações aos grupos formados, depois que todos os objetos estiverem alocados. Caso existam observações que precisam ser realocadas, fazê-lo e atualizar novamente os protótipos dos grupos formados;
4. repetir o passo anterior até que nenhuma observação possa ser realocada.

Em Huang (1998) dois bancos de dados reais foram usados para testar o desempenho dos métodos k-Modas e k-Protótipos. São estes: "doença do grão de soja" e "avaliação de crédito", respectivamente. Os resultados experimentais do artigo indicam que os algoritmos são acurados e eficientes na análise de uma grande e complexa massa de dados. O k-Modas é, entretanto, mais rápido que o k-Protótipos por necessitar de menos iterações para convergir. Uma desvantagem apontada pelo próprio autor é a escolha do peso ξ. Em Huang (1997) sugere-se que a média dos desvios padrões das variáveis contínuas pode ser usada para especificar ξ, mas isso não é uma regra geral.

3.1.3 *Fuzzy c*-Modas e *k*-Populações

3.1.3.1 *Fuzzy c*-Modas

O algoritmo *Fuzzy* c-Modas proposto em Huang e Ng (1999) é uma generalização do algoritmo k-Modas (HUANG, 1998), apresentado na seção 3.1.2, e tem como suporte o método *Fuzzy* c-Médias. São duas as principais contribuições dessa metodologia: os graus de pertinência das observações aos grupos (w), ao invés de assumirem os valores 0 ou 1, podem assumir qualquer valor no intervalo $[0, 1]$ e a quantidade chamada parâmetro *fuzzy* é incorporada à formulação.

No k-Modas as observações somente poderiam ser alocadas em um único grupo, àquele em que $w = 1$, e nessa nova metodologia o que cada observação possui é um peso ou um grau de pertinência em relação a todos grupos, podendo pertencer a mais um grupo ao mesmo tempo. Tanto no *Fuzzy* c-Modas quanto no *Fuzzy* c-Médias, a soma desses graus para uma dada observação X_i em relação aos k grupos, dada por $\sum_{l=1}^{k} w_{i,l}$ deve ser igual a 1. Os pesos w podem então ser interpretados como probabilidades e a sua análise contribui para uma melhor avaliação a respeito dos pontos do interior e da fronteira dos grupos formados. A idéia é que esses últimos pertençam ao grupo com um nível menor de atração do que pontos do interior desse.

O parâmetro *fuzzy* α determina o grau de confusão dos grupos finais. Isto é, ele regula o grau de sobreposição entre os grupos. A solução com maior grau de confusão está relacionado com α se aproximando do infinito (MINGOTI; LIMA, 2006). Na literatura é comum usar $\alpha = 2$ e foi esse o valor fixado para as simulações e também para as aplicações.

Dados o parâmetro *fuzzy* $\alpha \in (1, \infty)$ e a medida de proximidade definida na Equa-

ção (3.8), a função objetivo a ser minimizada passa então a ser escrita como:

$$\begin{aligned} P(W,\mathcal{Q}) &= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}^{\alpha} \times\ d(X_i, Q_l) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,l}^{\alpha} \times \delta(x_{i,j}, q_{l,j}) \\ \text{sujeito a} \quad & w_{i,l} \in [0,1]\ ,\ i \in [1,n]\ \text{ e } l \in [1,k] \\ & \sum_{l=1}^{k} w_{i,l} = 1\ ,\ i \in [1,n] \end{aligned} \tag{3.16}$$

Teorema 2 *A função P(W,Q) definida pela Equação (3.16) é minimizada se e somente se* $q_{l,j} = c_{t,j}$, $t \in\ DOM(A_j)$, *em que*

$$\sum_{i;x_{i,j}=c_{t,j}} w_{i,l}^{\alpha} \geq \sum_{i;x_{i,j}\neq\ c_{t,j}} w_{i,l}^{\alpha} \tag{3.17}$$

Pelo Teorema 2, enunciado e provado em Huang e Ng (1999), a moda do grupo l é constituído pelos valores $q_{l,j}$, $j \in [1,m]$, em que $q_{l,j} = c_{t,j}$ se a soma dos $w_{i,l}$ em relação às observações é máxima para a categoria t da variável j. Novamente, ele pode não ser único e, segundo a sugestão dos autores, a moda pode ser arbitrariamente definida por aquelas categorias que satisfaçam primeiro a desigualdade (3.17). No sentido de contornar esse problema Kim *et al.* (2004) e Kim *et al.* (2005) têm propostas que se assemelham bastante e que serão apresentadas na seção 3.1.3.2.

3.1.3.1.1 Algoritmo *Fuzzy c*-Modas .

Os passos do algoritmo *Fuzzy c*-Modas não estão descritos no artigo que apresenta o método. Mas, uma vez que ele se baseia no *Fuzzy c*-Médias, os passos desse algoritmo, apresentados em Lima (2001), podem ser adaptados e o método pode então ser devidamente formalizado.

As duas principais diferenças entre o *Fuzzy c*-Modas e o *Fuzzy c*-Médias são a obtenção da sementes dos grupos e a comparação entre as observações e essas sementes. No primeiro utiliza-se o Teorema 2 e o coeficiente de discordância simples (Equação (3.8) na página 19), respectivamente. Já no segundo, as sementes são as médias dos valores assumidos pelas variáveis e o coeficiente de dissimilaridade é, em geral, a distância euclidiana (Equação (3.2) na página 14).

O método *Fuzzy c*-Modas é dado pelos seguintes passos:

1. definir a quantidade de grupos k a serem obtidos e o parâmetro *fuzzy* α;
2. selecionar k sementes iniciais aleatoriamente ou determinando-se os pesos iniciais a partir de um gerador de números aleatórios. Suponhamos a segunda situação e consideremos a distribuição Uniforme[0,1] como geradora dos pesos iniciais;
3. padronizar os pesos iniciais para cada elemento amostral;
4. obter as sementes iniciais de acordo com o Teorema 2;
5. guardar as sementes;
6. comparar as observações às sementes de acordo com a medida de proximidade adotada;
7. recalcular os pesos;
8. alocar as observações aos grupos segundo algum critério, se for de interesse do pesquisador. Em geral, a observação será alocada ao grupo para o qual o seu grau de pertinência é máximo;
9. atualizar as sementes dos grupos de acordo com o Teorema 2;
10. repetir os passos de 5 a 9 até que algum critério de parada seja satisfeito.

Para recalcular os pesos (passo 7) será aplicada a expressão a seguir, apresentada em Bezdek (1973), a qual corresponde à solução (partição) ótima na minimização da função objetivo do método *Fuzzy* c-Médias (LIMA, 2001):

$$w_{i,l} = \frac{\left[d\left(X_i, Q_l\right)\right]^{-2/(\alpha-1)}}{\sum_{l=1}^{k}\left[d\left(X_i, Q_l\right)\right]^{-2/(\alpha-1)}} \tag{3.18}$$

3.1.3.2 *k*-Populações

A primeira abordagem (KIM *et al.*, 2004) introduz o conceito de *fuzzy* centróide e a segunda (KIM *et al.*, 2005) incorpora, entre outras modificações, principalmente nas restrições do *fuzzy* centróide, fatores de normalização na obtenção desses centróides e também na medida de proximidade entre observações e grupos de observações. Segundo os autores, a noção de população minimiza a incerteza e a imprecisão na representação das modas dos grupos e que o novo método, chamado k-Populações, tem menor tendência a encontrar um mínimo local como solução da função objetivo.

3.1.3.2.1 *Fuzzy* Centróide .

A moda do grupo l agora chamda *fuzzy* centróide vai ser dada por V_l, definido assim: $V_l = \{v_{l,1}, \ldots, v_{l,m}\}$ em que, $\forall\ j \in [1,m]$, $v_{l,j} = \{(c_{t,j}, \nu_{t,j}) \,|t \in\ DOM\,(A_j)\}$ sujeito a $0 \leq \nu_{t,j} \leq\ 1$ e também a $0 < \sum_{t=1}^{n_{A_j}} \nu_{t,j} <\ n$ (KIM *et al.*, 2005). A composição dos *fuzzy* centróides é determinada pelas categorias e pelos graus de contribuição dessas categorias para o grupo ($\nu_{t,j}$). Essas quantidades são obtidas a partir do grau de pertinência das observações ao grupos. Tem-se:

$$\nu_{t,j} \;=\; \frac{1}{\lambda_l} \times \sum_{i=1}^{n} \gamma\,(x_{i,j}) \tag{3.19}$$

em que

$$\lambda_l \;=\; \sqrt{\sum_{i=1}^{n} w_{i,l}^{2\alpha}} \tag{3.20}$$

$$\gamma\,(x_{i,j}) \;=\; \begin{cases} w_{i,l}^{\alpha}, & \text{se } x_{i,j} = c_{t,j} \\ 0, & \text{se } x_{i,j} \neq\ c_{t,j} \end{cases} \tag{3.21}$$

Resumidamente, para cada categoria t da variável j soma-se o grau de pertinência (w) da observação i ao grupo l se a categoria dessa variável nessa observação é t e pondera-se pelo fator de normalização (λ_l). Assim, $\nu_{t,j}$ descreve a distribuição das categorias da variável j com base nas observações que pertencem ao grupo l. O fator de normalização é o comprimento do vetor $w_l^{\alpha} = \left(w_{1,l}^{\alpha}, w_{2,l}^{\alpha}, \ldots, w_{n,l}^{\alpha}\right)'$.

O exemplo a seguir é bastante parecido com os apresentados nos artigos originais (Kim *et al.* (2004) e Kim *et al.* (2005)) em que a metodologia é apresentada.

Exemplo 5 *Imaginemos que em um dado grupo estão presentes três observações, caracterizadas por uma única variável que possui duas categorias* ($\{a, b\}$). *São estas:* $X = \{a, b, a\}$. *Sejam* $\alpha = 2$ *e* $w_{1,l} = 0,70$, $w_{2,l} = 0,80$ *e* $w_{3,l} = 0,15$.

O fuzzy centróide é então dado por:

$$V_l \;=\; \{(c_{1,1}; \nu_{1,1})\,; (c_{2,1}; \nu_{2,1})\}$$

em que

$$\nu_{1,1} = \frac{1}{\lambda_l} \times [\gamma(x_{1,1}) + \gamma(x_{2,1}) + \gamma(x_{3,1})]$$

Mas

$$\begin{aligned} \lambda_l &= \sqrt{w_{1,l}^4 + w_{2,l}^4 + w_{3,l}^4} = \sqrt{0,70^4 + 0,80^4 + 0,15^4} \approx 0,81 \\ \gamma(x_{1,1}) &= w_{1,l}^2 \ pois \ x_{1,1} = c_{1,1} \\ \gamma(x_{2,1}) &= 0 \ pois \ x_{2,1} \neq c_{1,1} \\ \gamma(x_{3,1}) &= w_{3,l}^2 \ pois \ x_{3,1} = c_{1,1} \end{aligned}$$

Assim,

$$\nu_{1,1} = \frac{1}{0,81} \times \left[w_{1,l}^2 + w_{3,1}^2\right] = \frac{1}{0,81} \times \left[\left(0,70^2\right) + \left(0,15^2\right)\right] \approx 0,63$$

Conseqüentemente,

$$\nu_{2,1} = \frac{1}{0,81} \times \left[w_{2,l}^2\right] = \frac{1}{0,81} \times \left(0,80^2\right) \approx 0,79$$

Finalmente,

$$V_l = \{(c_{1,1}; 0,63); (c_{2,1}; 0,79)\}$$

3.1.3.2.2 Medida de Proximidade .

Devido às modificações implementadas para a obtenção do *fuzzy* centróide, a medida de proximidade entre a observação X_i e o grupo l, caracterizado pelo *fuzzy* centróide V_l, passa a ser escrita como:

$$d(X_i, V_l) = \sum_{j=1}^{m} \delta(x_{i,j}, v_{l,t}) \tag{3.22}$$

em que

$$\delta\left(x_{i,j}, v_{l,t}\right) = \frac{1}{\eta_l} \times \sum_{t=1}^{n_{A_j}} \tau\left(x_{i,j}, c_{t,j}\right) \tag{3.23}$$

e

$$\eta_l = \sqrt{\sum_{t=1}^{n_{A_j}} \nu_{t,j}^2} \tag{3.24}$$

$$\tau\left(x_{i,j}, c_{t,j}\right) = \begin{cases} \nu_{t,j}, & \text{se } x_{i,j} \neq c_{t,j} \\ 0, & \text{se } x_{i,j} = c_{t,j} \end{cases} \tag{3.25}$$

em que η_l é o fator de normalização, que representa o comprimento do vetor $\nu_j = \left(\nu_{1,j}, \nu_{2,j}, \ldots, \nu_{n_{A_j},j}\right)'$.

3.1.3.2.3 Algoritmo *k*-Populações .

A formulação proposta faz com que $v_{l,t} \in V_l$ contenha a distribuição das categorias de A_j, ou seja, faz com que $v_{l,t}$ armazene cada categoria e a sua contribuição na formação do grupo. No artigo, Kim *et al.* (2005) sugere que o algoritmo seja inicializado atribuindo-se aleatoriamente valores para essas quantidades e, após a primeira iteração, para cada observação X_i e o grupo l obtém-se $\nu_{t,j}$ (Equação (3.19)) a partir de:

$$w_{i,l} = \left\{ \sum_{z=1}^{k} \left[\left(\frac{d\left(X_i, V_l\right)}{d\left(X_i, V_z\right)} \right)^{1/(\alpha-1)} \right] \right\}^{-1} \tag{3.26}$$

Não há modificações na função objetivo e o algoritmo possui os mesmos passos gerais: inicializar, alocar, atualizar, parar se há a impossibilidade de realocação das observações analisadas (KIM *et al.*, 2004):

1. determinar o número k de grupos e o parâmetro *fuzzy* α;
2. escolher *fuzzy* centróides iniciais de tal forma que para cada elemento de V_l são atribuídos graus de contribuição aleatórios;
3. calcular a matriz de proximidades entre as observações e os *fuzzy* centróides, de acordo com a Equação (3.22);
4. calcular o grau de pertinência entre as observações e os k grupos, de acordo com a Equação (3.26);

5. atualizar o *fuzzy* centróide a partir da Equação (3.19);

6. verificar se alguma observação pode ser realocada. Se sim, voltar ao passo 3. Se não, interromper o algoritmo.

A metodologia proposta dificulta os erros de classificação e evita que o algoritmo encontre um mínimo local pois mantém a informação a respeito da composição da moda do grupo de uma iteração para outra. Os autores (KIM *et al.*, 2005) compararam a metodologia proposta com três técnicas convencionais: alguma hierárquica não especificada no artigo, k-Modas e *Fuzzy* c-Modas. Eles obtiveram resultados bastante satisfatórios e concluíram pela superioridade do novo algoritmo. É importante ressaltar que a comparação é feita com base em alguns bancos de dados reais particulares, o que compromete a generalização dos resultados obtidos.

A situação proposta a seguir tenta exemplificar a metodologia apresentada.

Exemplo 6 *Suponhamos que se deseja agrupar três observações caracterizadas por uma única variável que possui duas categorias ($\{a, b\}$) em dois grupos. São estas: $X = \{a, b, a\}$. Seja $\alpha = 2$. Passo 1 (escolher fuzzy centróides iniciais):*

$$\begin{aligned} V_1 &= \{(c_{1,1}, \nu_{1,1}) ; (c_{2,1}, \nu_{2,1})\} = \{(a; 0,9) ; (b; 0,7)\} \\ V_2 &= \{(c_{1,1}, \nu_{1,1}) ; (c_{2,1}, \nu_{2,1})\} = \{(a; 0,6) ; (b; 0,8)\} \end{aligned}$$

Passo 2 (calcular graus de pertinência):
Como

$$\frac{1}{\alpha - 1} = \frac{1}{2-1} = \frac{1}{1} = 1$$

então o grau de pertinência da primeira observação ($i = 1$) ao primeiro grupo ($l = 1$) é:

$$w_{1,1} = \left[\frac{d(X_1, V_1)}{d(X_1, V_1)} + \frac{d(X_1, V_1)}{d(X_1, V_2)}\right]^{-1}$$

Mas

$$d(X_1, V_1) = \delta\left(x_{1,1}, v_{1,t}\right)$$

em que

$$\begin{aligned}
\delta\left(x_{1,1}, v_{1,t}\right) &= \frac{1}{\eta_1} \times \left[\tau\left(x_{1,1}, c_{1,1}\right) + \tau\left(x_{1,1}, c_{2,1}\right)\right] \\
\tau\left(x_{1,1}, c_{1,1}\right) &= 0 \ \textit{pois}\ x_{1,1} = c_{1,1} \\
\tau\left(x_{1,1}, c_{2,1}\right) &= \nu_{2,1} \ \textit{pois}\ x_{1,1} \neq c_{2,1} \\
\eta_1 &= \sqrt{\nu_{1,1}^2 + \nu_{2,1}^2} = \sqrt{0,90^2 + 0,70^2} \approx 1,14
\end{aligned}$$

Assim,

$$\begin{aligned}
\delta\left(x_{1,1}, v_{1,t}\right) &= \frac{1}{1,14} \times \left[0 + \nu_{2,1}\right] = \frac{1}{1,14} \times 0,70 \approx 0,61 \\
d(X_1, V_1) &= 0,61
\end{aligned}$$

Em relação ao segundo grupo ($l = 2$) tem-se:

$$\begin{aligned}
d(X_1, V_2) &= \frac{1}{\eta_2} \times 0,80 \\
\eta_2 &= \sqrt{0,60^2 + 0,80^2} = 1 \\
d(X_1, V_2) &= \frac{1}{1} \times 0,80 = 0,80
\end{aligned}$$

Finalmente,

$$\begin{aligned}
w_{1,1} &= \left[1 + \frac{0,61}{0,80}\right]^{-1} \approx 0,57 \\
w_{1,2} &= \left[1 + \frac{0,80}{0,61}\right]^{-1} \approx 0,43
\end{aligned}$$

De forma análoga obtêm-se os graus de pertinência para a terceira observação ($i = 3$):

$$\begin{aligned}
w_{3,1} &\approx 0,57 \\
w_{3,2} &\approx 0,43
\end{aligned}$$

O grau de pertinência da segunda observação ($i = 2$) ao primeiro grupo ($l = 1$) é:

$$w_{2,1} = \left[\frac{d(X_2, V_1)}{d(X_2, V_1)} + \frac{d(X_2, V_1)}{d(X_2, V_2)}\right]^{-1}$$

Mas

$$d(X_2, V_1) = \delta\left(x_{2,1}, v_{1,t}\right)$$

em que

$$\begin{aligned}
\delta\left(x_{2,1}, v_{1,t}\right) &= \frac{1}{\eta_1} \times \left[\tau\left(x_{2,1}, c_{1,1}\right) + \tau\left(x_{2,1}, c_{2,1}\right)\right] \\
\tau\left(x_{2,1}, c_{1,1}\right) &= \nu_{1,1} \ \textit{pois} \ x_{2,1} \neq c_{1,1} \\
\tau\left(x_{2,1}, c_{2,1}\right) &= 0 \ \textit{pois} \ x_{2,1} = c_{2,1} \\
\eta_1 &= \sqrt{\nu_{1,1}^2 + \nu_{2,1}^2} = \sqrt{0,90^2 + 0,70^2} \approx 1,14
\end{aligned}$$

Conseqüentemente,

$$\eta_2 = 1$$

Assim,

$$\delta\left(x_{2,1}, v_{1,t}\right) = \frac{1}{1,14} \times \left[\nu_{1,1} + 0\right] = \frac{1}{1,14} \times 0,90 \approx 0,79$$

Em relação ao segundo grupo ($l = 2$) tem-se:

$$\begin{aligned}
d(X_2, V_1) &= 0,79 \\
d(X_2, V_2) &= \frac{1}{1} \times 0,60 = 0,60
\end{aligned}$$

Finalmente,

$$\begin{aligned}
w_{2,1} &= \left[1 + \frac{0,79}{0,60}\right]^{-1} \approx 0,43 \\
w_{2,2} &= \left[1 + \frac{0,60}{0,79}\right]^{-1} \approx 0,57
\end{aligned}$$

Passo 3 (Atualizar os fuzzy centróides): Para o primeiro grupo ($l = 1$)

$$V_1 = \{(c_{1,1}, \nu_{1,1}); (c_{2,1}, \nu_{2,1})\}$$

em que

$$\nu_{1,1} = \frac{1}{\lambda_1} \times [\gamma(x_{1,1}) + \gamma(x_{2,1}) + \gamma(x_{3,1})]$$

Mas

$$\begin{aligned} \lambda_1 &= \sqrt{w_{1,1}^4 + w_{2,1}^4 + w_{3,1}^4} = \sqrt{0,57^4 + 0,43^4 + 0,57^4} \approx 0,50 \\ \gamma(x_{1,1}) &= w_{1,1}^2 \ \textit{pois}\ x_{1,1} = c_{1,1} \\ \gamma(x_{2,1}) &= 0 \ \textit{pois}\ x_{2,1} \neq c_{1,1} \\ \gamma(x_{3,1}) &= w_{3,1}^2 \ \textit{pois}\ x_{3,1} = c_{1,1} \end{aligned}$$

Assim,

$$\nu_{1,1} = \frac{1}{0,50} \times \left[w_{1,1}^2 + w_{3,1}^2\right] = \frac{1}{0,50} \times \left[2 \times \left(0,57^2\right)\right] \approx 1,30$$

Conseqüentemente,

$$\nu_{2,1} = \frac{1}{0,50} \times \left[w_{2,1}^2\right] = \frac{1}{0,50} \times \left(0,43^2\right) \approx 0,37$$

Analogamente para o segundo grupo ($l = 2$), obtêm-se:

$$\begin{aligned} \nu_{1,2} &= \frac{1}{0,42} \times \left[2 \times \left(0,43^2\right)\right] \approx 0,88 \\ \nu_{2,2} &= \frac{1}{0,42} \times \left[\left(0,57^2\right)\right] \approx 0,77 \end{aligned}$$

Pois

$$\lambda_2 = \sqrt{w_{1,2}^4 + w_{2,2}^4 + w_{3,2}^4} = \sqrt{0,43^4 + 0,57^4 + 0,43^4} \approx 0,42$$

Ao final do passo 2 observa-se que os graus de pertinência das observações que contêm a categoria *a* são maiores para o primeiro grupo, a que foi atribuído (aleatoriamente) maior

peso para essa categoria ($\nu_{1,1} = 0,9$ para $l = 1$). Por outro lado, o grau de pertinência da observação que contém a categoria b é maior para o segundo grupo ($\nu_{2,1} = 0,8$ para $l = 2$).

No passo 3 nota-se que o valor da contribuição da categoria a para o primeiro grupo é 1,30 ($\nu_{1,1} = 1,30$), mas na formulação do *fuzzy* centróide está especificado que essa quantidade deve estar contida no intervalo $[0, 1]$. É facilmente possível encontrar outros exemplos em que o mesmo ocorre. Assim, acredita-se que houve uma imprecisão por parte dos autores do artigo na descrição das restrições. No entanto, tal violação não é de grande importância pois não comprometeu a implementação dessa técnica.

3.1.4 *ROCK*

O *ROCK* é um procedimento para a análise de agrupamentos que incorpora uma nova concepção à medida de proximidade. Desenvolvido por Guha *et al.* (2000), emprega *links* ao invés de medidas de proximidade para agrupar os elementos em análise. Além desse novo elemento, o artigo de 2000 apresenta também o conceito de vizinho, a função objetivo, a medida de qualidade e os passos para a implementação da metodologia.

3.1.4.1 Vizinhança e *Links*

A medida de proximidade entre duas observações X_q e X_r é dada por uma medida de similaridade (s). O coeficiente de *Jaccard* é a medida usada no artigo original, porém mais detalhes da sua formulação serão apresentados na seção 3.1.4.4. Por enquanto vamos considerar a medida simplesmente como uma quantidade s. Assim, duas observações q e r serão consideradas vizinhas se $s\left(X_q, X_r\right) \geq \theta \in [0, 1]$, em que θ é um parâmetro a ser definido e que pode ser usado para controlar o quão próximo duas observações têm que estar para serem consideradas vizinhas.

O $link\left(X_q, X_r\right)$ é, então, o número de vizinhos comuns entre essas observações. Assim, quanto maior essa quantidade, mais provável é que X_q e X_r pertençam ao mesmo grupo. Essa concepção incorpora à metodologia um conhecimento maior a respeito da estrutura de relacionamento entre as observações, essa informação mais global faz com que *ROCK* seja um procedimento robusto.

O conceito de robustez está relacionado, portanto, com o fato do método *ROCK* utilizar *links* ao invés de uma medida de dissimilaridade para decidir se observações ou grupos devam ser unidos. Suponhamos que dois grupos não sejam tão bem separados e que

entre eles existam algumas observações que sejam bastante similares. Com o *ROCK*, esses conjuntos de objetos, mesmo estando próximos e sendo parecidos, não serão agrupados porque têm na verdade poucos vizinhos comuns e conseqüentemente um baixo número de *links*.

3.1.4.2 Função Objetivo e Medida de Qualidade

O objetivo é ainda manter em um mesmo grupo os elementos que são muito similares e em grupos distintos elementos não tão similares. Aplicando a definição de vizinhos e *links*, deseja-se então maximizar a soma de *links* entre observações que pertençam a um mesmo grupo e minimizar a soma de *links* entre observações de grupos diferentes. Assim, a função objetivo da Equação (3.27) deve ser maximizada para os k grupos:

$$P = \sum_{l=1}^{k} n_{G_l} \times \sum_{X_q, X_r \in\ G_l} \frac{link\,(X_q, X_r)}{n_{G_l}^{1+2f(\theta)}} \tag{3.27}$$

em que n_{G_l}é o número de observações no grupo G_l e $(1 - 2f\,(\theta))$ é o número esperando de *links* entre os pares de pontos em G_l. A função $f\,(\theta)$ pode não ser facilmente determinada, mas Guha *et al.* (2000) afirmam que se os grupos são razoavelmente bem definidos, qualquer estimativa para essa função, mesmo que viciada, pode funcionar bem na prática. Além disso, como todos os grupos são normalizados pelo denominador da Equação (3.27), todos eles são igualmente afetados e nem um nem outro é excessivamente penalizado. No artigo esses mesmos autores sugerem a função $f(\theta) = \frac{1-\theta}{1+\theta}$. Intuitivamente, percebe-se que essa é uma função interessante. Quando $\theta = 1$, ou seja, quando o único vizinho de uma determinada observação é ela própria, $f\,(\theta) = 0$ e o número esperado de *links* em G_l é $n_{G_l}^{1+2f(\theta)} = n_{G_l}$, em que n_{G_l} é o número de observações do grupo l. Quando $\theta = 0$, todas as observações de G_l são vizinhas umas das outras e agora como $f\,(\theta) = 1$, $n_{G_l}^{1+2f(\theta)} = n_{G_l}^3$.

Outra medida desenvolvida para a apresentação do *ROCK* é a medida de qualidade, que avalia se um dado par de grupos deve ou não ser unido. O seu valor máximo é dado pelo melhor par de grupos que devem ser agrupados naquele passo do algoritmo, ou seja, pelo par que possua o maior número de *links* cruzados (*cross links*). Esse número, dados dois grupos G_1 e G_2, é definido por: $link\,(G_1, G_2) = \sum_{X_q \in\ G_1, X_r \in\ G_2} link\,(X_q, X_r)$. A medida de qualidade é então dada por:

$$g\,(G_1, G_2) = \frac{link\,(G_1, G_2)}{(n_{G_1} + n_{G_2})^{1+2f(\theta)} - n_{G_1}^{1+2f(\theta)} - n_{G_2}^{1+2f(\theta)}}, \tag{3.28}$$

em que o denominador representa o número esperado de *links* cruzados entre os grupos

comparados. Essa ponderação faz-se necessária para que grupos com poucos *links* sejam mantidos separados.

Para o cálculo do número de *links* cruzados entre os grupos G_1 e G_2, cabe observar que se uma dada observação do grupo G_1, X_q por exemplo, tem como vizinha um outra observação do grupo G_2, X_r por exemplo, então todas as demais observações que são vizinhas de X_r passam, também, a serem vizinhas de X_q.

3.1.4.3 Algoritmo *ROCK*

O artigo que propõe o método *ROCK* apresenta detalhadamente os passos do algoritmo. Esses porém podem ser reestruturados de tal forma que a implementação computacional se torne mais viável:

1. determinar, se necessário, o número k de grupos a serem obtidos e fixar o parâmetro θ;
2. calcular a matriz de proximidades baseada em algum coeficiente de similidade;
3. obter a matriz de vizinhos A comparando a matriz de proximidades com o parâmetro θ fixado;
4. encontrar a matriz de *links* L de acordo com a sua definição. Ela pode ser facilmente encontrada fazendo-se $L = A'A$;
5. estabelecer a matriz que contém as medidas de qualidade para cada um dos pares de pontos;
6. unir as observações ou os grupos que possuem a maior entrada dessa última matriz;
7. recalcular a matriz de *links*;
8. repetir os três passos anteriores até que algum critério de parada seja satisfeito. Até que o número de grupos fixado k seja atingido ou até que a função objetivo seja nula.

Os autores afirmam a partir dos resultados obtidos e apresentados no artigo que a metodologia proposta não só possui resultados de melhor qualidade como também tem boas propriedades. Duas são as dificuldades na aplicação do *ROCK*: a primeira diz respeito à escolha do parâmetro θ, que é arbitrária e a outra, ao desempenho do algoritmo devido à sua complexidade e ao seu tempo de processamento.

Em Pantuzzo (2002), há entre outros pontos uma discussão sobre a escolha de θ. Nesse trabalho, sugere-se um procedimento empírico baseado nos valores que a medida de proximidade entre as observações pode assumir. Além desses valores, Pantuzzo testou nos estudos de caso selecionados outros pontos do domínio de θ. O desempenho do *ROCK* foi bastante satisfatório porém, notou-se uma forte dependência entre ele e a escolha do parâmetro θ. Os melhores resultados foram obtidos com θ em torno de $0,33$, o que corresponde, nos casos do trabalho dele, à metade dos atributos comuns possíveis entre duas observações quaisquer. Cabe observar que esse valor é inferior ao sugerido no artigo original do *ROCK*. Nesta dissertação fixou-se $\theta = 0,30$ para a realização das simulações. Nas aplicações, além desse valor, $\theta = 0,70$ também foi utilizado.

O exemplo a seguir, baseado no exemplo numérico apresentado no Anexo B de Pantuzzo (2002), contribui para a melhor compreensão da metodologia.

Exemplo 7 *Suponhamos que* 6 *observações, caracterizadas por* 5 *variáveis binárias, estejam organizadas na matriz* X *a seguir. Objetiva-se agrupar essas observações utilizando o método ROCK com* $\theta = 0,50$.

$$\mathbf{X} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

- *Medida de proximidade*

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 1 & 0,6666667 & 0,3333333 & 0,6666667 & 0 & 0,3333333 \\ 0,6666667 & 1 & 0,25 & 0,5 & 0,2 & 0,6666667 \\ 0,3333333 & 0,25 & 1 & 0,25 & 0,25 & 0,3333333 \\ 0,6666667 & 0,5 & 0,25 & 1 & 0,2 & 0,25 \\ 0 & 0,2 & 0,25 & 0,2 & 1 & 0,25 \\ 0,3333333 & 0,6666667 & 0,3333333 & 0,25 & 0,25 & 1 \end{pmatrix}$$

- *Matriz de vizinhos*

$$\mathbf{A} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- *Matriz de links*

$$\mathbf{L} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 & 3 & 0 & 3 & 0 & 1 \\ 3 & 0 & 0 & 3 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- *Medida de qualidade*

$$\mathbf{G} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 & 2,470291 & 0 & 2,470291 & 0 & 0,8234303 \\ 2,470291 & 0 & 0 & 2,470291 & 0 & 1,6468606 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 2,470291 & 2,470291 & 0 & 0 & 0 & 0,8234303 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0,8234303 & 1,6468606 & 0 & 0,8234303 & 0 & 0 \end{pmatrix}$$

- *Passo 1: Grupos unidos: 1 (observação 1) e 2 (observação 2)*

  . *Número de links*

$$\mathbf{L} = \begin{array}{c} \{1;2\} \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{pmatrix} 0 & 0 & 6 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 \end{pmatrix}$$

. *Medida de qualidade*

$$\mathbf{G} = \begin{matrix} \{1;2\} \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 0 & 2,791049 & 0 & 1,3955245 \\ 0 & 0 & 0 & 0 & 0 \\ 2,791049 & 0 & 0 & 0 & 0,8234303 \\ 0 & 0 & 0 & 0 & 0 \\ 1,3955245 & 0 & 0,8234303 & 0 & 0 \end{pmatrix}$$

- *Passo 2: Grupos unidos: 1 (observações 1 e 2) e 3 (observação 4)*

. *Número de links*

$$\mathbf{L} = \begin{matrix} \{1;2;4\} \\ 3 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

. *Medida de qualidade*

$$\mathbf{G} = \begin{matrix} \{1;2;4\} \\ 3 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 0 & 0 & 1,3475224 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1,3475224 & 0 & 0 & 0 \end{pmatrix}$$

- *Passo 3: Grupos unidos: 1 (observações 1, 2 e 4) e 4 (observação 6)*

. *Número de links*

$$\mathbf{L} = \begin{matrix} \{1;2;4;6\} \\ 3 \\ 5 \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

. *Medida de qualidade*

$$\mathbf{G} = \begin{matrix} \{1;2;4;6\} \\ 3 \\ 5 \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- *Passo 4: Função G é nula*

- *Resultado Final: 3 grupos foram formados:* $\{1, 2, 4, 6\}$, $\{3\}$ *e* $\{5\}$.

3.1.4.4 Medidas de Proximidade

A medida de similaridade proposta no artigo original (GUHA *et al.*, 2000), baseada no coeficiente de *Jaccard* (DUDA; HART, 1973), é dada por:

$$s\left(X_q, X_r\right) = \frac{|X_q \cap\ X_r|}{|X_q \cup\ X_r|} \tag{3.29}$$

Na Equação (3.29), o numerador representa as categorias coincidentes nas duas observações e o denominador, todos as categorias diferentes que ocorrem. A seguir apresenta-se um exemplo.

Exemplo 8 *Suponhamos 3 observações:* $X_1 = (a, c, e)$, $X_2 = (a, d, e)$ *e* $X_3 = (a, c, h)$, *sendo o domínios das 3 variáveis que as caracterizam dados pelos seguintes conjuntos:* $DOM\left(A_1\right) = \{a, b\}$, $DOM\left(A_2\right) = \{c, d\}$ *e* $DOM\left(A_3\right) = \{e, f, g, h\}$. *Usando a medida em (3.29) tem-se:*

$$s\left(X_1, X_2\right) = s\left(X_1, X_3\right) = \frac{2}{4} = 0,5$$

Observa-se que o denominador de (3.29) corresponde à quantidade de categorias diferentes que aparecem na comparação das observações. No primeiro caso tem-se $X_q \cup\ X_r = \{a, c, d, e\}$ *e no segundo,* $X_q \cup\ X_r = \{a, c, e, h\}$.

A idéia por trás da medida de similaridade proposta por Dutta *et al.* (2005) no artigo sobre o *QROCK* é uma ponderação com relação ao "tamanho" das variáveis que caracterizam as observações, ou seja, ao número de categorias presentes nos domínios dessas variáveis. Suponhamos que duas observações possuem categorias coincidentes em $(m-1)$ variáveis e que na *m-ésima*, cujo domínio possui duas categorias, essas observações sejam discordantes. Suponhamos novamente duas outras observações em que o mesmo ocorre, mas agora há discordância em outra variável, cujo domínio possui 50 categorias. A medida de similaridade ponderada definida a seguir trata diferentemente essas duas situações, fazendo com que o primeiro grupo de observações seja menos similar que o segundo grupo.

$$s\left(X_q, X_r\right) = \frac{|X_q \cap\ X_r|}{|X_q \cap\ X_r| + 2\sum_{j \notin\ X_q \cap\ X_r} \left(|DOM(A_j)|\right)^{-1}} \tag{3.30}$$

em que $|DOM(A_j)|$ é o número de categorias da variável j. Observa-se que a medida de similaridade previamente devida é um caso particular dessa, em que $|DOM(A_j)| = 1$ com $j \in [1, m]$.

Exemplo 9 *Usando as observações e variáveis do Exemplo 8 anterior, a similaridade passa a ser dada por:*

$$s(X_1, X_2) = \frac{2}{2 + 2 \times \frac{1}{2}} = \frac{2}{3} \approx 0,67$$
$$s(X_1, X_3) = \frac{2}{2 + 2 \times \frac{1}{4}} = \frac{4}{5} = 0,80$$

Observa-se que a similaridade entre as duas primeiras observações, que discordam na variável A_2, cujo domínio possui menos categorias, é menor que a similaridade entre as outras duas observações, que discordam na variável A_3, cujo domínio possui mais categorias. Ou seja, a nova medida de similaridade "penaliza" as observações cuja chance de serem concordantes é maior.

Ao contrário do que ocorre na análise de discriminante (MINGOTI, 2005; JOHNSON; WICHERN, 2002; TIMM, 2002; JOBSON, 1991-92), na análise de agrupamentos utilizando os métodos apresentados, não se faz necessário conhecer a distribuição de probabilidades das variáveis aleatórias medidas nos elementos amostrais. Assim, a análise de agrupamentos pode ser vista como um procedimento não-paramétrico.

3.2 Extensões do *ROCK*

Além da comparação de alguns métodos de agrupamentos para variáveis categóricas propostos na literatura, nesta dissertação pretende-se avaliar como algumas metodologias de comportam quando se tem também variáveis contínuas. Duas técnicas voltadas para essa abordagem já foram apresentadas na seção 3.1: Ligação Média e k-Protótipos. O *ROCK*, conforme originalmente apresentado na mesma seção, permite agrupar apenas observações de variáveis categóricas. Objetiva-se com a extensão desse método torná-lo aplicável a situações em que ocorram variáveis categóricas, contínuas ou ambas. A idéia é substituir a medida de similaridade s definida na seção 3.1.4 (página 34) pelas devidas quantidades aplicadas a cada contexto.

Assim, quando as observações são caracterizadas por variáveis contínuas, exclusivamente, a medida de proximidade adotada é a distância euclidiana (Equação (3.2)) transformada em similaridade (Equações (3.6) e (3.7)). E quando as categóricas também estão presentes, utiliza-se a medida combinada (Equações (3.4), (3.5) e (3.14)), que agrega a distância euclidiana transformada e, por exemplo, a medida de proximidade ponderada (Equação (3.30)). Essas últimas medidas de proximidade têm a mesma direção e estão definidas no mesmo intervalo de variação e, portanto, podem ser combinadas.

A partir de então é possível usar o parâmetro θ para definir se duas observações são ou não vizinhas e estabelecer o número de *links* entre elas. De forma geral, percebe-se que o algoritmo não sofre modificações substanciais em relação ao apresentado na seção 3.1.4. E isso se deve ao fato da matriz de proximidades só ser usada pelo *ROCK* nos passos iniciais do algoritmo. Depois de encontrada a matriz de vizinhos, os demais passos se baseiam somente na matriz de *links*, obtida a partir dessa última. Assim, a essência do método está preservada e isso é interessante porque as extensões incorporam as características do método original e podem ser também consideradas como metodologias robustas (seção 3.1.4.1).

3.3 Implementação Computacional

São apresentadas neste capítulo as modificações realizadas nas versões originais dos algoritmos. Elas foram necessárias para que a implementação computacional se tornasse mais viável e também para uma melhor comparação das metodologias.

3.3.1 Medida de Proximidade

O artigo do método das k-Modas sugere a utilização do coeficiente de discordância simples (Equação (3.8)) como medida de proximidade entre as observações de variáveis categóricas. Assim, para a sua generalização (*Fuzzy* c-Modas) e para a sua extensão (k-Protótipos), manteve-se a medida originalmente proposta. Também não foram feitas alterações no método k-Populações, que é uma modificação do *Fuzzy* c-Modas. Quando necessário, o coeficiente de discordância simples foi transformado utilizando-se as Equações (3.6) e (3.7).

Quanto aos algoritmos Ligação Média e *ROCK*, decidiu-se utilizar a medida de similaridade ponderada (Equação (3.30)), uma vez que a sua interpretação é mais interessante que a dada pelo coeficiente de *Jaccard* (exemplos 8 e 9). Com respeito às variáveis contínuas, utilizou-se a distância euclidiana (Equação (3.2)) transformada (Equações (3.6) e (3.7)).

3.3.2 Critérios de Parada e Número de Grupos

O Método da Ligação Média, como todos os procedimentos hierárquicos aglomerativos, tem como critério de parada o número de grupos desejado. Esse é também um dos critérios de parada do *ROCK*. O outro é a função objetivo nula ($g = 0$). Nesse caso verifica-se a nulidade das matrizes de vizinhos e de *links*. Ou seja, nenhuma observação pode ser considerada vizinha de outra e, dessa forma, nenhum outro agrupamento pode ser feito. Essa é uma solução interessante quando se estuda bancos de dados reais, pois isso demonstra que o *ROCK* não procede com o agrupamento se não há razão para fazê-lo (PANTUZZO, 2002). Já nos estudos de simulação, quando isso ocorre, é muito provável que o número de grupos pré-estabelecido na situação simulada não tenha sido alcançado. E essa solução deixa de ser interessante uma que vez que não será possível compará-la com aquelas dos demais métodos analisados.

As metodologias não-hierárquicas têm originalmente como critério de parada a im-

possibilidade de realocação. Além desse critério foram implementados outros dois para os métodos k-Modas, k-Protótipos e *Fuzzy* c-Modas: a comparação entre as sementes dos grupos e o número máximo de iterações. A idéia é que o algoritmo seja interrompido caso não haja, de um passo para outro, mudanças na alocação das observações aos grupos, ou que essas não alterem a estrutura dos grupos (comparação entre sementes), ou caso o número máximo de iterações seja atingido.

Fixou-se o número máximo de iterações em 1.000 e a comparação entre as sementes do passo anterior e do atual foi feita com base no coeficiente de concordância simples, para as variáveis categóricas, e também com base na distância euclidiana, transformada em similaridade, para os dois tipos de variáveis. Caso essas medidas fossem simultaneamente maiores que $0,80$, considerou-se que não houve mudanças entre as sementes comparadas.

Para o k-Populações, devido à sua formulação, não foi possível implementar a comparação entre as modas. Os critérios de interrupção desse algoritmo foram somente a realocação e o número máximo de iterações.

Conforme já é sabido, o número de grupos precisa ser previamente fixado nos métodos não-hierárquicos. Essa quantidade definirá quantas sementes iniciais serão necessárias para inicializar o algoritmo. Apesar dessa prévia definição, é possível que esses métodos convirjam para soluções em que o número de grupos não é o fixado. As simulações nas quais isso ocorreu foram descartadas, mas foi feita uma contabilização para cada um dos métodos analisados. O mesmo foi feito com as simulações em que o critério de parada do *ROCK* foi da função objetivo nula.

Cabe aqui destacar que as duas próximas modificações comentadas neste capítulo foram desenvolvidas devido ao alto número de descartes obtidos nas primeiras simulações feitas. A primeira diz respeito à inicialização do algoritmo k-Populações e a segundo à alocação aleatória nos métodos *Fuzzy*.

3.3.3 Inicialização

Para uniformizar as técnicas não-hierárquicas, optou-se pela inicialização aleatória das sementes. Para isso, retirou-se uma amostra sem reposição de observações, do tamanho do número de grupos fixado, que foram então usadas como sementes iniciais dos métodos. Assim, ao invés de (ver seção 3.1.3.2.3):

1. determinar o número k de grupos e o parâmetro *fuzzy* α;

2. escolher *fuzzy* centróides iniciais de tal forma que para cada elemento de V_l são atribuídos graus de contribuição aleatórios;

3. calcular a matriz de proximidades entre as observações e os *fuzzy* centróides, de acordo com a Equação (3.22);

4. calcular o grau de pertinência entre as observações e os k grupos, de acordo com a Equação (3.26);

5. atualizar o *fuzzy* centróide a partir da Equação (3.19);

6. verificar se alguma observação pode ser realocada. Se sim, voltar ao passo 3. Se não, interromper o algoritmo.

o que se fez foi:

1. determinar o número k de grupos e o parâmetro *fuzzy* α;

2. escolher k observações aleatoriamente e usá-las como sementes iniciais;

3. calcular a matriz de proximidades entre as observações e as sementes iniciais, com base no coeficiente de discordância simples (Equação (3.8));

4. calcular o grau de pertinência entre as observações e os k grupos, de acordo com a Equação (3.26);

5. atualizar o *fuzzy* centróide a partir da Equação (3.19);

6. calcular a matriz de proximidades entre as observações e os *fuzzy* centróides, de acordo com a Equação (3.22);

7. repetir os passos de 4 a 6 até que algum critério de parada seja satisfeito.

O algoritmo *Fuzzy c*-Modas (seção 3.1.3.1.1) também pode ser reescrito, com pequenas modificações, considerando-se que as sementes iniciais são selecionadas aleatoriamente:

1. definir a quantidade de grupos k a serem obtidos e o parâmetro *fuzzy* α;

2. selecionar k sementes iniciais aleatoriamente e defini-las como as modas iniciais do algoritmo;

3. guardar as sementes;

4. comparar as observações às sementes de acordo com a medida de proximidade adotada;

5. recalcular os pesos;

6. alocar as observações aos grupos segundo algum critério, se for de interesse do pesquisador. Em geral, a observação será alocada ao grupo para o qual o seu grau de pertinência é máximo;

7. padronizar os pesos para cada grupo, objetivando satisfazer $\sum_{l=1}^{k} w_{i,l} = 1$, com $i \in [0, n]$;

8. atualizar as sementes dos grupos de acordo com o Teorema 2;

9. repetir os passos de 3 a 8 até que algum critério de parada seja satisfeito.

Resumidamente, os passos 2 e 3 da seção 3.1.3.1.1 são desconsiderados e antes da atualização das sementes (passo 9 da mesma seção), é necessário padronizar os pesos para cada grupo.

3.3.4 Alocação

No estudo de bancos de dados reais tem-se como resultado dos métodos *Fuzzy c*-Modas e k-Populações os graus de pertinência das observações aos grupos. A análise de observações que pertençam a mais de um grupo é um dos ganhos da aplicação dessas técnicas. Nesse caso, a alocação não é obrigatória. Já nos estudos de simulação, faz-se necessário decidir a qual grupo a observação pertence. Em geral, destina-se a observação ao grupo cujo grau de pertinência é máximo. Se essa situação ocorre em dois ou mais grupos, ou seja, se uma observação possui o mesmo grau de pertinência para dois ou mais grupos, é costume alocá-la arbitrariamente ao "primeiro" grupo (HUANG; NG, 1999).

Na modificação sobre a alocação implementada nesta dissertação, ao invés da observação ser classificada sempre no "primeiro" grupo, caso haja empates entre os graus de pertinência, escolheu-se aleatoriamente entre os possíveis destinos aquele ao qual a observação será alocada.

3.3.5 Modificações no *k*-Modas e no *k*-Protótipos

Existem diversos tipos de algoritmo das k-Médias e eles diferem na seleção das sementes iniciais, nos cálculos de dissimilaridade e nas estratégias para se calcular as médias dos

grupos (ANDERBERG, 1973; HUANG, 1998). Segundo Sharma (1996), Timm (2002) e Mingoti (2005), os passos tradicionalmente seguidos são estes:

1. definir a quantidade de grupos k a serem obtidos e selecionar k sementes iniciais;
2. alocar as observações aos k grupos de acordo com a medida de proximidade adotada (em geral utiliza-se a distância euclidiana);
3. atualizar as médias dos novos grupos;
4. repetir os passos 2 e 3 até que não seja possível fazer nenhuma outra alocação.

Se fôssemos aplicar diretamente as modificações propostas na elaboração do k-Modas e do k-Protótipos nesse algoritmo, os resultados finais seriam um pouco diferentes das sequências de passos apresentadas nas seções 3.1.2.1.3 e 3.1.3.1.1. Mas uma vez que a utilização desse algoritmo é mais usual e que o seu desenvolvimento computacional é mais fácil, optou-se por fazer pequenas alterações nos métodos originais, mantendo-se, é claro, as inovações dos autores dos artigos. Os passos dos algoritmos modificados podem ser assim descritos:

1. definir a quantidade de grupos k a serem obtidos e selecionar k sementes iniciais;
2. alocar as observações aos k grupos, de acordo com a medida de proximidade adotada (coeficiente de discordância simples ou medida de proximidade combinada);
3. atualizar as modas ou os protótipos dos grupos formados, depois que todos os objetos tiverem sido alocados;
4. comparar novamente as observações às sementes atualizadas. Caso existam observações que estejam mais próximas a outras sementes que não àquelas do próprio grupo, realocá-las e, então, atualizar novamente as modas ou os protótipos dos grupos formados;
5. repetir o passo anterior até que algum critério de parada seja satisfeito.

Contribuiu também para que as alterações acima fossem feitas o fato do algoritmo *Fuzzy* c-Modas (seção 3.1.3.1.1), ter como suporte o *Fuzzy* c-Médias (LIMA, 2001). Esse último, por sua vez, está baseado na sequência mais usual do método das k-Médias, apresentado acima. Assim, tendo todas as metodologias a mesma base, a comparabilidade dos resultados obtidos a partir da sua aplicação refletem somente as diferenças existentes entre os métodos.

3.3.6 Comentários sobre o *ROCK*

Sabe-se que o parâmetro θ do *ROCK* (página 34) determina não somente se duas observações ou grupos são vizinhos, mas também a qualidade da solução obtida por esse método. Supondo-se que o valor do θ "ótimo" seja conhecido, aumentado-se o seu valor a partir deste ponto observa-se que o *ROCK* converge primeiro para a solução em que todas as observações são agrupadas em um grupo "maior". Aumentado-se ainda mais o seu valor, a convergência se dará mais rapidamente pelo critério da medida de qualidade nula. Isso porque somente grupos ou observações que tenham alta similaridade serão considerados vizinhos.

Isso pôde ser observado nas primeiras simulações feitas. Nesses testes foi possível notar também que o valor ideal de θ para a combinação usando pesos e para a proposta de *Gower* não é o mesmo. Uma conclusão preliminar é que quanto menor o valor desse parâmetro para o segundo caso (*Gower*) em relação ao primeiro (Pesos), melhor foram os resultados finais. Análises mais aprofundadas precisam ser feitas para que conclusões mais sólidas a esse respeito possam ser obtidas.

Outro ponto do método *ROCK* que não será discutido nesta dissertação é a determinação da função $f(\theta)$, presente no cálculo da função-objetivo e na medida de qualidade do algoritmo. Optou-se por usar a sugestão dos autores $f(\theta) = \frac{1-\theta}{1+\theta}$, uma vez que a sua interpretação intuitiva faz sentido. O estudo de outras funções fazem parte dos futuros trabalhos a serem desenvolvidos.

4 *Estudos de Simulação e de Aplicação*

4.1 Estudos de Simulação

Foram realizados dois estudos de simulação. O primeiro diz respeito à análise de agrupamentos de observações de variáveis categóricas e o segundo, de observações caracterizadas pelos dois tipos de variáveis. As seções desse capítulo contêm as situações simuladas, algumas definições, a geração aleatória das observações e os resultados obtidos.

Nos dois estudos e em cada uma das situações foram feitas 1.000 simulações de Monte Carlo de tal forma que os resultados obtidos retratam a média de todas elas. As simulações correspondem ao número de bancos de dados gerados e, considerando a simulação como um todo, 490.000 bancos foram criados, sendo 130.000 da primeira parte, e 360.000 da segunda. Vale a pena explicitar que após gerado, um mesmo banco foi usado por todos os métodos de agrupamentos. Além disso, depois de escolhidas as sementes iniciais, as mesmas são usadas nas metodologias que delas precisam.

Cabe observar que todos os métodos, incluindo pequenos procedimentos, programas para a geração de dados categóricos e para simulação tiveram que ser implementados dado que atualmente quase nada, comparando-se à quantidade de técnicas existentes, está disponível nos *softwares* estatísticos. Os programas para a geração de dados contínuos foram obtidos após adaptação dos códigos usados por Lima (2001). A confiabilidade dos programas foi testada em um estudo de consistência. Nele foi possível verificar se os métodos alocavam corretamente observações em grupos completamente distintos. Para isso, dentre as situações propostas, selecionaram-se duas (a serem identificadas nas seções 4.1.3 e 4.1.4), e novamente procedeu-se 1.000 simulações. Assim, mais 2.000 bases foram simuladas.

4.1.1 Tempo de Processamento

Um único modelo foi selecionado para que se pudesse avaliar o tempo de processamento dos algoritmos. O recurso computacional usado nessa etapa foi um computador doméstico cuja placa mãe e processador são, respectivamente: ASUS modelo A8V e Athlon 64 3200+ (OS 32 Bits). A máquina dispõe de 1Gb de Memória DDR (Dual) e aproximadamente 95Gb no disco rígido. Todos os aplicativos foram finalizados antes que o teste fosse começado e os resultados representam o tempo médio em 10 simulações. As situações utilizadas também serão identificados nas seções correspondentes.

Nota-se que o método computacionalmente mais demorado é o *ROCK*. As metodologias não-hierárquicas possuem tempos de processamento inferiores, sendo o *Fuzzy c*-Modas o menos demorado entre as técnicas para variáveis categóricas (somente) e o k-Protótipos entre os métodos para ambos os tipos de variáveis.

Tabela 4.1: Tempo de processamento (em segundos) dos métodos estudados

Método	Tempo (em segundos)
Variáveis categóricas (somente)	
Ligação Média	5,3
ROCK	41,9
k-Modas	0,5
Fuzzy c-Modas	0,1
k-Populações	0,9
Variáveis categóricas e contínuas	
Ligação Média	5,9
ROCK	42,5
k-Protótipos	0,8

4.1.2 Métodos de Avaliação

Para avaliar o isolamento externo dos grupos finais obtidos, utiliza-se a "Taxa de Alocação Correta", que é a proporção de observações alocadas corretamente aos grupos originalmente simulados. O valor um (ou 100%) é o valor de referência dessa medida, bastante usada na literatura.

Para estudar a coesão interna no agrupamento de variáveis contínuas, é comum usar a "Taxa de Dispersão Interna" ou *Intraclass Dispersion Rate* (ICDR) (RENCHER, 1995).

Essa taxa é a razão entre a soma de quadrados dentro dos grupos e a soma de quadrados total. Em outras palavras, $ICDR = \frac{SQD}{SQT} = (1 - R^2)$, em que R^2 é o coeficiente de determinação do modelo. Quanto maior o valor de R^2, menor é a dispersão interna dos grupos e, portanto, maior é a coesão dentro deles.

Para as variáveis categóricas ainda não foi desenvolvido um método para avaliar a coesão interna dos grupos formados. Assim, somente a "Taxa de Alocação Correta" será usada na comparação dos desempenhos dos métodos. Vale lembrar que essa avaliação é feita com base nas 1.000 simulações efetuadas.

4.1.3 Variáveis Categóricas

4.1.3.1 Situações Simuladas

No total, 65 situações foram construídas para simulação. Tais situações podem ser classificadas em nove casos gerais, com uma característica de extrema importância: a separação ou a intersecção dos grupos. Os casos de 1 a 3 dizem respeito ao grupos separados, os casos de 4 a 7 dizem respeito aos grupos sobrepostos e os casos 8 e 9 tentam representar situações "reais".

Além dessa característica, as situações diferem umas das outras no número de grupos, que varia entre dois e cinco, de variáveis, que varia entre duas e quatro, e de categorias para cada variável, que varia de duas a 10. Há situações em que a quantidade de categorias é igual para todas, para parte ou para nenhuma variável. Tantas situações foram delineadas tentando-se expor os algoritmos aos mais diversos tipos de dados, permitindo dessa forma a generalização dos resultados e uma análise mais detalhada do desempenho dos métodos.

A Tabela 4.2 descreve as 65 situações simuladas. No caso 1, por exemplo, as situações 2 e 3 têm o mesmo número de grupos e de variáveis que a situação 1. O mesmo ocorre entre os casos 13 a 16. Preferiu-se não preencher a tabela complemente para facilitar a sua leitura.

Tabela 4.2: Número de grupos (k), de variáveis categóricas (m) e de categorias dessas variáveis ($|DOM(A_j)|$, com $j \in [1, m]$) das situações estudadas

Caso	Situação	k	m	$\|DOM(A_1)\|$	$\|DOM(A_2)\|$	$\|DOM(A_3)\|$	$\|DOM(A_4)\|$
Grupos separados na 1^a variável							
1	1	2	2	2	2		
	2			2	5		
	3			2	10		
	4	3	2	3	3		
	5			3	5		
	6			3	10		
	7	5	2	5	5		
	8			5	10		
2	9	2	4	2	2	2	2
	10			2	2	3	3
	11			2	2	3	4
	12			2	5	5	5
	13	3	4	3	3	3	3
	14			3	2	3	3
	15			3	2	5	10
	16			3	5	5	5
	17	5	4	5	5	5	5
	18			5	2	4	4
	19			5	3	4	10
	20			5	2	2	2
	21			5	8	8	8
Grupos separados nas duas 1^{as} variáveis							
3	22	2	4	2	2	2	2
	23			2	2	3	3
	24			2	2	5	10
	25			2	2	5	5
	26	3	4	3	3	3	3
	27			3	3	2	2
	28			3	3	5	10
	29	5	4	5	5	5	5

continua na próxima página

Tabela 4.2 – continuação da página anterior

Caso	Situação	k	m	$\|DOM(A_1)\|$	$\|DOM(A_2)\|$	$\|DOM(A_3)\|$	$\|DOM(A_4)\|$
	30			5	5	3	3
	31			5	5	5	10
	32			5	5	2	2
Grupos sobrepostos na 1^a variável							
4	33	2	2	2	2		
	34			2	5		
	35			2	10		
	36	3	2	3	3		
	37			3	5		
	38			3	10		
	39	5	2	5	5		
	40			5	10		
5	41	2	4	2	2	3	3
	42			2	2	5	10
	43			2	5	5	5
	44	3	4	3	2	3	3
	45			3	2	5	10
	46			3	5	5	5
	47	5	4	5	2	3	3
	48			5	2	5	10
	49			5	5	5	5
Grupos sobrepostos nas duas 1^{as} variáveis							
6	50	2	4	2	2	2	2
	51			2	2	3	3
	52			2	2	5	10
	53			2	2	5	5
	54	3	4	3	3	3	3
	55			3	3	2	2
	56			3	3	5	10
	57	5	4	5	5	5	5
	58			5	5	3	3
	59			5	5	5	10

continua na próxima página

Tabela 4.2 – continuação da página anterior

Caso	Situação	k	m	$\|DOM(A_1)\|$	$\|DOM(A_2)\|$	$\|DOM(A_3)\|$	$\|DOM(A_4)\|$
	60			5	5	2	2
Grupos sobrepostos nas três 1^{as} variáveis							
7	61	5	4	5	2	3	3
	62			5	2	5	10
	63			5	5	5	5
Não há controle sobre a separação e sobreposição dos grupos							
8	64	[2,8]	[2,15]	[2,10]	[2,10]	[2,10]	[2,10]
9	65	2	15	2	...		2

A separação entre os grupos foi construída de forma que todas as observações de um mesmo grupo tinham a mesma categoria para a primeira (casos 1 e 2) ou para as duas primeiras variáveis (caso 3). As categorias das demais variáveis, se essas existissem, foram escolhidas de forma aleatória. Por exemplo, na situação 1 (caso 1), em que há dois grupos, (suponhamos $\{G_1, G_2\}$) e duas variáveis, cada uma com duas categorias (suponhamos $\{a, b\}$), a todas as observações do grupo G_1 foi atribuída a categoria $\{a\}$ e a todas as observações do grupo G_2, a categoria $\{b\}$ para a primeira variável. Para a segunda variável, as categorias foram atribuídas aleatoriamente. Nessa situação e nas demais do caso 1 os grupos são separados na primeira variável.

Analogamente, para a situação 32 (caso 3), em que há cinco grupos e quatro variáveis, as duas primeiras com quatro categorias (suponhamos $\{a, b, c, d\}$) e as outras duas com duas categorias (suponhamos $\{e, f\}$), a todas as observações do grupo G_1 foram atribuídas as categorias $\{a, a\}$, a todas do grupo G_2, as categorias $\{b, b\}$ e, sucessivamente, a todas do grupo G_5, as categorias $\{e, e\}$ para as duas primeiras variáveis. Para as duas outras variáveis, as categorias foram atribuídas aleatoriamente. Em todas as situações do caso 3, nessa inclusive, os grupos são separados nas duas primeiras variáveis.

Já a sobreposição entre os grupos foi construída de forma que as observações de um mesmo grupo tinham proporcionalmente todas as categorias possíveis da primeira (casos 4 e 5) ou das duas primeiras (caso 6) ou das três primeiras variáveis (caso 7). A proporcionalidade se dava de acordo com o número de categorias. Por exemplo, na situação 33 (caso 4) que possui as mesmas especificações da situação 1 comentada anteriormente, ou seja, dois grupos e duas variáveis, cada uma com duas categorias. À metade das observações do grupo G_1 foi atribuída a categoria $\{a\}$ e às demais, a categoria $\{b\}$. O mesmo foi feito para as observações do grupo G_2. Novamente as categorias da outra variável foram

escolhidas aleatoriamente. Nas demais situações do caso 4, assim como nessa, os grupos são sobrepostos na primeira variável. Como essa contém apenas duas categorias, as observações dos grupos foram divididas ao meio para que uma parte recebesse a primeira e a outra parte, a segunda categoria.

A idéia é a mesma para situação 63 (caso 7), em que há cinco grupos, quatro variáveis e cada variável com cinco categorias. Agora as observações dos grupos precisam ser divididas em cinco subgrupos, cada um com 20% das observações. Ao primeiro subconjunto de observações foi atribuída a categoria $\{a\}$, às observações subsequentes, a categoria $\{b\}$ e, sucessivamente, ao último subconjunto foi atribuída a categoria $\{e\}$. Isso para os cinco grupos e para as três primeiras variáveis. Assim, todas as categorias estão presentes em cada um dos grupos e essa ocorrência é proporcional ao número de categorias. Nesse último exemplo, como são cinco categorias a proporção é 20% das observações.

Se o número de categorias não for o mesmo, a proporcionalidade em relação a esse número ainda se preserva. Por exemplo, na situação 61, em que a primeira variável possui cinco, a segunda, duas e a terceira, três categorias, a atribuição se deu da seguinte forma: suponhamos que sejam apenas dois grupos, cada um com 10 observações, e as mesmas três variáveis da situação 61. Assim, a matriz de dados será:

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} \textit{Grupo} & A_1 & A_2 & A_3 \end{array} \\ \left(\begin{array}{cccc} G_1 & a & f & h \\ G_1 & a & f & h \\ G_1 & b & f & h \\ G_1 & b & f & h \\ G_1 & c & f & i \\ G_1 & c & g & i \\ G_1 & d & g & i \\ G_1 & d & g & j \\ G_1 & e & g & j \\ G_1 & e & g & j \\ G_2 & a & f & h \\ G_2 & a & f & h \\ G_2 & b & f & h \\ G_2 & b & f & h \\ G_2 & c & f & i \\ G_2 & c & g & i \\ G_2 & d & g & i \\ G_2 & d & g & j \\ G_2 & e & g & j \\ G_2 & e & g & j \end{array} \right) \end{array}$$

No exemplo, a "primeira" categoria ($\{h\}$) da variável A_3 foi atribuída a um número maior de observações (quatro das 10, o que corresponde a 40%). Nos estudos de simulação realizados o que ocorreu foi o inverso. Isso porque fixou-se em 50 e não em 10 o número de observações para cada grupo e, assim sendo, as proporções foram $0,32$, $0,33$ e $0,33$, o que corresponde a 16, 17 e 17 observações, após o arredondamento. É importante notar que é sempre a "primeira" categoria que é a menos freqüente e portanto os grupos não são diferentes nisso.

Ressalta-se também que, no término da geração dos dados, um número aleatório foi gerado para cada observação e em seguida o banco foi organizado em ordem crescente segundo essa quantidade. O objetivo disso era desorganizar os dados e não tê-los na forma como a matriz apresenta, por grupo e por categoria. Os bancos de dados são assim mais "reais" pois em termos práticos as observações estão dispostas sem critério algum. É importante ressaltar que uma vez desordenadas, as observações são agrupadas por todos os métodos em questão, sem nenhuma outra alteração.

Nas situações 61 e 63 e nas demais do caso 7 os grupos são sobrepostos nas três primeiras observações. Esse é o pior caso, pois há maior confusão.

No caso 8 (situação 64) o número de grupos, de variáveis e de categorias foram escolhidos aleatoriamente nos intervalos $[2, 8]$, $[2, 15]$ e $[2, 10]$ respectivamente. Aleatória era também a atribuição das categorias das variáveis. Não houve, portanto, um controle da separação e da sobreposição dos grupos. Além desse caso, o 9 (situação 65) também tenta representar uma situação "real". Nele há dois grupos e 15 variáveis, cada uma com duas categorias. Novamente a atribuição das categorias foi feita de forma aleatória.

Os casos analisados não esgotam todas as possíveis situações reais. Porém, com base em um esquema de análise estudou-se o efeito do aumento do número de grupos, de variáveis e de categorias segundo a disposição do grupos (separados ou sobrepostos). Estudou-se também o grau de separação e sobreposição dos grupos, mantendo-se o número de variáveis constante. A partir da estratégia disposta a seguir, em que se levou em consideração as estruturas das bases de dados, foi possível se ter uma noção do desempenho dos algoritmos nos tantos casos propostos.

- **Estratégia para análise de desempenho**

  1. Grau de separação e sobreposição, mantendo-se o número de variáveis constante
     - grupos separados: situações $\{9 - 21\}$ *versus* situações $\{22 - 32\}$
     - grupos sobrepostos: situações $\{41 - 49\}$ *versus* situações $\{50 - 60\}$ *versus* situações $\{61 - 63\}$
  2. Número de grupos
     - grupos separados: situações $\{1 - 3, 9 - 12, 22 - 25\}$ *versus* situações $\{4 - 6, 13 - 16, 26 - 28\}$ *versus* situações $\{7 - 8, 17 - 21, 29 - 32\}$
     - grupos sobrepostos: situações $\{33 - 35, 41 - 43, 50 - 53\}$ *versus* situações $\{36 - 38, 44 - 46, 54 - 56\}$ *versus* situações $\{39 - 40, 47 - 49, 57 - 63\}$
  3. Número de variáveis
     - grupos separados: situações $\{1 - 8\}$ *versus* situações de $\{9 - 21\}$
     - grupos sobrepostos: situações $\{33 - 40\}$ *versus* situações $\{41 - 49\}$
  4. Número de categorias: análise de cada um dos nove casos

4.1.3.2 Geração Aleatória das Observações Categóricas

A forma de geração das variáveis que determinam a separação e a sobreposição dos grupos já foi detalhada na seção anterior. Para finalizar a formação do banco de dados, basta definir como que, aleatoriamente, são geradas as demais variáveis, quando existirem. Dois modelos foram implementados, ambos com a mesma estrutura. Para o grupo k e para a variável j, com $j \in [1, m]$, seja t uma possível categoria de A_j ($t \in DOM(A_j)$):

1. determinar a probabilidade de ocorrência de cada uma das suas categorias: $p_{c_{t,j}}$;
2. normalizar, se for preciso, e obter as novas probabilidades $p^*_{c_{t,j}}$. Assim, $P[A_j = t] = p^*_{c_{t,j}}$, $\forall t \in DOMA_j$ e $\sum_t p^*_{c_{t,j}} = 1$;
3. gerar aleatoriamente as observações da distribuição $p^*_{c_{t,j}}$.

Os termos "modelo Beta" e "modelo Uniforme" serão usados daqui por diante para designar as duas formas de geração de observações das variáveis que não determinam a separação ou a sobreposição dos grupos. Tais modelos diferem apenas na determinação das probabilidades de ocorrência das categorias (passo 1). No modelo Beta, as probabilidades de ocorrência foram geradas aleatoriamente de acordo com uma distribuição Beta com parâmetros especificados (α e β). No modelo Uniforme, a probabilidade de ocorrência era inversamente proporcional ao número de categorias $\left(p_{c_{t,j}} = \frac{1}{|DOM(A_j)|}\right)$. A normalização (passo 2) foi aplicada somente ao primeiro modelo. No segundo, bastou fazer $p^*_{c_{t,j}} = p_{c_{t,j}}$.

Para gerar as ocorrências da variável j a partir da distribuição final (passo 3), utilizou-se o Método da Transformação Inversa para Variáveis Discretas, que consiste em:

1. a partir das probabilidades $p^*_{c_{t,j}}$, encontrar a distribuição acumulada;
2. gerar um número aleatório U a partir da distribuição Uniforme $(0, 1)$;
3. comparar o valor obtido com a distribuição acumulada e atribuir a categoria adequada:

$$A_j = \begin{cases} c_{1,j}, & \text{se } 0 < U < p^*_{c_{1,j}} \\ c_{2,j}, & \text{se } p^*_{c_{1,j}} < U < p^*_{c_{1,j}} + p^*_{c_{2,j}} \\ \vdots & \\ c_{t,j}, & \text{se } p^*_{c_{t-1,j}} < U < 1 \end{cases} \tag{4.1}$$

Um breve estudo sobre a distribuição Beta foi feito para auxiliar na determinação dos parâmetros a serem usados na simulação. Como o objetivo era ter probabilidades de ocorrência maiores e menores, de tal forma que o modelo Beta fosse bastante diferente do Uniforme, arbitrariamente os valores $\alpha = 1$ e $\beta = 0,1$ foram escolhidos. A Figura 4.1 ilustra a distribuição Beta, dependendo dos seus parâmetros (α,β).

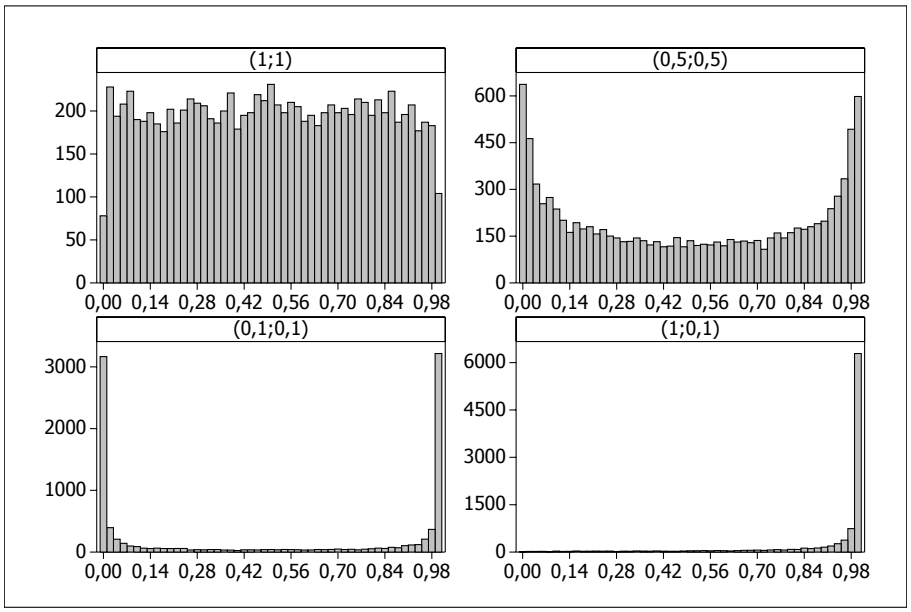

Figura 4.1: Distribuição Beta(α, β)

Na literatura não há nenhum esquema que sugira a geração desse tipo de dado. As idéias propostas no trabalho tentam representar uma situação em que algumas categorias têm maior probabilidade de ocorrer (modelo Beta) e outra em que todas as possíveis opções são igualmente prováveis (modelo Uniforme), como ilustra a Figura 4.2.

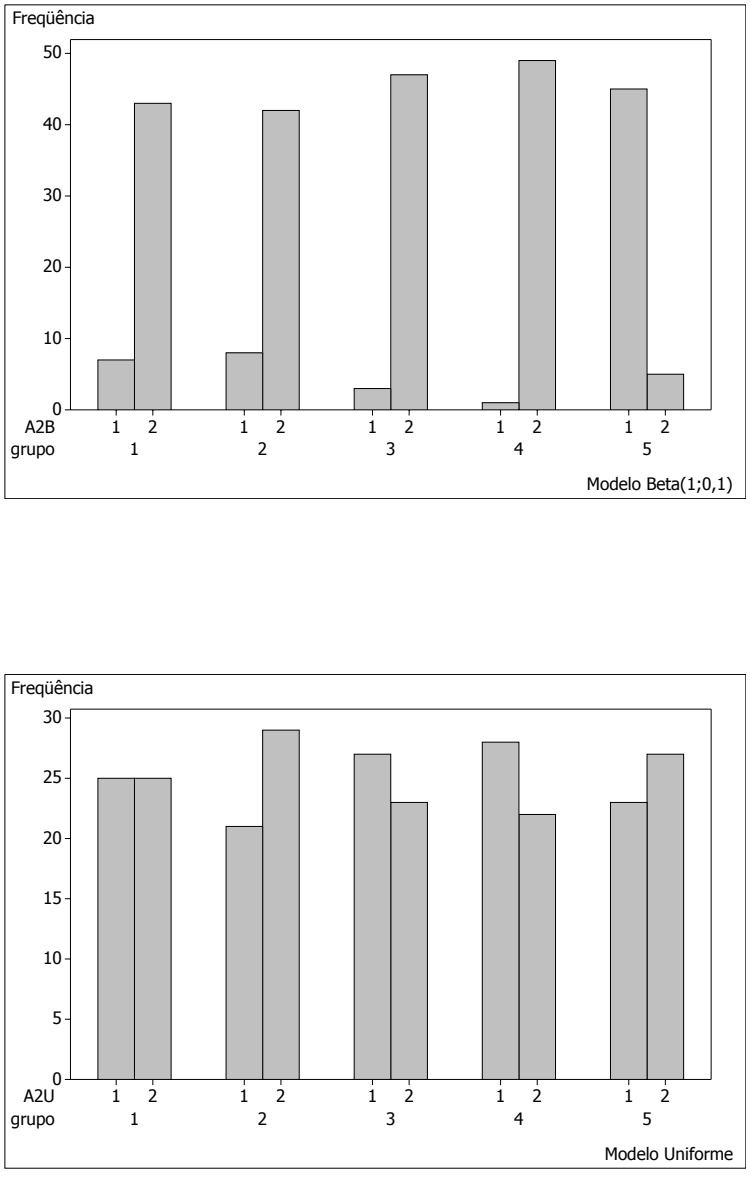

Figura 4.2: Distribuição de freqüências da variável A_2 para os modelos Beta$(1; 0, 1)$ (A_2B) e Uniforme (A_2U) segundo 5 grupos, cada um com 50 observações

.

A geração aleatória de observações foi focada nas variáveis categóricas nominais. Segundo Anderberg (1973), variáveis cuja escala de mensuração é ordinal também fazem parte das chamadas variáveis categóricas. Essas, porém, na análise de agrupamentos de observações, não contribuem com a propriedade que as difere das nominais, a ordenação. Isso porque nos métodos existentes não há nada especificadamente desenvolvido para usar essa informação adicional no processo de agrupamento. Outra possibilidade muita usada no tratamento desse tipo de atributo é considerá-lo como uma variável discreta. Assim sendo, a distância euclidiana (Equação (3.2)) pode ser usada para agrupar observações ou grupos. Se as variáveis ordinais puderem ser tratadas simplesmente como nominais, as conclusões desta dissertação se aplicam também à ocorrência desse tipo de atributo.

As situações 22 e 63, ambas para o modelo Beta, foram as selecionadas para compor a análise de consistência dos programas e do tempo de processamento das metodologias, respectivamente. Foi preciso fazer uma pequena modificação na situação 22 para que as observações estivessem alocadas em dois grupos completamente separados. O número de variáveis para a análise de consistência passou de quatro para duas. Na situação 63, nada foi alterado.

O número de simulações para a análise de consistência foi igual a 1.000 e como resultado tem-se que todos os métodos, sem exceção, tiveram taxa de alocação correta igual a 1 ou 100%. Isso significa que todos os procedimentos foram corretamente implementados.

4.1.3.3 Algumas definições

O número de grupos (k), de variáveis categóricas (m) e de categorias dessas variáveis ($|DOM(A_j)|$, com $j \in [1, m]$) são definidos pela situação simulada. Como o número de observações para cada um dos grupos foi fixado em 50, na situação mais simples, 100 observações foram geradas enquanto que na mais complexa, 250 observações.

Também já foram estabelecidos a medida de proximidade (seção 3.3.1), a escolha e comparação das sementes (seção 3.3.2), o número máximo de iterações (mesma seção), o número de simulações (seção 4.1) e, finalmente, os parâmetros $\theta = 0,30$ do *ROCK* (página 34) e $\alpha = 2$ dos métodos *Fuzzy* c-Modas e k-Populações (páginas 24 e 26).

4.1.3.4 Resultados

Para simplificar a apresentação dos resultados e não tornar a análise repetitiva, optou-se por classificar as estruturas dos dados da seguinte forma:

- os grupos podem estar dispostos de dois tipos: “cenário 1” quando os grupos são separados e “cenário 2” quando esses são sobrepostos;

- essa estrutura pode variar de acordo com o número de variáveis que determinam a separação e a sobreposição. O “grau de separação 1” diz respeito às situações em que uma única variável controla a separação dos grupos. Analogamente, o “grau de separação 2”, às situações em que duas variáveis proporcionam a separação dos grupos. Da mesma forma são definidos os graus de sobreposição 1, 2 e 3;

Cabe ressaltar aqui que uma vez que os grupos são separados em uma ou duas variáveis, isso não significa que eles são totalmente separados. Isso ocorre pelo fato de existirem outras variáveis, geradas aleatoriamente conforme apresentado nessa mesma seção, e que também caracterizam as observações dos grupos existentes. Esse procedimento de simulação implementado nesta dissertação é muito usual na área de análise de agrupamentos e é devido a Milligan (1980, 1985). Conforme será mostrado posteriormente, os atributos não controlados incorporam mais informações às observações e conseqüentemente podem auxiliar ou não no processo de agrupamento.

A análise do número de grupos, de variáveis e do grau de separação e sobreposição dos grupos é feita com base nas tabelas presentes no corpo do texto. Essas são resumos das tabelas apresentadas no Anexo, nas quais se baseia a análise do número de categorias.

4.1.3.4.1 Grau de separação e sobreposição .

O grau de separação e sobreposição está então relacionado com a quantidade de variáveis que determinam tais características. O aumento do primeiro grau indica a maior separação entre os grupos. Por outro lado, o aumento do segundo grau indica uma maior confusão entre os grupos, uma vez que há mais sobreposição. As Tabelas 4.3 e 4.4 apresentam as taxas de alocação correta de cada um dos métodos segundo o grau de separação e sobreposição. A primeira refere-se ao modelo Beta e a outra, ao Uniforme. Para que análise não fosse influenciada pelos demais fatores, optou-se por fixar o número de variáveis categóricas em quatro ($m = 4$) e construir a tabela segundo o número de grupos k. As situações em que há apenas duas variáveis categóricas ($m = 2$) foram desconsideradas pelo fato de se referirem somente ao caso em que os grupos são separados na primeira variável (grau de separação 1), não sendo possível assim compará-las com o grau de separação 2. Tabelas que levam em consideração todas as situações, ou seja, agregando-as independente do número de variáveis ($m = 2$ ou $m = 4$), foram elaboradas e a partir delas obtêm-se resultados bastante próximos dos comentados nesta seção.

Tabela 4.3: Taxa de alocação correta dos métodos segundo grau de separação e sobreposição e número de grupos (k), para quatro variáveis categóricas ($m = 4$) e modelo Beta$(1; 0, 1)$

	Separação		Sobreposição		
Método	1	2	1	2	3
			$k = 2$		
Ligação Média	91,83	**100,00**	**58,90**	50,00	-
ROCK	84,99	**96,64**	**76,46**	52,61	-
k-Modas	84,46	**98,21**	**66,92**	52,71	-
Fuzzy c-Modas	86,98	**98,73**	**59,73**	51,07	-
k-Populações	90,17	**98,08**	**60,19**	53,40	-
			$k = 3$		
Ligação Média	75,43	**100,00**	**52,79**	34,65	-
ROCK	78,22	**79,31**	**75,20**	68,56	-
k-Modas	68,82	**88,62**	**54,46**	44,00	-
Fuzzy c-Modas	71,84	**82,39**	**49,09**	44,65	-
k-Populações	83,71	**94,48**	**50,35**	39,66	-
			$k = 5$		
Ligação Média	59,52	**99,62**	**43,98**	20,02	26,70
ROCK	**64,79**	56,75	**67,06**	61,45	43,63
k-Modas	54,56	**72,89**	**42,29**	34,69	34,22
Fuzzy c-Modas	54,27	**64,86**	**36,64**	32,72	34,63
k-Populações	67,46	**87,29**	**44,00**	25,71	29,96
			Média		
Ligação Média	74,36	**99,86**	**51,89**	34,91	26,70
ROCK	75,14	**77,41**	**72,91**	60,18	43,63
k-Modas	68,15	**86,39**	**54,56**	43,78	34,22
Fuzzy c-Modas	69,74	**81,96**	**48,49**	42,65	34,63
k-Populações	79,45	**93,17**	**51,51**	39,58	29,96

Nota: "-" indica que nenhuma situação foi simulada com a configuração

Tabela 4.4: Taxa de alocação correta dos métodos segundo grau de separação e sobreposição e número de grupos (k), para quatro variáveis categóricas ($m = 4$) e modelo Uniforme

	Separação		Sobreposição		
Método	1	2	1	2	3
			$k = 2$		
Ligação Média	80,80	**100,00**	**51,91**	50,00	-
ROCK	79,08	**100,00**	**64,10**	50,00	-
k-Modas	71,43	**93,55**	**62,27**	53,94	-
Fuzzy c-Modas	75,69	**95,86**	**54,51**	51,57	-
k-Populações	79,76	**95,40**	**52,44**	50,75	-
			$k = 3$		
Ligação Média	64,33	**100,00**	**45,86**	34,66	-
ROCK	66,80	**69,78**	**47,23**	47,17	-
k-Modas	55,78	**77,40**	**49,29**	44,62	-
Fuzzy c-Modas	55,47	**69,51**	42,10	**43,01**	-
k-Populações	68,00	**91,63**	**41,63**	36,61	-
			$k = 5$		
Ligação Média	46,36	**100,00**	**34,49**	20,00	26,83
ROCK	**51,12**	39,44	**38,09**	35,23	34,61
k-Modas	44,21	**61,21**	**36,41**	33,04	33,91
Fuzzy c-Modas	41,20	**56,47**	29,41	30,26	**34,12**
k-Populações	51,99	**86,98**	**38,29**	23,86	29,51
			Média		
Ligação Média	62,49	**100,00**	**44,08**	34,91	26,83
ROCK	64,55	**69,74**	**49,81**	43,86	34,61
k-Modas	56,14	**77,38**	**49,32**	43,80	33,91
Fuzzy c-Modas	56,20	**74,35**	**42,00**	41,49	34,12
k-Populações	62,84	**91,31**	**44,12**	37,12	29,51

Nota: "-" indica que nenhuma situação foi simulada com a configuração

Conforme era esperado, observa-se que os métodos comparados, independente do número de grupos e também em média, têm desempenho melhor quando o grau de separação aumenta, e pior quando o grau de sobreposição aumenta. Destacam-se, no entanto, algumas exceções. Tanto no modelo Beta$(1; 0, 1)$ quanto no Uniforme, ao se passar do grau de separação 1 para o 2 e $k = 5$, o ROCK apresenta queda ($12,41\%$ e $22,85\%$). Ao se passar do grau de sobreposição 2 para o 3 e, de novo, $k = 5$, a Ligação Média, o *Fuzzy* c-Modas e o k-Populações apresentam aumento ($33,37\%$ e $34,15\%$, $5,84\%$ e $12,76\%$, e $16,53\%$ e $23,68\%$, para os modelos Beta e Uniforme respectivamente).

Apesar de todos os métodos terem, em média, o mesmo padrão, há alguns que se beneficiam mais e alguns que se prejudicam menos com o aumento desses graus. Em ambos os modelos, o método da Ligação Média é o que apresenta maior aumento nas taxas de alocação. Passando-se do grau de separação 1 para 2 a taxa de alocação média desse método aumenta $34,29\%$ no modelo Beta e $60,03\%$ no modelo Uniforme. Também em um e em outro modelo o método que menos se prejudica com o aumento do grau de sobreposição de 1 para 2 é o *Fuzzy c-Modas*. A redução da taxa de alocação correta desse método é de $12,04\%$ e $1,21\%$ nos modelos Beta e Uniforme, respectivamente.

O aumento e a diminuição da taxa de alocação quando se passa do grau 1 para o grau 2 de separação e de sobreposição variam de método para método. Porém, a redução ao se passar do grau 2 para o grau 3 de sobreposição está na faixa de 20%, para todas as metodologias.

Percebe-se ainda pela análise das Tabelas 4.3 e 4.4 que os métodos de agrupamentos tiveram um desempenho pior no modelo Uniforme, em comparação com o modelo Beta. Isso se deve ao fato de todas as categorias terem aproximadamente a mesma probabilidade de ocorrência. Outra observação que se faz é com respeito ao método *ROCK* no cenário em que há sobreposição. Nessas situações esse método é mais estável que as demais metodologias uma vez que suas taxas são pouco afetadas nas mudanças de um grau.

4.1.3.4.2 Número de grupos .

A Tabela 4.5 apresenta os resultados do modelo Beta segundo número de grupos, de variáveis e grau de separação e de sobreposição. Já a Tabela 4.6, segundo o número de grupos e cenário, independente do grau. Tanto em uma quanto na outra tabela, todas as situações simuladas estão sendo consideradas.

A segunda tabela (4.6) resume as informações da primeira (4.5) em relação ao número de grupos. Ela foi construída tirando-se uma média das situações de cada cenário, não

levando em conta o grau de separação e de sobreposição. Observa-se que a comparação do número de variáveis somente é feita na primeira tabela. Esses dados não aparecem na tabela-resumo porque eles ocorrem em apenas um dos graus, não necessitando assim que seja feita uma média global para cada cenário.

Tabela 4.5: Taxa de alocação correta dos métodos segundo número de grupos, de variáveis e disposição dos grupos, para modelo Beta$(1; 0, 1)$

	Grupos			Média	Variáveis	
Método	$k = 2$	$k = 3$	$k = 5$	global	$m = 2$	$m = 4$
Grau de separação 1						
Ligação Média	**93,09**	**82,05**	**65,34**	**80,16**	**89,59**	**74,36**
ROCK	88,07	**80,62**	**67,25**	**78,65**	**84,35**	**75,14**
k-Modas	85,49	70,13	55,83	70,48	74,27	68,15
Fuzzy c-Modas	**88,26**	72,02	54,75	71,68	74,83	69,74
k-Populações	**92,17**	**84,11**	**68,95**	**81,74**	**85,46**	**79,45**
Grau de separação 2						
Ligação Média	**100,00**	**100,00**	**99,62**	**99,86**	-	**99,86**
ROCK	96,64	79,31	56,75	77,41	-	77,41
k-Modas	**98,21**	**88,62**	**72,89**	**86,39**	-	**86,39**
Fuzzy c-Modas	**98,73**	82,39	64,86	81,96	-	81,96
k-Populações	98,08	**94,48**	**87,29**	**93,17**	-	**93,17**
Grau de sobreposição 1						
Ligação Média	56,86	46,57	37,91	47,66	42,89	**51,89**
ROCK	**68,72**	**62,91**	**55,27**	**62,71**	**51,24**	**72,91**
k-Modas	**65,00**	**54,17**	**41,95**	**54,40**	**54,21**	**54,56**
Fuzzy c-Modas	57,29	**47,39**	35,09	47,26	**45,89**	48,49
k-Populações	**58,28**	47,30	**39,39**	**48,85**	45,85	51,51
Grau de sobreposição 2						
Ligação Média	50,00	34,65	20,02	34,91	-	34,91
ROCK	**52,61**	**68,56**	**61,45**	**60,18**	-	**60,18**
k-Modas	**52,71**	**44,00**	**34,69**	**43,78**	-	**43,78**
Fuzzy c-Modas	51,07	**44,65**	**32,72**	**42,65**	-	**42,65**
k-Populações	**53,40**	39,66	25,71	39,58	-	39,58
Grau de sobreposição 3						
Ligação Média	-	-	26,70	26,70	-	26,70
ROCK	-	-	**43,63**	**43,63**	-	**43,63**
k-Modas	-	-	**34,22**	**34,22**	-	**34,22**
Fuzzy c-Modas	-	-	**34,63**	**34,63**	-	**34,63**
k-Populações	-	-	29,96	29,96	-	29,96

Nota: "-" indica que nenhuma situação foi simulada com a configuração

Tabela 4.6: Média da taxa de alocação correta dos métodos segundo número de grupos e disposição dos grupos, para modelo Beta$(1; 0, 1)$

Método	Grupos $k=2$	Grupos $k=3$	Grupos $k=5$	Média global
Grupos separados				
Ligação Média	**95,60**	**87,43**	**77,81**	**86,93**
ROCK	**91,19**	**80,23**	**63,43**	**78,22**
k-Modas	90,12	75,67	62,03	75,95
Fuzzy c-Modas	92,07	75,13	58,43	75,21
k-Populações	**94,32**	**87,22**	**75,62**	**85,67**
Grupos sobrepostos				
Ligação Média	54,12	42,60	29,14	41,11
ROCK	**62,27**	**64,79**	**54,42**	**59,97**
k-Modas	**60,08**	**50,78**	**37,60**	**48,68**
Fuzzy c-Modas	54,80	**46,47**	**34,18**	**44,40**
k-Populações	**56,33**	44,76	32,47	43,73

Percebe-se que em todas as situações, independente do cenário, o aumento do número de grupos afeta negativamente o desempenho dos métodos, sem exceção. Ou seja, a taxa de alocação correta diminui com o aumento dessa quantidade.

Chamam a atenção os seguintes resultados:

- no grau de separação 2, o aumento do número de grupos tem efeito bem maior nos métodos ROCK, k-Modas e *Fuzzy* c-Modas, enquanto as demais metodologias conseguem manter uma boa taxa de alocação correta (Ligação Média acima de 99% e k-Populações acima de 87%). O ROCK tem taxa de $56,75\%$, o k-Modas $72,89\%$ e o *Fuzzy*, $64,86\%$, quando $k = 5$ (Tabela 4.5);

- no cenário em que os grupos são separados, os melhores desempenhos são obtidos pela Ligação Média que, em média, tem taxa de alocação correta de $86,93\%$, pelo k-Populações, cuja taxa média é $85,67\%$ e pelo ROCK, com $78,22\%$ de taxa (Tabela 4.6);

- no grau de sobreposição 1, os métodos que agora sofrem mais o efeito do número de grupos são: Ligação Média, *Fuzzy* c-Modas e k-Populações. O k-Modas também compõe essa lista quando o grau de sobreposição é 2. No 3, ele e o *Fuzzy* c-Modas

deixam de fazer parte dos métodos que têm o desempenho bastante afetado pelo aumento dos grupos. De forma geral, considerando-se todos os graus de sobreposição, a Ligação Média e o k-Populações são as metodologias que sofrem mais o efeito aqui comentado. Enquanto o ROCK consegue manter a taxa de alocação em $61,45\%$ no grau de sobreposição 2 e $k = 5$, a Ligação Média tem taxa de $20,02\%$ e o k-Populações, de $25,71\%$ (Tabela 4.5);

- no cenário de grupos sobrepostos os melhores desempenhos são obtidos pelo ROCK, com taxa de alocação média de $59,97\%$, pelo k-Modas, que tem taxa de $48,68\%$ e pelo *Fuzzy* c-Modas, cuja taxa é $44,40\%$ (Tabela 4.6);

Em resumo, para o modelo Beta, salvo o *ROCK*, os métodos com melhor desempenho no cenário 1 (Ligação Média e k-Populações) não são os mesmos com melhor desempenho no cenário 2 (k-Modas e k-Populações).

Para o modelo Uniforme (Tabelas 4.7 e 4.8), de forma geral, a análise é bastante análoga. As maiores taxas de alocação correta quando os grupos são separados são obtidas pelos métodos Ligação Média, k-Populações, *ROCK* e k-Modas. Quando os grupos são sobrepostos, pelos métodos *ROCK*, pelo k-Modas e pelo *Fuzzy* c-Modas. Destacam-se nesse modelo o *ROCK* e o k-Modas porque, independente da disposição dos grupos, têm bons resultados. Em média, o *ROCK* continua tendo destaque uma vez que as suas taxas são superiores as do k-Modas. Assim, novamente, essa seria a única técnica que tem desempenhos satisfatórios quando os grupos são separados e quando esses são sobrepostos.

Tabela 4.7: Taxa de alocação correta dos métodos segundo número de grupos, de variáveis e disposição dos grupos, para modelo Uniforme

	Grupos			Média	Variáveis	
Método	$k = 2$	$k = 3$	$k = 5$	global	$m = 2$	$m = 4$
Grau de separação 1						
Ligação Média	**85,95**	**74,87**	**55,15**	**71,99**	**87,44**	**62,49**
ROCK	**83,89**	**74,77**	**57,23**	**71,97**	**84,02**	**64,55**
k-Modas	76,47	60,59	47,91	61,66	70,62	56,14
Fuzzy c-Modas	80,70	59,24	44,01	61,32	69,62	56,20
k-Populações	**84,81**	**73,88**	**51,99**	**70,23**	**82,22**	**62,84**
Grau de separação 2						
Ligação Média	**100,00**	**100,00**	**100,00**	**100,00**	-	**100,00**
ROCK	**100,00**	69,78	39,44	69,74	-	69,74
k-Modas	93,55	**77,40**	**61,21**	**77,38**	-	**77,38**
Fuzzy c-Modas	**95,86**	69,51	56,47	74,35	-	74,35
k-Populações	95,40	**91,63**	**86,98**	**91,31**	-	**91,31**
Grau de sobreposição 1						
Ligação Média	51,44	41,77	31,07	42,03	39,73	44,08
ROCK	**60,61**	**44,01**	**33,56**	**46,79**	43,41	**49,81**
k-Modas	**62,81**	**50,89**	**37,41**	**51,13**	**53,17**	**49,32**
Fuzzy c-Modas	**53,90**	**42,61**	29,52	42,75	**43,58**	42,00
k-Populações	53,09	41,65	**35,34**	**43,83**	**43,51**	**44,12**
Grau de sobreposição 2						
Ligação Média	50,00	34,66	20,00	34,91	-	34,91
ROCK	50,00	**47,17**	**35,23**	**43,86**	-	**43,86**
k-Modas	**53,94**	**44,62**	**33,04**	**43,80**	-	**43,80**
Fuzzy c-Modas	**51,57**	**43,01**	**30,26**	**41,49**	-	**41,49**
k-Populações	**50,75**	36,61	23,86	37,12	-	37,12
Grau de sobreposição 3						
Ligação Média	-	-	26,83	26,83	-	26,83
ROCK	-	-	**34,61**	**34,61**	-	**34,61**
k-Modas	-	-	**33,91**	**33,91**	-	**33,91**
Fuzzy c-Modas	-	-	**34,12**	**34,12**	-	**34,12**
k-Populações	-	-	29,51	29,51	-	29,51

Nota: "-" indica que nenhuma situação foi simulada com a configuração

Tabela 4.8: Média da taxa de alocação correta dos métodos segundo número de grupos e disposição dos grupos, para modelo Uniforme

	Grupos			Média
Método	$k = 2$	$k = 3$	$k = 5$	global
Grupos separados				
Ligação Média	**91,06**	**82,41**	**71,46**	**81,62**
ROCK	**89,75**	**73,28**	50,76	**71,20**
k-Modas	82,68	65,63	**52,75**	67,06
Fuzzy c-Modas	86,21	62,32	48,54	65,80
k-Populações	**88,66**	**79,20**	**64,71**	**77,47**
Grupos sobrepostos				
Ligação Média	50,86	39,40	26,32	38,03
ROCK	**56,37**	**45,06**	**34,38**	**44,57**
k-Modas	**59,26**	**48,80**	**35,08**	**46,86**
Fuzzy c-Modas	**52,97**	**42,74**	**30,92**	**41,46**
k-Populações	52,15	39,97	30,06	40,06

4.1.3.4.3 Número de variáveis .

As Tabelas de 4.5 e 4.7 da análise do número de grupos trazem também informações referentes ao número de variáveis, para os modelos Beta e Uniforme. Em ambos os modelos, no grau de separação 1 (cenário 1), aumentado-se o número de variáveis, diminui-se a taxa de alocação correta. Já no grau de sobreposição 2 (cenário 2), ocorre o inverso. Aumentando a quantidade de atributos, melhora-se o desempenho dos métodos comparados.

Destaca-se porém a pequena redução da taxa de alocação correta dos métodos k-Modas e *Fuzzy* c-Modas no grau de sobreposição 1 para o modelo Uniforme. Ao se aumentar o número de variáveis de duas para quatro as taxas desses métodos passam de $53,17\%$ e $43,58\%$ para $49,32\%$ (redução de $7,81\%$) e 42% (redução de $3,76\%$), respectivamente. Enquanto isso, as demais técnicas têm os seus desempenhos melhorados.

As novas variáveis, ao serem incorporadas ao banco de dados, contribuem com o aumento da informação a ser tratada. Nenhuma suposição é feita com respeito a ocorrência das novas variáveis no que se refere à separação e à sobreposição dos grupos. O que se percebe é que a informação adicional, no primeiro cenário, prejudica a estrutura de separação imposta aos grupos e por isso os resultados são inferiores. Por outro lado, ela

ameniza a estrutura de sobreposição e conseqüentemente os métodos conseguem melhor agrupar as observações.

Ainda analisando-se os resultados referentes ao número de variáveis, chamam a atenção os seguintes resultados:

- no modelo Beta, os métodos que possuem melhor desempenho no cenário 1 quando $p = 2$ são a Ligação Média $(89,59\%)$, k-Populações $(85,46\%)$ e *ROCK* $(84,35\%)$. Quando $p = 4$, destacam-se os mesmos métodos, porém em outra ordem: k-Populações $(79,45\%)$, *ROCK* $(75,14\%)$ e Ligação Média $(74,36\%)$.

- no cenário 2, os melhores resultados são alcançados pelos métodos k-Modas $(54,21\%)$, *ROCK* $(51,24\%)$, *Fuzzy* c-Modas $(45,89\%)$ e k-Populações $(45,85\%)$ quando $p = 2$. Para $p = 4$, o *Fuzzy* tradicional e a sua modificação dão lugar para a Ligação Média, porém o k-Populações tem taxa próxima dessa última técnica. Além disso, o *ROCK* tem melhor desempenho que o k-Modas. Nesse contexto tem-se: *ROCK* $(72,91\%)$, k-Modas $(54,56\%)$, Ligação Média $(51,89\%)$ e k-Populações $(51,51\%)$;

- em ambos cenários no modelo Uniforme, as conclusões são análogas às obtidas no modelo Beta. No cenário 1 e $p = 2$ tem-se: Ligação Média $(87,44\%)$, *ROCK* $(84,02\%)$ e k-Populações $(82,22\%)$. Para $p = 4$ tem-se: *ROCK* $(64,55\%)$, k-Populações $(62,84\%)$ e Ligação Média $(62,49\%)$.

- no cenário 2 e $p = 2$ tem-se: k-Modas $(53,17\%)$, *Fuzzy* c-Modas $(43,58\%)$, k-Populações $(43,51\%)$ e *ROCK* $(43,41\%)$. Para $p = 4$ tem-se: ROCK $(49,81\%)$, k-Modas $(49,32\%)$, k-Populações $(44,12\%)$ e Ligação Média $(44,08\%)$.

4.1.3.4.4 Número de categorias .

As Tabelas A.1 e A.2 do Anexo apresentam as taxas de alocação correta dos métodos em cada uma das situações descritas na Tabela 4.2. Elas estão divididas de acordo com a estrutura dos dados gerados (cenários e graus), mas para facilitar a análise do número de categorias, as situações estão dispostas em ordem crescente de acordo com essa quantidade.

Para o modelo Beta (Tabela A.1) e cenário 1 (grupos separados), observa-se o seguinte:

- os métodos Ligação Média e k-Populações têm as suas maiores taxas de alocação correta nas situações em que o número de categorias das variáveis é maior.

Percebe-se que o aumento dessa quantidade proporciona melhor desempenho dessas metodologias;

- a taxa de alocação correta do *ROCK*, por sua vez, também aumenta com o aumento do número de categorias das variáveis. Porém esse método possui uma peculiaridade. Se essa última quantidade aumenta igualmente, por exemplo, passando-se da situação 20 ($Dom(A_i) = 2,\ \forall\ i = 2, \ldots, 4$) , vide Tabela 4.2 na página 52, para a 17 ($Dom(A_i) = 5,\ \forall\ i = 2, \ldots, 4$) e em seguida para a 21 ($Dom(A_i) = 8,\ \forall\ i = 2, \ldots, 4$), a taxa aumenta de $64,67\%$ para $66,36\%$ e para $70,92\%$. Mas se aquela mesma quantidade não aumenta igualmente, por exemplo, passando-se da situação 18 ($Dom(A_2) = 2, Dom(A_i) = 4,\ \forall\ i = 3, 4$) para a 19 ($Dom(A_2) = 3, Dom(A_3) = 4\ e\ Dom(A_4) = 10$), a taxa cai de $62,72\%$ para $59,27\%$. Outro exemplo é a quando se aumenta o número de categorias de três (situação 30) para cinco e 10 (situação 31) nas duas últimas de quatro variáveis. Nesse caso a taxa de alocação correta cai de $58,53\%$ para $54,01\%$. As exceções ocorrem quando $k = 3$, ou seja, passando-se da situação 13 para a 16 e da 26 ou da 27 para a 28;

- em geral, os métodos k-Modas e *Fuzzy* c-Modas têm as suas taxas um pouco diminuídas quando o número de categorias aumenta igualmente. Em algumas situações, como nas mesmas situações 13 e 16 do parágrafo anterior, as taxas não se alteram. O k-Modas passa de $68,69\%$ para $68,86\%$ e no *Fuzzy* c-Modas, de $72,13\%$ para $72,15\%$. Quando uma variável possui domínio maior que as demais, ou seja, quando há desbalanceamento no número de categorias, não se pode observar um comportamento único para esses métodos. Os desempenhos tanto melhoram (passando-se, por exemplo, da situação 22 para a 24, em ambos os métodos) quanto pioram (da situação 14 para a 15 no k-Modas e da 18 para a 19 no *Fuzzy*) como também não se alteram (da situação 18 para a 19 no k-Modas e da 14 para a 15 no *Fuzzy*).

No mesmo modelo, mas agora no cenário 2 (grupos sobrepostos), o que ocorre é o inverso:

- os métodos Ligação Média, k-Modas, *Fuzzy* c-Modas e k-Populações têm desempenhos piores quando o número de categorias é maior. Destaca-se que há pouquíssima diferença entre as taxas de alocação da Ligação Média e do *Fuzzy* c-Modas nas situações em que o grau de sobreposição é 2;

- as taxas do *ROCK* também diminuem quando o aumento desse número é desigual. Por exemplo, ao se passar da situação 44 ($Dom(A_2) = 2\ e\ Dom(A_i) = 3,\ \forall\ i = 3, 4$)

para a 45 ($Dom(A_2) = 2, Dom(A_3) = 5$ *e* $Dom(A_4) = 10$) ou da 54 ($Dom(A_i) = 3, \forall\ i = 2, \ldots, 4$) para a 56. ($Dom(A_2) = 3, Dom(A_3) = 5$ *e* $Dom(A_4) = 10$). No primeiro caso a taxa de alocação do *ROCK* cai de $81,78\%$ para $70,50\%$ e no segundo, de $84,54\%$ para $44,15\%$;

- para $k = 3$ e $k = 5$, quando o aumento do número de categorias é igual, a tendência do *ROCK* é aumentar a taxa de alocação correta. Como exemplo têm-se as situações $\{55, 54\}$ e $\{60, 58, 57\}$. Passando-se da 55 para a 54 a taxa aumenta de 77% para $84,54\%$ e da 60 para a 58 e em seguida para a 57, de $58,10\%$ para $61,69\%$ e finalmente para $69,91\%$. Nesses casos, o aumento do número de categorias acontece nas duas últimas, das quatro variáveis existentes. Nas primeiras situações essa quantidade passa de duas para três e nas outras situações, de duas para três e finalmente para cinco;

- para $k = 2$ ocorre o inverso. O aumento igual do número de categorias diminui as taxas de alocação dos métodos analisados. Um exemplo disso é o que acontece entre as situações $\{50, 51, 53\}$. O *ROCK* na primeira situação tem taxa de $57,88\%$, na segunda de $50,30\%$ e na última de 50%.

Em resumo, no modelo Beta o aumento do número de categorias influencia os resultados das metodologias. E essa influência acontece de modo diferente, dependendo do método. O aumento dessa quantidade implica nos seguintes resultados:

- melhora o desempenho da Ligação Média e do k-Populações quando os grupos são separados e o piora quando esses são sobrepostos;

- melhora o desempenho do *ROCK* quando o aumento é igual é o cenário é o 1 e o piora quando é desigual, no mesmo cenário. No 2 o desempenho também piora quando o aumento do número de categorias é desigual. E quando é igual, nada se pode concluir pelos resultados obtidos;

- piora o desempenho do k-Modas e do *Fuzzy* c-Modas quando o aumento é igual no cenário 1 e, independente da forma, no cenário 2 também. Quando os grupos são separados e o aumento do número de categorias é desigual, nada se pode afirmar.

Para o modelo Uniforme (Tabela A.2) e cenário 1, valem os comentários referentes aos métodos Ligação Média e k-Populações para o modelo Beta e mesmo cenário. Tais

comentários também se aplicam às demais metodologias. Assim, o aumento do número de categorias proporciona melhoria no desempenho dos métodos analisados.

Cabe ressaltar, no entanto, que nas situações em que os grupos são separados na primeira de quatro variáveis (grau de separação 1 e $m = 4$), o *ROCK* e o k-Modas não possuem esse comportamento tão evidente. Pode-se até pensar que o aumento do número de categorias não melhora o desempenho dos métodos. O fato é que nesses casos faz-se necessário analisar separadamente o aumento igual e o aumento desigual desse número. A partir daí é possível constatar que as taxas realmente elevam-se.

Quando os grupos são sobrepostos a análise é um pouco mais complexa. No cenário 2, grau de sobreposição 1 e $m = 2$, o aumento do número de categorias tende a diminuir as taxas de alocação correta da Ligação Média, do *ROCK* e do k-Modas. O *Fuzzy* c-Modas e o k-Populações têm o desempenho um pouco melhorado nesse contexto. O aumento das taxas das situações 38 e 40 demonstra esse fato. Já quando $m = 4$, destacam-se os seguintes resultados:

- o desempenho de todos os métodos comparados ou se mantém (situações 41 e 42) ou é melhorado (situações 47 e 48) com o aumento desigual do número de categorias;

- ao se alterar os domínios das variáveis de $\{k, 2, 3, 3\}$ para $\{k, 5, 5, 5\}$ (situações $\{41, 43\}$, $\{44, 46\}$ e $\{47, 49\}$), as taxas da Ligação Média e do *Fuzzy* c-Modas diminuem. Por outro lado, as taxas do *ROCK* (para $k = 5$), do k-Modas (para qualquer número de grupos) e do k-Populações (exceto quando $k = 2$) aumentam. Os desempenhos melhoram pouco, mas chama a atenção o método k-Populações ao se passar da situação 47 $(28,30\%)$ para a 49 $(49,14\%)$.

No grau de sobreposição 2, o aumento do número de categorias não parece afetar o desempenho dos métodos Ligação Média e *Fuzzy* c-Modas. Quando esse aumento é igual, porém, o k-Modas e o k-Populações têm as suas taxas diminuídas. Observe as situações 55 e 54. As duas últimas variáveis, das quatro existentes, passa a ter três categorias ao invés de duas e as taxas dessa técnicas passam de $45,93\%$ e $37,41\%$ para $44,84\%$ e $35,53\%$. Enquanto isso os resultados do *ROCK* aumentam. Nas mesmas situações a taxa desse método passa de $50,26\%$ para $56,59\%$.

Quando o aumento é desigual, o desempenho do *ROCK* e do k-Modas, em geral, diminuem. Para o k-Populações não é possível determinar um único efeito, pois as suas taxas tanto se mantém (situações 51 e 52), quanto aumentam (situações 58 e 59) como também diminuem (situações 55 e 56) com o aumento o número de categorias.

No grau de sobreposição 3, todos os métodos comparados têm as taxas diminuídas quando o aumento do número de categorias ocorre de forma desigual (passando-se da situação 61 para a 62).

No modelo Uniforme, assim como no Beta, o número de categorias é um fator que influencia o desempenho das técnicas estudadas. O aumento dessa quantidade causa nesse contexto a melhora dos resultados quando os grupos são separados e, de forma geral, a piora quando esses são sobrepostos. As taxas nesse cenário tendem também a se manter. Somente o *ROCK* quando o aumento é igual no grau de sobreposição 2 apresenta aumento das taxas quando o número de categorias é maior.

4.1.3.4.5 Situações gerais .

As situações 64 e 65, correspondentes aos casos 8 e 9, foram incluídas nos estudos de simulação objetivando verificar o desempenho dos métodos primeiro em 1.000 bancos diferentes na estrutura, no número de grupos, de variáveis e de categorias e segundo em 1.000 outros bancos, também diferentes na estrutura. Esse agora continha dois grupos e 15 variáveis, cada uma com duas categorias (Tabela 4.2). Em ambas situações, nenhuma estrutura de separação ou sobreposição foi imposta aos grupos construídos.

Para o modelo Beta, a ordem das metodologias da mais para a menos eficiente na situação 64 é: *ROCK*, k-Populações, Ligação Média, k-Modas e *Fuzzy* c-Modas. Na situação 65: k-Populações, *ROCK*, Ligação Média, k-Modas e *Fuzzy* c-Modas. Para o Uniforme, tem-se na situação 64: k-Populações, *ROCK*, Ligação Média, k-Modas e *Fuzzy* c-Modas. Na situação 65: *ROCK*, Ligação Média, k-Modas, k-Populações e *Fuzzy* c-Modas.

Observa-se que a Ligação Média e o *ROCK* são os dois, dentre os cinco métodos comparados, que aparecem entre as três metodologias mais eficientes em ambas situações e modelos.

4.1.3.4.6 Breve Estudo dos Parâmetros θ e α .

Conforme se sabe, os métodos *ROCK*, *Fuzzy* c-Modas e k-Protótipos dependem dos parâmetros envolvidos nos respectivos algoritmos. Esta dissertação baseou-se na literatura para fixar o valor dessas quantidades em $\theta = 0,30$ (página 34) e $\alpha = 2$ (páginas 24 e 26), porém fez-se um breve estudo considerando algumas das situações apresentadas na Tabela 4.2. Escolheu-se de cada caso, desconsiderando-se os casos gerais (8 e 9),

aquelas situações que preservavam o número de grupos ($k = 5$), de variáveis ($m = 2$ ou $m = 4$) e de categorias ($|DOM(A_j)| = 5$, $\forall\ j \in [1, m]$). As situações escolhidas foram $\{7; 17; 29; 39; 49; 57; 63\}$. E os conjuntos $\{0,15; 0,25; 0,30; 0,33; 0,50; 0,67; 0,70; 0,75; 0,90\}$ e $\{0,5; 2; 5; 10; 50\}$ contêm os valores testados para os parâmetros θ e α, respectivamente, em ambos os modelos.

Pelos resultados obtidos confirmou-se que realmente o desempenho desses métodos é altamente dependente da escolha dos seus parâmetros. Tanto no modelo Beta quanto no Uniforme, observou-se que valores de θ entre $0,50$ e $0,75$ fazem com que o *ROCK* tenha altas taxas de alocação correta quando os grupos são separados (situações 7, 17 e 29) e quando dois grupos são sobrepostos na primeira de duas (situação 39) e de quatro variáveis (situação 49). Valores desse parâmetro entre $0,30$ e $0,33$ fazem com que essa técnica tenha melhor rendimento quando três ou cinco grupos são sobrepostos, ou seja, quando há maior confusão.

Em ambos os modelos, os métodos *Fuzzy* c-Modas e k-Protótipos têm bons desempenhos quando os grupos são separados (situações 7, 17 e 29) e $\alpha \geq\ 2$. As taxas de alocação correta desses métodos nesse cenário e α igual a qualquer um dos valores do conjunto $\{2; 5; 10; 50\}$ são bastante próximas. No cenário de sobreposição, os resultados do *Fuzzy* parecem independer da escolha do seu parâmetro. Isso porque as taxas são, de novo, muito próximas. No k-Protótipos, por outro lado, $\alpha = 0,5$ e $\alpha = 50$ fornecem os melhores resultados.

Percebe-se que estudos mais aprofundados são necessários. Talvez seja possível encontrar uma regra geral para se escolher o valor que se deve atribuir aos parâmetros θ e α desses métodos e assim resolver a principal dificuldade na aplicação dessas metodologias.

4.1.4 Variáveis Categóricas e Contínuas

4.1.4.1 Situações Simuladas

Quanto à separação e à sobreposição dos grupos, quatro estruturas podem acontecer agora: os grupos podem ser separados ou sobrepostos, simultaneamente, nos dois tipos de variáveis ou podem ser sobrepostos nas variáveis categóricas e separados nas contínuas ou o inverso, separados nas categóricas e sobrepostos nas contínuas. Selecionou-se para este trabalho a situação em que os grupos têm interseção nos dois tipos de variáveis.

Para construir as situações a serem simuladas nessa etapa, combinaram-se aquelas em que ocorre sobreposição nas categóricas (casos de 4 a 7 da seção 4.1.3) com estruturas

para variáveis contínuas utilizadas nas dissertações de Lima (2001) e Felix (2004). Essas últimas situações têm como suporte o trabalho desenvolvido por Milligan (1985). Nesse artigo os autores desenvolvem uma estratégia de simulação de dados para análise de agrupamentos, que vem sendo aplicada em vários trabalhos.

As situações categóricas escolhidas foram aquelas em que o número de categorias das variáveis era idêntico, para todos os atributos existentes (Tabela 4.9). Assim, as situações estudadas foram $\{1, 4, 7, 9, 12, 17, 18, 22, 25, 31\}$ e, portanto, o número de categorias não é um fator a ser estudado nesta simulação. Quando variáveis contínuas são incorporadas ao banco de dados, quatro novos aspectos podem ser estudados: o número de variáveis desse tipo, o grau de sobreposição delas, a correlação entre elas e a atribuição de pesos da medida de proximidade combinada. A combinação desses fatores determina o número total de situações distintas que foram construídas para o estudo.

Tabela 4.9: Número de grupos (k), de variáveis (p) e de categorias das variáveis categóricas ($|DOM(A_j)|$, com $j \in [1, m]$) das situações estudadas

Caso	Situação	k	p	$\|DOM(A_1)\|$	$\|DOM(A_2)\|$	$\|DOM(A_3)\|$	$\|DOM(A_4)\|$
Grupos sobrepostos na 1^a variável							
1	1	2	4	2	2		
	4	3	4	3	3		
	7	5	4	5	5		
2	9	2	8	2	2	2	2
	12	3	8	3	3	3	3
	17	5	8	5	5	5	5
Grupos sobrepostos nas duas 1^{as} variáveis							
3	18	2	8	2	2	2	2
	22	3	8	3	3	3	3
	25	5	8	5	5	5	5
Grupos sobrepostos nas três 1^{as} variáveis							
4	31	5	8	5	5	5	5

Mais uma vez, é claro que todas as possíveis situações reais não estão sendo abordadas, mas com base em um esquema de análise foi possível estudar o efeito do número de grupos e de variáveis. E ainda o efeito da correlação entre as variáveis contínuas, dos pesos da medida de proximidade combinada e da sobreposição nas duas variáveis. Assim foi possível se ter uma noção do desempenho das metodologias nos tantos casos propostos.

- **Estratégia para análise de desempenho**

  1. Grau de sobreposição:
     - situações $\{1, 4, 7\}$ e grau 1 de sobreposição das contínuas *versus* situações $\{9, 12, 17\}$ e grau 2 $(2 > 1)$ de sobreposição das contínuas *versus* situações $\{18, 22, 25, 31\}$ e grau 3 $(3 > 2)$ de sobreposição das contínuas;
  2. Número de grupos: situações $\{1, 9, 18\}$ *versus* situações $\{4, 12, 22\}$ *versus* situações $\{7, 17, 25, 31\}$;
  3. Número de variáveis: situações $\{1, 4, 7\}$ *versus* situações $\{9, 12, 17\}$;
  4. Estruturas de correlação e pesos.

A situação 31 foi a escolhida para se avaliar o tempo de processamento dos métodos comparados e uma outra situação hipotética foi construída para avaliar a consistência dos programas desenvolvidos. Nessa situação haviam dois grupos e eles eram completamente separados tanto nas variáveis categóricas quanto nas contínuas. A quantidade de cada tipo era 2, assim como o número de categorias dos atributos categóricos. Todos os métodos, se exceção, tiveram taxa de alocação correta igual a 1, ou 100%. De novo foram 1.000 o número de simulações feitas para a análise de consistência.

4.1.4.2 Geração Aleatória das Observações Contínuas

A estratégia comumente usada para gerar observações de variáveis contínuas baseia-se na idéia de que, para cada um dos k grupos, as $(p - m)$ variáveis contínuas possuem distribuição normal $(p - m)$-variada. Assim, basta definir os parâmetros que determinam a estrutura dos grupos, que são o vetor de médias e a matriz de covariâncias. Esses fatores podem ou não variar de um grupo para outro.

Neste texto serão apresentadas somente as informações necessárias para entender os procedimentos utilizados. Para detalhes sugere-se consultar as dissertações de Lima (2001) e Felix (2004).

Os grupos separados não é foco do trabalho, porém entende-se que é necessário explicar essa estrutura para melhor compreender os grupos sobrepostos. Definidos o número de grupos k e de variáveis $(p - m)$ contínuas e a estrutura de correlação entre essas variáveis, os passos seguidos para gerar observações de grupos separados são estes:

1. inicializar o gerador de números aleatórios e as matrizes necessárias para o desenvolvimento do algoritmo;

2. gerar os limites dos grupos para a primeira variável da seguinte forma:

   - aleatorizar os grupos para essa variável;
   - gerar o desvio padrão a partir de uma distribuição Uniforme $(10, 40)$;
   - definir que a amplitude desse grupo será três vezes o desvio padrão obtido;
   - calcular a média como sendo o ponto médio desse intervalo;
   - determinar as amplitudes dos demais grupos de maneira idêntica;
   - separar os grupos usando uma quantidade aleatória $QA_{(l)}$, em que (l) é a posição do grupo l (G_l). Sejam f uma variável aleatória da distribuição Uniforme $\left(\frac{1}{4}, \frac{3}{4}\right)$ e s_z e s_l os desvios padrões de G_l e G_z, nessa ordem. Assim $QA_{(l)} = f \times (s_l + s_z)$. O grupo G_z é o imediatamente seguinte a G_l.

3. repetir o passo anterior para as demais variáveis. Só que agora a amplitude máxima dos dados é limitada em duas ou três vezes a amplitude da primeira variável;

4. obter a matriz de covariâncias a partir da estrutura de correlação definida e as variâncias das variáveis geradas nos passos 2 e 3.

A Figura 4.3 ilustra a configuração dos grupos separados para uma determinada variável. Se $l = 3$, então $z = 1$ e portanto $QA_{(3)} = QA_1 = f \times (s_3 + s_1)$.

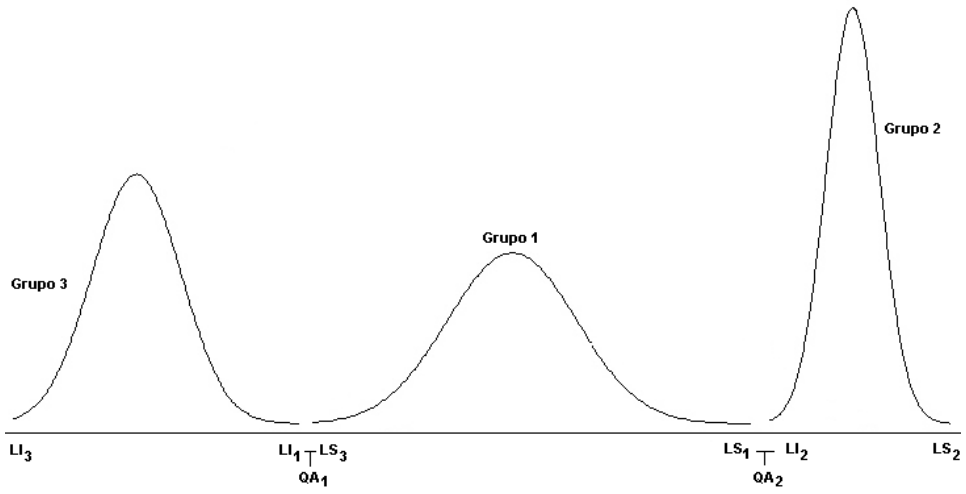

Figura 4.3: Grupos separados em uma variável contínua

Para gerar observações de grupos sobrepostos, é necessário modificar o último item do passo 2 anterior. Agora o limite inferior do grupo G_z não mais será maior que o limite superior do grupo G_l. Para melhor controlar o grau de sobreposição, utiliza-se a taxa de interseção dos grupos (int): $LI_z = (1 - int) \times (LS_l - LI_l) + LI_l$, em que LI_l e LI_z são os limites inferiores de G_l e G_z, respectivamente, e LS_l o limite superior de G_l.

A Figura 4.4 ilustra a configuração dos grupos sobrepostos para uma determinada variável. A quantidade $R_{(l)}$ que aparece na figura é igual $LS_l - LI_z$, em que (l) é, de novo, a posição de G_l. Aplicando a fórmula de LI_z do parágrafo anterior, tem-se: $R_{(l)} = LS_l - [(1 - int) \times (LS_l - LI_l) + LI_l] = int \times (LS_l - LI_l)$. Se $l = 3$, então $z = 1$ e portanto $R_{(3)} = R_1 = LS_3 - LI_1 = int \times (LS_3 - LI_3)$.

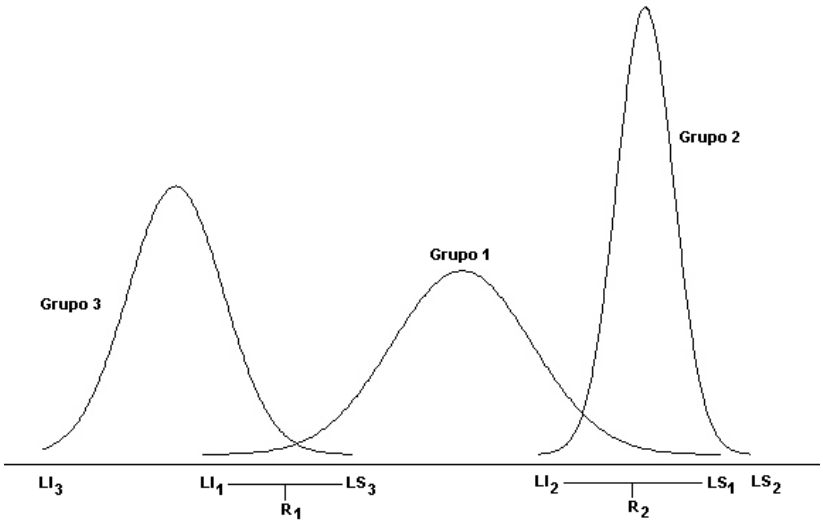

Figura 4.4: Grupos sobrepostos em uma variável contínua

As possíveis estruturas dos grupos listadas no início da seção 4.1.4 estão representadas nas Figuras 4.5 a 4.8. Em cada gráfico há 100 observações, sendo 50 de cada grupo de uma variável categórica e duas contínuas (B_1 e B_2). Duas são as categorias do primeiro tipo de atributo e elas são identificadas pelos símbolos "bola" e "estrela". Analisando-se estas figuras é possível entender como se dá a separação e a sobreposição em um e em outro tipo de atributo. Ressalta-se que o foco desta dissertação será a estrutura "sobrepostos em ambas variáveis" (Figura 4.8) e que a sobreposição das variáveis contínuas ocorre entre todos os atributos desse tipo enquanto que a sobreposição das variáveis categóricas ocorre entre, no máximo, os três primeiros de quatro atributos.

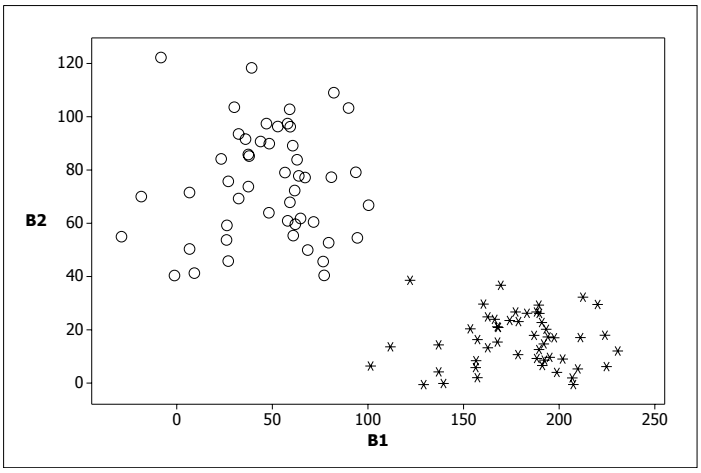

Figura 4.5: Grupos separados em ambas variáveis

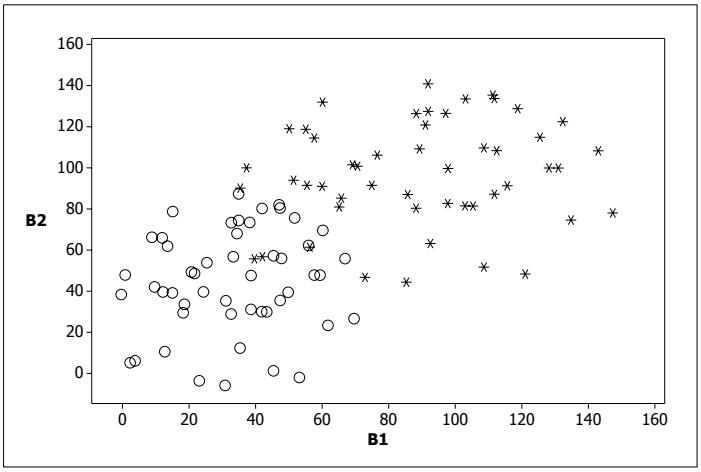

Figura 4.6: Grupos separados nas categóricas e sobrepostos nas contínuas

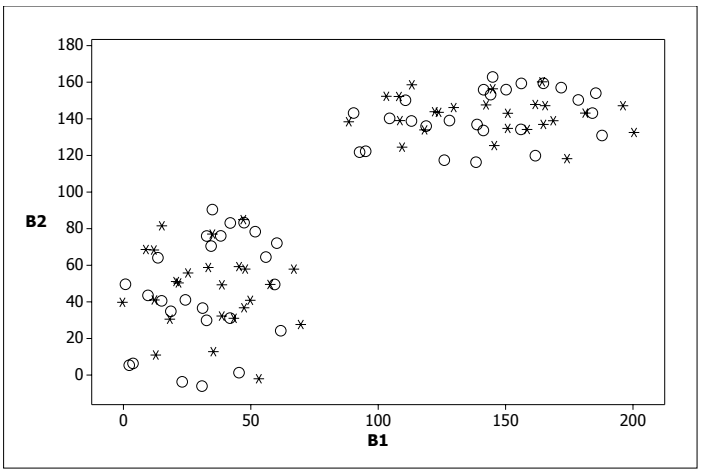

Figura 4.7: Grupos sobrepostos nas categóricas e separados nas contínuas

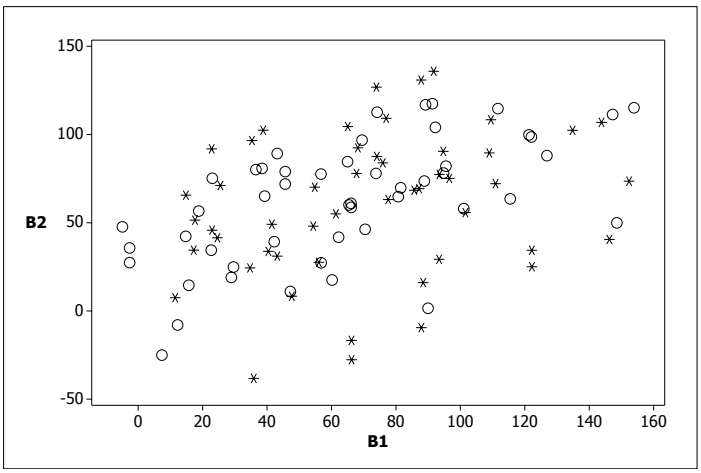

Figura 4.8: Grupos sobrepostos em ambas variáveis

4.1.4.3 Algumas Definições

Cada situação da Tabela 4.9 determina o número de grupos (k), de variáveis categóricas ($m = \frac{p}{2}$) e de categorias dessas variáveis ($|DOM(A_j)|$). Quanto à parte contínua, optou-se por fixar a quantidade de variáveis desse tipo como sendo o mesmo número de categóricas ($p - m = m$). Em relação à correlação, duas estruturas foram usadas. Na primeira, as variáveis eram independentes, ou seja, a matriz de correlação das variáveis contínuas era igual à Identidade. Na segunda, essa matriz era gerada aleatoriamente de uma distribuição Uniforme $(0, 1)$. No primeiro caso todos os grupos possuíam a mesma matriz. Já no segundo, as matrizes eram diferentes para cada um deles.

Com respeito aos graus de sobreposição das variáveis contínuas, três estruturas foram usadas. Elas foram obtidas atribuindo-se diferentes valores para a taxa de interseção dos grupos (int). Foram gerados grupos com 20%, 40% e 80% de sobreposição.

Finalmente, os pesos determinam como as medidas de proximidade das variáveis categóricas (medida de similaridade ponderada e coeficiente de discordância simples) e contínuas (distância euclidiana transformada em similaridade) devem ser combinadas (Equação (3.4)). E é com base nesse resultado que o agrupamento é realizado. Estudaram-se três estruturas de pesos. Na primeira foram dados pesos iguais às medidas de cada uma das variáveis ($w_{cat} = w_{cont} = 0,5$) e nas outras duas, ora as categóricas tiveram mais peso ($w_{cat} = 0,75$ e $w_{cont} = 0,75$), ora as contínuas tiveram peso maior ($w_{cat} = 0,25$ e $w_{cont} = 0,75$).

4.1.4.4 Resultados

A análise do grau de sobreposição, dos números de grupos e de variáveis, da correlação e dos pesos é realizada com base em tabelas-resumo, apresentadas no corpo do texto. As tabelas em anexo apresentam os resultados dos métodos comparados segundo situações, sobreposição das variáveis categóricas e contínuas, estruturas de correlação e pesos para cada um dos modelos.

Dependendo do aspecto de interesse, os demais fatores são desconsiderados, fazendo com que as tabelas-resumo retratem somente o ponto analisado. Por exemplo, as Tabelas 4.10 e 4.11 relacionadas com a análise do grau de sobreposição independem do número de grupos e de variáveis, da correlação entre elas e da estrutura de pesos. Assim como a tabela do número de grupos e de variáveis (4.12), que independe do grau de sobreposição, da correlação entre as variáveis contínuas e da estrutura de pesos.

4.1.4.4.1 Grau de sobreposição .

O grau de sobreposição pode ser estudado focando-se nas variáveis categóricas ou nas contínuas ou em ambas. Dado que o objetivo dessa análise é comparar o desempenho das metodologias na presença dos dois tipos de atributos, serão levados em consideração os dois graus de sobreposição conjuntamente.

Assim sendo, os dados das Tabelas 4.10 e 4.11 que devem ser observados são os que estão destacados em negrito. Eles mostram as taxas de alocação correta para os métodos Ligação Média, *ROCK* e k-Protótipos quando os graus de sobreposição das variáveis aumentam simultaneamente. O primeiro bloco de taxas corresponde ao grau de sobreposição 1 (e $m = 4$) para as variáveis categóricas e 20% de sobreposição nas contínuas, o segundo bloco, ao grau 2 e 40% e o terceiro e último bloco, ao grau 3 e 80%.

No modelo Beta (Tabela 4.10) as taxas de alocação diminuem com o aumento da sobreposição. O método *ROCK* sofre pequena queda de $1,11\%$ ao passar do bloco 1 para o 2. A sua taxa diminui de $89,35\%$ para $88,36\%$. Enquanto isso a Ligação Média tem queda de $36,58\%$ e o k-Protótipos, de $20,53\%$. Do bloco 2 para o 3, essas metodologias possuem reduções maiores ($52,93\%$ e $42,55\%$, respectivamente) e o *ROCK* acompanha esse comportamento ($10,51\%$). A Ligação Média e o *ROCK* são os métodos que mais e menos se prejudicam com o aumento da sobreposição, respectivamente.

Tabela 4.10: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas, para modelo Beta$(1; 0, 1)$

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas)			
		1 ($m = 2$)	1 ($m = 4$)	2	3
20%	Ligação Média	61,18	**79,30**	53,86	36,81
	ROCK	82,13	**89,35**	86,61	70,25
	k-Protótipos	61,93	**75,41**	63,20	50,24
40%	Ligação Média	59,07	73,82	**50,29**	32,97
	ROCK	83,59	89,90	**88,36**	73,08
	k-Protótipos	58,96	72,27	**59,93**	47,66
80%	Ligação Média	56,48	64,93	45,61	**23,67**
	ROCK	83,99	88,86	89,34	**79,07**
	k-Protótipos	49,69	57,36	47,54	**34,43**

Tabela 4.11: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas, para modelo Uniforme

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas) 1 ($m = 2$)	1 ($m = 4$)	2	3
20%	Ligação Média	54,35	**71,67**	52,47	35,81
	ROCK	76,65	**80,90**	76,94	63,76
	k-Protótipos	58,56	**66,07**	60,13	48,91
40%	Ligação Média	52,55	65,96	**49,25**	32,05
	ROCK	78,48	81,75	**79,56**	66,66
	k-Protótipos	55,41	62,34	**56,90**	45,82
80%	Ligação Média	51,26	59,80	45,24	**23,16**
	ROCK	79,40	82,24	81,89	**72,70**
	k-Protótipos	46,06	47,74	45,29	**32,95**

No modelo Uniforme (Tabela 4.11) as taxas também diminuem. O *ROCK* novamente destaca-se pelo fato das suas taxas sofrerem poucas alterações. As suas taxas de alocação são $80,90\%$, $79,56\%$ e $72,80\%$ nos blocos 1, 2 e 3. As demais técnicas continuam tendo as suas taxas muito reduzidas com o aumento da sobreposição. Analogamente ao modelo Beta, a Ligação Média tem maiores quedas que o k-Protótipos. As taxas desse primeiro método são $71,67\%$, $49,25\%$ e $23,16\%$ e as do segundo, $66,07\%$, $56,90\%$ e $32,95\%$.

4.1.4.4.2 Número de grupos e de variáveis .

A partir da Tabela 4.12, referente ao desempenho dos métodos com enfoque no número de grupos e de variáveis, conclui-se que o aumento do número de grupos diminui a taxa de alocação correta dos métodos comparados e o aumento do número de variáveis, aumenta essa taxa.

O efeito do número de grupos é esperado uma vez que existência de mais grupos dificulta a identificação deles pelas metodologias. Observa-se que todas elas sofrem esse efeito, independente do modelo. Porém, a redução do desempenho não acontece equivalentemente entre elas.

Tabela 4.12: Taxa de alocação correta dos métodos segundo número de grupos e de variáveis, para modelos Beta(1;0,1) e Uniforme

	Grupos			Média	Variáveis	
Método	$k=2$	$k=3$	$k=5$	global	$p=4$	$p=8$
Modelo Beta						
Ligação Média	**70,76**	60,47	45,50	57,57	58,91	**72,68**
ROCK	**93,28**	88,60	77,65	85,63	83,24	**89,37**
k-Protótipos	**68,34**	61,59	50,16	59,04	56,86	**68,35**
Modelo Uniforme						
Ligação Média	**65,51**	56,20	41,94	53,29	52,72	**65,81**
ROCK	**90,27**	80,59	68,24	78,55	78,18	**81,63**
k-Protótipos	**65,26**	55,03	45,05	54,10	53,34	**58,72**

No modelo Beta, a Ligação Média é o método que tem maior redução ($24,76\%$) da taxa de alocação correta, passa de $60,47\%$ para $45,50\%$ quando o número de grupos aumenta de $k=3$ para $k=5$. O *ROCK* é o que tem menor redução ($5,02\%$), passando de $93,28\%$ para $88,60\%$ quando o número de grupos aumenta de $k=2$ para $k=3$.

No modelo Uniforme tanto a Ligação Média quanto o k-Protótipos têm grandes reduções. O *ROCK*, de novo, destaca-se como o menos afetado, mas os percentuais de diminuição ($10,72\%$ e $15,32\%$) são agora mais próximos dos percentuais das demais técnicas ($14,21\%$ e $25,37\%$ para a Ligação Média e $15,68\%$ e $18,14\%$ para o k-Protótipos).

O efeito do número de variáveis bastante se assemelha às conclusões da comparação quando somente variáveis categóricas estavam presentes nos bancos de dados. O aumento do número de atributos contribui para que os grupos sejam mais facilmente identificados. Acredita-se que esse efeito está relacionado com as variáveis categóricas, dado que algumas delas determinam a sobreposição enquanto outras não, sendo geradas aleatoriamente de acordo com o modelo selecionado.

Ainda com base na Tabela 4.12 nota-se que em ambos os modelos, a Ligação Média é o método que mais se beneficia com o acréscimo das novas variáveis. O *ROCK* possui resultados próximos quando $p=4$ e quando $p=8$. No modelo Uniforme, por exemplo, a taxa de alocação correta desse método é $78,18\%$ no primeiro caso e $81,63\%$ no segundo.

Além da tabela já apresentada, as Tabelas B.1 a B.12 em anexo permitem também que seja feita a análise do número de grupos e de variáveis. O desempenho dos métodos em relação ao número de grupos pode ser estudado comparando-se as situações dentro

dos graus de sobreposição das variáveis categóricas. E em relação ao número de variáveis, comparando-se as situações do "grau de sobreposição 1 e $m = 2$" com "grau de sobreposição 1 e $m = 4$". É importante nessas análises manter fixo o percentual de sobreposição nas contínuas.

Independente desses fatores, o *ROCK* destaca-se como sendo o método mais eficiente para agrupar observações de variáveis categóricas e contínuas. As suas taxas de alocação são, em geral, superiores às taxas dos demais métodos. Quando elas não são as maiores, estão bem próximas do máximo. Como por exemplo em (os melhores desempenhos estão entre parênteses):

1. Modelo Beta

   - sem correlação e grau de sobreposição 20%, situação 17 quando os pesos são iguais e situações 12, 17 e 25 quando as contínuas têm mais peso (Ligação Média) (Tabelas B.1 e B.3);
   - com correlação e grau de sobreposição 20%, situação 17 quando os pesos são iguais e situações 17 e 25 quando contínuas têm mais peso (Ligação Média) (Tabelas B.4 e B.6).

2. Modelo Uniforme

   - sem correlação e grau de sobreposição 20%, situação 17 quando os pesos são iguais e quando as contínuas têm mais peso (Ligação Média) (Tabelas B.7 e B.9);
   - sem correlação, situação 7 quando as categóricas têm mais peso nos graus de sobreposição 20% e 40% (k-Protótipos) (Tabela B.8);
   - com correlação e grau de sobreposição 20%, situação 7 quando as categóricas têm mais peso (k-Protótipos) e situação 17 quando as contínuas têm mais peso (Ligação Média) (Tabelas B.11 e B.12).

4.1.4.4.3 Correlação .

Duas estruturas de correlação foram estudadas. Em uma delas as variáveis contínuas são independentes e na outra, a correlação entre elas foi determinada aleatoriamente, a partir de uma variável aleatória Uniforme $(0, 1)$ (página 84).

As Tabelas 4.13 e 4.14 trazem as taxas de alocação correta para os métodos comparados segundo os graus de sobreposição e as estruturas de correlação impostas às variáveis

contínuas. Percebe-se que em ambos os modelos não há diferença entre as taxas de alocação quando existe ou não correlação. Esse aspecto, portanto, não influencia o desempenho de nenhum método em nenhuma das estruturas estudadas. Isso já havia sido evidenciado no caso contínuo, segundo o artigo de Mingoti e Lima (2006).

Tabela 4.13: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas e estruturas de correlação, para modelo Beta$(1; 0, 1)$

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas)			
		1 ($m = 2$)	1 ($m = 4$)	2	3
Contínuas independentes					
20%	Ligação média	61,12	**79,89**	**54,11**	**37,04**
	ROCK	82,03	89,32	86,33	69,96
	k-Protótipos	**62,11**	**75,69**	**63,39**	**50,32**
40%	Ligação média	**59,18**	**74,59**	**50,46**	**33,06**
	ROCK	83,53	**90,07**	88,15	72,70
	k-Protótipos	**59,13**	**72,73**	**60,26**	**47,91**
80%	Ligação média	**56,63**	**64,95**	45,54	22,94
	ROCK	**84,48**	**90,00**	**90,15**	**80,29**
	k-Protótipos	**49,84**	**57,98**	**47,86**	**34,71**
Contínuas correlacionadas					
20%	Ligação média	**61,24**	78,71	53,61	36,59
	ROCK	**82,23**	**89,37**	**86,89**	**70,55**
	k-Protótipos	61,76	75,13	63,02	50,15
40%	Ligação média	58,96	73,04	50,12	32,88
	ROCK	**83,65**	89,74	**88,56**	**73,45**
	k-Protótipos	58,80	71,80	59,60	47,41
80%	Ligação média	56,33	64,92	**45,67**	**24,41**
	ROCK	83,51	87,72	88,52	77,85
	k-Protótipos	49,54	56,74	47,23	34,14

Tabela 4.14: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas e estruturas de correlação, para modelo Uniforme

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas) 1 ($m = 2$)	1 ($m = 4$)	2	3
Contínuas independentes					
20%	Ligação média	54,32	**72,28**	**52,71**	**36,20**
	ROCK	76,59	80,83	76,35	63,33
	k-Protótipos	**58,66**	**66,38**	**60,39**	**49,11**
40%	Ligação média	**52,60**	**66,37**	**49,32**	**32,21**
	ROCK	78,47	**81,81**	79,22	66,31
	k-Protótipos	**55,61**	**62,87**	**57,27**	**46,17**
80%	Ligação média	**51,37**	59,77	45,17	22,44
	ROCK	**79,70**	**83,32**	**82,51**	**73,95**
	k-Protótipos	**46,25**	**48,22**	**45,61**	**33,03**
Contínuas correlacionadas					
20%	Ligação média	**54,38**	71,06	52,24	35,43
	ROCK	**76,71**	**80,98**	**77,54**	**64,20**
	k-Protótipos	58,46	65,75	59,86	48,72
40%	Ligação média	52,50	65,54	49,19	31,89
	ROCK	**78,50**	81,69	**79,89**	**67,02**
	k-Protótipos	55,20	61,81	56,53	45,47
80%	Ligação média	51,15	**59,82**	**45,32**	**23,88**
	ROCK	79,10	81,15	81,26	71,44
	k-Protótipos	45,87	47,25	44,96	32,87

A Tabela 4.15 agrega todos os resultados das simulações e a partir dela ratifica-se a análise quanto à correlação, pois as taxas em um contexto e no outro são bastante próximas.

Tabela 4.15: Taxa de alocação correta dos métodos segundo estruturas de correlação, para modelos Beta$(1; 0, 1)$ e Uniforme

	Variáveis contínuas	
Métodos	Independentes	Correlacionadas
Modelo Beta		
Ligação média	**57,75**	57,39
ROCK	**85,84**	85,41
k-Protótipos	**59,33**	58,75
Modelo Uniforme		
Ligação Média	**53,42**	53,16
ROCK	**78,67**	78,44
k-Protótipos	**54,40**	53,81

4.1.4.4.4 Pesos .

As Tabelas 4.16 e 4.17 apresentam as taxas de alocação correta dos métodos Ligação Média, *ROCK* e k-Protótipos segundo os graus de sobreposição e as estruturas de pesos.

Tanto para o modelo Beta (Tabela 4.16) quanto para o Uniforme (Tabela 4.17), as menores taxas de alocação correta foram obtidas quando o maior peso era dado às variáveis categóricas. Por outro lado, maiores taxas estão associadas ao maior peso dado às variáveis contínuas. Os resultados referentes aos pesos iguais são intermediários a esses dois.

Quando os grupos são sobrepostos em 80% nas variáveis contínuas, o *ROCK* tem melhores resultados quando o peso é igual (acima de 93%). Percebe-se que ao incorporar as informações das variáveis categóricas, essa técnica passa a ter desempenho melhor, independente do grau de sobreposição delas. Quando os grupos são sobrepostos em três de quatro variáveis categóricas, a Ligação Média tem os piores resultados para pesos iguais e para peso maior para variáveis categóricas (20%).

Tabela 4.16: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas e estruturas de pesos, para modelo Beta$(1; 0, 1)$

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas) 1 ($m = 2$)	1 ($m = 4$)	2	3
Peso igual					
20%	Ligação Média	53,77	83,40	36,64	20,00
	ROCK	86,42	88,67	85,58	72,42
	k-Protótipos	54,80	75,52	55,76	36,66
40%	Ligação Média	52,70	75,44	35,27	20,00
	ROCK	90,10	89,80	88,86	76,99
	k-Protótipos	52,85	70,48	51,43	35,92
80%	Ligação Média	50,64	65,63	35,23	20,00
	ROCK	**96,54**	**93,96**	**96,15**	**96,62**
	k-Protótipos	47,98	55,71	43,62	32,91
Peso maior para variáveis categóricas					
20%	Ligação Média	52,19	60,61	34,88	20,00
	ROCK	68,08	86,68	81,31	51,76
	k-Protótipos	53,80	59,33	45,48	33,55
40%	Ligação Média	52,00	59,79	34,88	20,00
	ROCK	67,98	86,64	82,24	52,39
	k-Protótipos	52,82	58,49	45,08	33,72
80%	Ligação Média	50,42	57,36	34,88	20,00
	ROCK	68,01	86,67	82,18	52,36
	k-Protótipos	47,95	54,37	43,30	32,48
Peso maior para variáveis contínuas					
20%	Ligação Média	**77,57**	**93,87**	**90,05**	**70,44**
	ROCK	**91,89**	**92,70**	**92,93**	**86,58**
	k-Protótipos	**77,21**	**91,39**	**88,36**	**80,50**
40%	Ligação Média	**72,51**	**86,22**	**80,72**	**58,91**
	ROCK	**92,69**	**93,27**	**93,98**	**89,85**
	k-Protótipos	**71,22**	**87,83**	**83,29**	**73,35**
80%	Ligação Média	**68,38**	**71,80**	**66,71**	**31,01**
	ROCK	87,43	85,95	89,67	88,23
	k-Protótipos	**53,13**	**62,00**	**55,71**	**37,88**

Tabela 4.17: Taxa de alocação correta dos métodos segundo sobreposição nas variáveis categóricas e nas contínuas e estruturas de pesos, para modelo Uniforme

Percentual de sobreposição (contínuas)	Métodos	Grau de sobreposição (categóricas) 1 ($m = 2$)	1 ($m = 4$)	2	3
Peso igual					
20%	Ligação média	44,48	75,99	35,20	20,00
	ROCK	84,91	85,81	83,83	71,45
	k-Protótipos	51,39	60,68	49,39	35,68
40%	Ligação Média	43,55	67,41	35,04	20,00
	ROCK	89,80	87,94	88,29	76,54
	k-Protótipos	49,48	55,68	46,70	34,76
80%	Ligação Média	43,08	61,62	35,11	20,00
	ROCK	**96,76**	**95,11**	**96,42**	**96,82**
	k-Protótipos	44,24	44,38	42,21	31,63
Peso maior para variáveis categóricas					
20%	Ligação Média	42,52	46,89	34,89	20,00
	ROCK	52,76	63,76	53,75	33,65
	k-Protótipos	50,39	48,10	43,99	32,13
40%	Ligação Média	42,52	46,76	34,88	20,00
	ROCK	52,61	63,59	56,27	33,46
	k-Protótipos	49,19	47,35	43,53	32,24
80%	Ligação Média	42,56	46,54	34,89	20,00
	ROCK	52,62	63,96	58,57	33,65
	k-Protótipos	44,26	43,62	41,85	31,08
Peso maior para variáveis contínuas					
20%	Ligação Média	**76,05**	**92,12**	**87,33**	**67,44**
	ROCK	**92,28**	**93,13**	**93,25**	**86,18**
	k-Protótipos	**73,90**	**89,42**	**87,00**	**78,93**
40%	Ligação Média	**71,58**	**83,70**	**77,85**	**56,16**
	ROCK	**93,04**	**93,74**	**94,12**	**90,00**
	k-Protótipos	**67,56**	**84,00**	**80,47**	**70,46**
80%	Ligação Média	**68,14**	**71,23**	**65,73**	**29,47**
	ROCK	88,81	87,64	90,68	87,63
	k-Protótipos	**49,66**	**55,21**	**51,80**	**36,13**

No modelo Beta (Tabela 4.16), o pior resultado do *ROCK* para grupos sobrepostos na primeira de quatro variáveis categóricas e 80% nas contínuas, é obtido quando é dado maior peso para as variáveis contínuas $(85, 95\%)$. Essa taxa de alocação correta não é muito maior do que a encontrada quando é dado maior peso para as variáveis categóricas $(86, 67\%)$.

No modelo Uniforme (Tabela 4.17), o pior resultado do k-Protótipos para grupos sobrepostos na primeira de duas variáveis categóricas e 80% nas contínuas, é obtido quando os pesos são iguais $(44, 24\%)$. De novo, esse resultado não é muito maior do que aquele encontrado quando o peso maior é dado para as variáveis categóricas $(44, 26\%)$.

Em média (Tabela 4.18), confirma-se a análise de que melhores e piores desempenhos estão relacionados ao maior peso para variáveis contínuas e categóricas, respectivamente. E que os pesos iguais proporcionam taxas de alocação nem tão boas nem tão ruins.

Tabela 4.18: Taxa de alocação correta dos métodos segundo estruturas de pesos, para modelos Beta$(1; 0, 1)$ e Uniforme

Métodos	Peso		
	Igual	Maior para categóricas	Maior para contínuas
Modelo Beta			
Ligação Média	50,87	45,70	**76,13**
ROCK	89,81	76,20	**90,87**
k-Protótipos	54,33	49,39	**73,41**
Modelo Uniforme			
Ligação Média	46,15	39,24	**74,48**
ROCK	89,05	55,15	**91,46**
k-Protótipos	47,82	44,41	**70,09**

4.2 Estudos de Aplicação

As aplicações foram selecionadas dentre os exemplos encontrados nos artigos de referência da dissertação e dentre os bancos disponíveis no repositório *UCI Machine Learning Repository* (http://www.ics.uci.edu/ mlearn/MLRepository.html). Todos os métodos avaliados foram aplicados aos dados e, ao final, comparou-se os grupos finais com os originalmente existentes nas bases. A identificação do grupo ao qual pertencia a observação foi desconsiderada durante o agrupamento. Assim como as observações que continham valores faltantes (*missings*).

Três bancos de dados foram estudados. Dois exclusivamente com variáveis categóricas e um com variáveis dos dois tipos. Cada uma das aplicações será apresentada a seguir, seguindo o seguinte esquema: descrição das informações e resultados. Esses últimos serão comentados com base na taxa de alocação correta (seção 4.1.2 na página 50).

Na descrição apresentam-se as variáveis, cujos nomes optou-se por não traduzir. Bem como os nomes das suas categorias. Isso não traz prejuízo ao trabalho uma vez que o objetivo não é justificar a composição dos grupos, mas o desempenho dos métodos na alocação das observações. Por outro lado, entende-se ser necessário a realização de uma breve análise descritiva das informações no sentido de compreender a disposição dos grupos existentes.

Pelo fato de algumas metodologias serem dependentes da escolha inicial das sementes, realizou-se a análise de agrupamentos 100 vezes em cada um dos bancos, mudando-se a ordenação das observações. A taxa de alocação correta a ser apresentada é a média das taxas obtidas em cada uma das vezes. Esse é um procedimento comum na literatura.

Estudos de aplicação foram feitos com o objetivo de ilustrar a utilização das metodologias de agrupamentos selecionadas para esta dissertação. A avaliação do desempenho dessas técnicas fica um pouco comprometida pelo fato de a partição original dos dados, que é aquela proveniente de alguma variável, é de certa forma subjetiva. Isso ficará mais claro na aplicação Hepatite (seção 4.2.3).

4.2.1 Doença do Grão de Soja

Essas informações são foco de grande maioria dos artigos encontrados na literatura (HUANG, 1998; HUANG; NG, 1999; KIM *et al.*, 2004; SAN *et al.*, 2004; KIM *et al.*, 2005). Elas dizem respeito a 47 grãos de soja divididas em 4 grupos de doenças: *Diaporthe Stem*

Canker, *Charcoal Rot*, *Rhizoctonia Root Rot* e *Phytophthora Rot*, sendo que o último grupo tem 17 observações e os demais, 10. Dado que não ocorrem valores faltantes, todos os grãos puderam ser usados.

Estão disponíveis no banco de dados original 35 variáveis categóricas que representam características locais (*plant local variables*) e globais (*plant global variables*) das plantas e descrevem também as condições ambientais (*environment variables*). Desses 35 atributos, somente 21 foram usados para o agrupamento uma vez que os demais (14) continham apenas uma categoria. O tamanho do domínio das 21 variáveis varia entre 2 e 7. Foi necessário alterar o domínio de 3 variáveis de 3 para 2. Isso porque nenhuma observação continha uma das categorias dessas variáveis.

Com um número de grupos igual a 4, somente através da análise descritiva é difícil perceber padrões de ocorrência de categorias nas variáveis estudadas. É possível notar alguns pontos que chamam a atenção na Tabela 4.19. Observa-se que em 15 das 21 variáveis, entre elas na A_{15}, todos os objetos de determinado(s) grupo(s) contêm a mesma categoria. As exceções são os atributos A_1, A_5, A_6, A_9, A_{10} e A_{12}.

Tabela 4.19: Distribuição de freqüências das variáveis categóricas do estudo sobre a doença do grão de soja

Variável	Categoria	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
A_1	1	.	.	5	5	10
	2	.	.	.	5	5
	3	.	.	3	4	7
	4	3	2	1	3	9
	5	2	2	1	.	5
	6	2	3	.	.	5
	7	3	3	.	.	6
A_2	1	10	10	2	.	22
	2	.	.	8	17	25
A_3	1	.	10	.	.	10
	2	.	.	.	4	4
	3	10	.	10	13	33
A_4	1	.	.	10	7	17
	2	10	4	.	10	24
	3	.	6	.	.	6

continua na próxima página

Tabela 4.19 – continuação da página anterior

Variável	Categoria	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
A_5	1	9	4	8	11	32
	2	1	6	2	6	15
A_6	1	.	2	3	2	7
	2	4	3	2	5	14
	3	3	2	2	4	11
	4	3	3	3	6	15
A_7	1	7	.	.	.	7
	2	3	.	10	16	29
	3	.	5	.	.	5
	4	.	5	.	1	6
A_8	1	7	10	5	6	28
	2	3	.	5	11	19
A_9	1	5	5	6	8	24
	2	5	5	4	9	23
A_{10}	1	2	4	.	7	13
	2	3	3	5	4	15
	3	5	3	5	6	19
A_{11}	1	.	.	9	.	9
	2	10	10	1	17	38
A_{12}	1	6	7	8	15	36
	2	4	3	2	2	11
A_{13}	1	.	10	.	.	10
	2	.	.	10	8	18
	3	.	.	.	9	9
	4	10	.	.	.	10
A_{14}	1	4	.	.	.	4
	2	6	.	10	.	16
	3	.	.	.	17	17
	4	.	10	.	.	10
A_{15}	1	.	10	10	17	37
	2	10	.	.	.	10

continua na próxima página

Tabela 4.19 – continuação da página anterior

Variável	Categoria	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Total
A_{16}	1	.	10	.	11	21
	2	10	.	10	6	26
A_{17}	1	10	10	5	17	42
	2	.	.	5	.	5
A_{18}	1	10	.	10	17	37
	2	.	10	.	.	10
A_{19}	1	10	.	10	17	37
	2	.	10	.	.	10
A_{20}	1	10	10	.	.	20
	2	.	.	10	17	27
A_{21}	1	10	10	9	.	29
	2	.	.	1	17	18

Outro ponto que se destaca na Tabela 4.19 é o fato de algumas categorias serem exclusivas de alguns grupos. Isto é, todos ou a maioria dos grãos de um mesmo grupo possuem uma determinada característica enquanto que nenhum outro grão, de outro grupo, também a possui. Isso é interessante porque contribui para a melhor identificação desse grupo. Se isso fosse observado mais vezes, poderia-se dizer que os grupos são mais facilmente identificados pelos métodos de agrupamentos analisados. A ocorrência se dá em 9 das 21 variáveis: a categoria 1 da variável A_3 está relacionada ao grupo 2, a categoria 3 de A_4 ao grupo 2, as categorias 1 e 3 de A_7 aos grupos 1 e 2 (nessa ordem), as categorias 1 e 4 de A_{13} aos grupos 2 e 1 (nessa ordem), as categorias 3 e 4 de A_{14} aos grupos 4 e 2 (nessa ordem), a categoria 2 de A_{15} ao grupo 1, a categoria 2 de A_{17} ao grupo 3 e, finalmente, a categoria 2 de A_{18} e A_{19} ao grupo 2.

Nenhum padrão diferenciado pôde ser notado nas demais variáveis. Frente à análise feita, pode-se concluir que se somente os atributos listados na último parágrafo fossem usados no agrupamento dos grãos, talvez os grupos seriam mais facilmente identificados. Mas uma vez que todas as informações são utilizadas, e isso inclui aquelas variáveis em que as observações de todos os grupos possuem as mesmas características, os grupos podem ainda ser encontrados, mas os métodos podem não ter um bom desempenho.

A Tabela 4.20 apresenta as taxas de alocação correta dos métodos comparados nesta aplicação:

Tabela 4.20: Taxa de alocação correta do estudo sobre a doença do grão de soja

Método	Taxa de alocação correta
Ligação Média	40,43
$ROCK$ $(\theta = 0,30)$	83,83
$ROCK$ $(\theta = 0,70)$	56,38
k-Modas	43,09
Fuzzy c-Modas	43,09
k-Populações	41,28

De forma geral, as metodologias realmente tiveram alguma dificuldade em classificar corretamente os grãos nos grupos de doenças. Com exceção do $ROCK$, a taxa de alocação correta dos métodos comparados fica entre 40% e 44%. O método $ROCK$ tem desempenho bastante superior comparando-o com os demais quando $\theta = 0,30$ $(83,83\%)$. Para $\theta = 0,70$ o resultado não é tão bom, apesar da taxa ainda ser superior a das demais técnicas $(56,38\%)$. Com isso, confirma-se o bom desempenho do $ROCK$ a existência de forte dependência entre o esse método e o seu parâmetro.

4.2.2 Votos de Congressistas

O banco de dados original (GUHA *et al.*, 2000) dessa aplicação contém 435 congressistas americanos classificados de acordo com a sua posição política: democrata ou republicana. O primeiro grupo continha 267 indivíduos e o segundo, 168. Devido à existência de valores faltantes, o banco foi reduzido a 232 observações, sendo 124 congressistas democratas (grupo 1) e 108 republicanos (grupo 2).

Na literatura, o objetivo dessa análise é traçar o perfil de voto de um e do outro grupo, a partir de 16 questões. Os congressistas as avaliaram e votaram contra ou a favor do assunto tratado. Assim, todas as variáveis contêm apenas duas categorias.

Tabela 4.21: Distribuição de freqüências das variáveis categóricas do estudo sobre os votos de congressistas (percentual em relação ao total de cada grupo)

Variável	Categoria	Democratas	Republicanos	Total
A_1	1	51 (41,13%)	85 (78,70%)	136
	2	73 (58,87%)	23 (21,30%)	96
A_2	1	68 (54,84%)	57 (52,78%)	125
	2	56 (45,16%)	51 (47,22%)	107
A_3	1	18 (14,52%)	91 (84,26%)	109
	2	106 (85,48%)	17 (15,74%)	123
A_4	1	118 (95,16%)	1 (0,93%)	119
	2	6 (4,84%)	107 (99,07%)	113
A_5	1	99 (79,84%)	5 (4,63%)	104
	2	25 (20,16%)	103 (95,37%)	128
A_6	1	69 (55,65%)	14 (12,96%)	83
	2	55 (44,35%)	94 (87,04%)	149
A_7	1	29 (23,39%)	79 (73,15%)	108
	2	95 (76,61%)	29 (26,85%)	124
A_8	1	21 (16,94%)	92 (85,19%)	113
	2	103 (83,06%)	16 (14,81%)	119
A_9	1	26 (20,97%)	93 (86,11%)	119
	2	98 (79,03%)	15 (13,89%)	113
A_{10}	1	58 (46,77%)	46 (42,59%)	104
	2	66 (53,23%)	62 (57,41%)	128
A_{11}	1	61 (49,19%)	91 (84,26%)	152
	2	63 (50,81%)	17 (15,74%)	80
A_{12}	1	108 (87,10%)	16 (14,81%)	124
	2	16 (12,90%)	92 (85,19%)	108
A_{13}	1	88 (70,97%)	17 (15,74%)	105
	2	36 (29,03%)	91 (84,26%)	127
A_{14}	1	81 (65,32%)	2 (1,85%)	83
	2	43 (34,68%)	106 (98,15%)	149
A_{15}	1	50 (40,32%)	96 (88,89%)	146
	2	74 (59,68%)	12 (11,11%)	86
A_{16}	1	7 (5,65%)	36 (33,33%)	43
	2	117 (94,35%)	72 (66,67%)	189

Pela Tabela 4.21 nota-se que somente em 3 das 16 questões (A_2, A_{10} e A_{16}), tanto republicanos quanto democratas deram os mesmos votos. Em 12 das 13 questões restantes a maioria dos democratas votaram diferentemente da maioria dos republicanos. Na questão restante (variável A_{11}), o primeiro grupo está igualmente dividido entre as duas opções de voto enquanto que a maioria dos republicanos optaram pela categoria 1.

Dessa forma os grupos de congressistas parecem ter perfis diferentes na avaliação dos assuntos votados e assim sendo os grupos "republicanos" e "democratas" estariam bem definidos, fazendo com que os métodos de análise de agrupamentos possam encontrá-los sem grandes dificuldades.

Tabela 4.22: Taxa de alocação correta do estudo sobre os votos de congressistas

Método	Taxa de alocação correta
Ligação Média	76,31
ROCK ($\theta = 0,30$)	68,53
ROCK ($\theta = 0,70$)	99,61
k-Modas	53,56
Fuzzy c-Modas	53,13
k-Populações	53,34

De fato, os métodos tiveram melhor desempenho nessa aplicação (Tabela 4.22) do que na anterior (Tabela 4.20 da página 99). A Ligação Média destaca-se com taxa de alocação correta de $76,31\%$. Em seguida tem-se o *ROCK* com taxa de $68,53\%$ e $99,61\%$ para $\theta = 0,30$ e $\theta = 0,70$, respectivamente. Por fim, as demais metodologias, que têm desempenhos bem menores e próximos. As suas taxas estão em torno de 53%.

De novo percebe-se a relação entre a taxa de alocação correta do *ROCK* e o seu parâmetro θ. Nesse estudo de aplicação o valor $0,70$ é o que fornece melhor resultado, diferentemente do exemplo anterior em que $0,30$ era o valor que produzia a maior taxa de alocação correta. Isso pode ser uma indicação de que valores de θ próximos a $0,70$ sejam melhores quando somente variáveis binárias estão presentes no banco de dados. Estudos mais detalhados precisam ser desenvolvidos para que resultados mais sólidos possam ser encontrados.

4.2.3 Hepatite

O banco de dados *Hepatitis* (ANDREOPOULOS *et al.*, 2006) contém 13 variáveis categóricas e 6 contínuas. As observações são pacientes, classificados nos grupos sobrevi-

ventes e não sobreviventes, cada um com 123 e 32 indivíduos, respectivamente, totalizando 155 observações. As variáveis categóricas representam algo observado nos pacientes enquanto que as contínuas, resultados de testes. O domínio de todas as primeiras variáveis é 2 e a escala de mensuração das segundas é contínua, com exceção da idade, que é discreta. Devido aos valores faltantes, a base foi reduzida a 80 pacientes, sendo 13 do grupo de não sobreviventes (Grupo 1) e 67 do grupo de sobreviventes (Grupo 2). A identificação do grupo (desfecho médico) foi desconsiderada durante o agrupamento.

Destaca-se na Tabela 4.23 que para certos atributos existe coincidência entre as categorias mais freqüentes dos grupos 1 e 2. Por exemplo, todos os 13 indivíduos do grupo 1 possuem as categorias 1, 2 e 2 para as variáveis A_1, A_6 e A_7. Para esses atributos, a maioria das observações do grupo 2 também estão concentradas nessas categorias (respectivamente, 56, 55 e 54 das 67 observações). O mesmo ocorre com outras quatro variáveis. São estas: A_3, A_4, A_9 e A_{12}.

Tabela 4.23: Distribuição de freqüências das variáveis categóricas do estudo sobre a hepatite (percentual em relação ao total do grupo)

Variável	Categoria	Grupo 1	Grupo 2	Total
A_1	1	13 (100,00%)	56 (83,58%)	69
	2	. (0,00%)	11 (16,42%)	11
A_2	1	8 (61,54%)	30 (44,78%)	38
	2	5 (38,46%)	37 (55,22%)	42
A_3	1	2 (15,38%)	19 (28,36%)	21
	2	11 (84,62%)	48 (71,64%)	59
A_4	1	11 (84,62%)	41 (61,19%)	52
	2	2 (15,38%)	26 (38,81%)	28
A_5	1	9 (69,23%)	22 (32,84%)	31
	2	4 (30,77%)	45 (67,16%)	49
A_6	1	. (0,00%)	12 (17,91%)	12
	2	13 (100,00%)	55 (82,09%)	68
A_7	1	. (0,00%)	13 (19,40%)	13
	2	13 (100,00%)	54 (80,60%)	67
A_8	1	7 (53,85%)	31 (46,27%)	38
	2	6 (46,15%)	36 (53,73%)	42
A_9	1	4 (30,77%)	11 (16,42%)	15
	2	9 (69,23%)	56 (83,58%)	65

continua na próxima página

Tabela 4.23 – continuação da página anterior

Variável	Categoria	Grupo 1	Grupo 2	Total
A_{10}	1	8 (61,54%)	17 (25,37%)	25
	2	5 (38,46%)	50 (74,63%)	55
A_{11}	1	7 (53,85%)	5 (7,46%)	12
	2	6 (46,15%)	62 (92,54%)	68
A_{12}	1	5 (38,46%)	5 (7,46%)	10
	2	8 (61,54%)	62 (92,54%)	70
A_{13}	1	1 (7,69%)	46 (68,66%)	47
	2	12 (92,31%)	21 (31,34%)	33

Já em relação às demais, essa coincidência não ocorre. Por exemplo, em A_2, 8 dos 13 $(61,54\%)$ indivíduos do grupo 1 possuem a categoria 1, enquanto que 37 dos 67 (55,22%) do grupo 2 possuem a categoria 2. Isso também ocorre nos atributos A_5, A_{10} e A_{13}. Essa situação é mais interessante porque, conforme já comentado na aplicação anterior, contribui para que os grupos sejam mais facilmente identificados.

Em ambos os grupos não há uma categoria muito mais freqüente que outra para a variável A_8. O mesmo ocorre entre as observações do grupo 1 em relação à variável A_{11}.

Assim, ao utilizar todas as informações categóricas conjuntamente não se pode afirmar que os grupos podem ser diretamente discriminados. As variáveis A_2, A_5, A_{10} e A_{13} poderiam separar os indivíduos mais facilmente uma vez que a maior parte das observações de um grupo não possui a mesma característica que a maior parte das observações do outro grupo.

Em se tratando das variáveis contínuas, observa-se na Figura 4.9 e na Tabela 4.24 que os grupos têm padrões diferenciados. Em relação às variáveis B_1, B_4 e B_6, o grupo 1 é mais homogêneo que o 2. O inverso ocorre nas variáveis B_2, B_3 e B_5. Não fosse o valor discrepante (*outlier*) na distribuição de B_3, os padrões dos grupos seriam mais próximos nessa variável. Os demais atributos, com exceção da B_1, possuem *outliers* em um ou em outro grupo e isso, conforme já se sabe, influencia nas médias (Tabela 4.24). Nota-se que as observações do grupo 2 têm valores mais elevados para três das seis variáveis. A saber, B_4, B_5 e B_6.

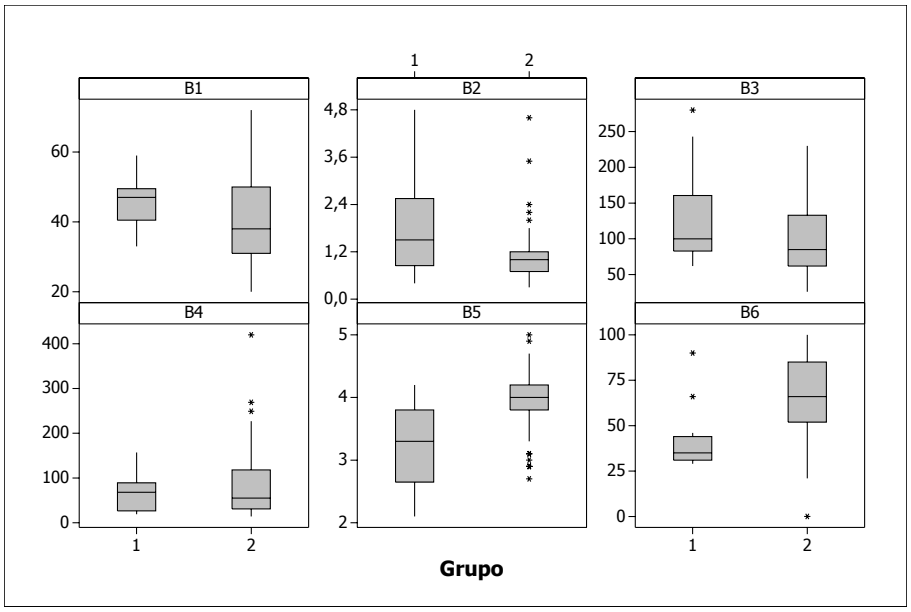

Figura 4.9: *Boxplots* das variáveis contínuas do estudo sobre a hepatite

Tabela 4.24: Média e desvio padrão (entre parênteses) das variáveis contínuas segundo grupo do estudo sobre a Hepatite

Variável	Grupo 1	Grupo 2
B_1	46,08 (7,24)	39,61 (11,66)
B_2	1,92 (1,42)	1,09 (0,66)
B_3	125,85 (68,27)	98,46 (49,78)
B_4	69,31 (46,21)	84,49 (75,57)
B_5	3,22 (0,64)	3,96 (0,48)
B_6	41,62 (17,65)	66,57 (22,31)

Os grupos nessa aplicação não estão tão bem separados e acredita-se que os métodos de agrupamentos terão alguma dificuldade em identificar os grupos de pacientes. A estrutura de pesos usada nesse exemplo é aquela dada pela Equação (3.4). Três conjuntos de valores foram escolhidos para as quantidades w_{cat} e w_{cont}, a saber: $\{0,50; 0,50\}$, $\{0,75; 0,25\}$ e $\{0,25; 0,75\}$. No primeiro, as medidas de proximidade de ambas variáveis têm o mesmo peso e nos dois outros conjuntos é dado peso maior para as variáveis categóricas e para as contínuas, respectivamente.

A Ligação Média é a metodologia que possui melhor resultado nessa aplicação (Tabela 4.25). As suas taxas de alocação, todas acima de 98%, são bem maiores que as taxas dos demais métodos. O *ROCK* tem taxa de 60%, para os dois valores de θ e as três configurações de pesos. O desempenho do k-Protótipos nessas situações também não muda muito. As suas taxas variam de $68,38\%$ (pesos iguais) a $69,88\%$ (peso maior para variáveis categóricas).

Tabela 4.25: Taxa de alocação correta do estudo sobre a hepatite

Método	Taxa de alocação correta
Peso igual	
Ligação Média	98,75
ROCK ($\theta = 0,30$)	60,00
ROCK ($\theta = 0,70$)	60,00
k-Protótipos	68,38
Peso maior para variáveis categóricas	
Ligação Média	98,88
ROCK ($\theta = 0,30$)	60,00
ROCK ($\theta = 0,70$)	60,00
k-Protótipos	69,88
Peso maior para variáveis contínuas	
Ligação Média	98,75
ROCK ($\theta = 0,30$)	60,00
ROCK ($\theta = 0,70$)	60,00
k-Protótipos	69,38

Nesse exemplo o *ROCK* não apresentou um desempenho tão bom quanto nos exemplos anteriores, uma vez que seus resultados nos estudos de aplicação quando somente variáveis categóricas estavam presentes nos bancos de dados e nos estudos de simulação envolvendo um único ou ambos os tipos de atributos foram melhores. Acreditando-se que estivesse havendo a influência de valores discrepantes em relação aos atributos contínuos. Fez-se um pequeno teste e retirou-se da base a 29ª observação, que continha valor bastante atípico para a variável B_4. Com isso as taxas da Ligação Média passam a configurar em torno de 74% enquanto que o *ROCK* e o k-Protótipos mantêm os mesmos patamares, próximos de 60% e de 70% respectivamente.

Para essa aplicação outros fatores foram estudados como por exemplo o tamanho dos grupos. O desbalanceamento, ou seja, o fato de um grupo tem muito mais observações que o outro, poderia afetar o desempenho das metodologias (LIMA, 2001). Porém, o número de grupos é o que parece melhor justificar os resultados não satisfatórios do *ROCK* e também o fato de não haver grandes diferenças entre as taxas de alocação correta das três estruturas de pesos.

Realizou-se a análise da agrupamentos pelo método da Ligação Média considerando somente as variáveis contínuas e a distância euclidiana como medida de proximidade. Observa-se pelo Dendograma (Figura 4.10), e também por alguns outros métodos de estimação do número de grupos k (MINGOTI, 2005), que o tamanho da partição não parece ser 2. Estima-se que hajam de 2 a 4 grupos de pacientes.

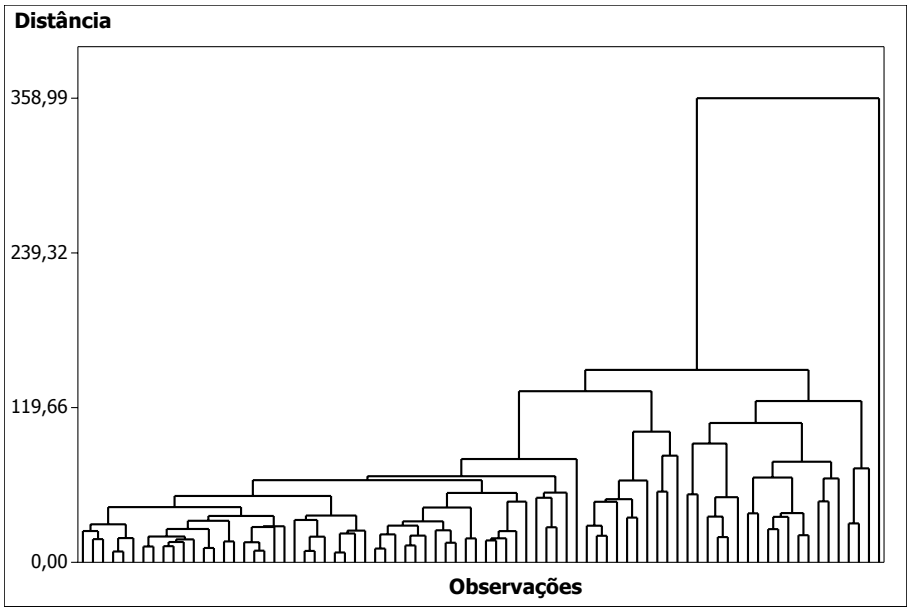

Figura 4.10: Dendograma do agrupamento utilizando apenas as variáveis contínuas do estudo sobre a hepatite, método da ligação média e distância euclidiana como medida de proximidade

Quando a análise de agrupamentos é feita somente com as variáveis categóricas, com exceção do *Fuzzy c*-Modas, todas as demais técnicas têm desempenhos satisfatórios (em torno de 70%). Assim, acredita-se que para os atributos categóricos pode ser que os dados estejam realmente estruturados de forma que existam 2 grupos de pacientes. Mas ao se utilizar os dois tipos de variáveis e estabelecer que hajam 2 grupos, as taxas de alocação correta dos métodos não são tão boas como o esperado.

5 *Conclusões e Considerações Finais*

5.1 Conclusões

5.1.1 Simulação de Variáveis Categóricas

Com base nos dados referentes aos graus de separação e sobreposição (seção 4.1.3.4.1 na página 62), percebeu-se que as taxas de alocação correta aumentam com o aumento da separação e diminuem com o aumento da sobreposição. O desempenho das metodologias depende também dos outros três fatores de análise, do número de grupos, de variáveis e de categorias.

O número de grupos (seção 4.1.3.4.2 na página 65) não tem o seu efeito influenciado pela estrutura do grupos. Ou seja, independentemente da disposição dos grupos e também do modelo (Beta ou Uniforme, especificados na seção 4.1.3.2 na página 58), o aumento dessa quantidade traz como conseqüência a diminuição das taxas de alocação correta dos métodos comparados.

O efeito da quantidade de atributos (seção 4.1.3.4.3 na página 71) nos métodos Ligação Média, *ROCK* e k-Populações depende da estrutura dos dados e o aumento dessa quantidade pode tanto contribuir quanto atrapalhar o processo de agrupamento. Para o k-Modas e o *Fuzzy* c-Modas o aumento do número de variáveis é prejudicial, independente dos grupos estarem separados ou sobrepostos. Como o número de variáveis pode ser determinado pelo pesquisador, recomenda-se a utilização do menor número possível de variáveis, para que o tempo computacional não seja muito elevado.

Em relação à eficiência dos métodos comparados, em ambos os modelos, percebeu-se que, quando havia separação dos grupos, a técnica *ROCK* tinha resultados próximos dos melhores métodos nesse cenário, a Ligação Média e o k-Populações. Quando os grupos são sobrepostos, no entanto, as taxas daquele método eram bem superiores em relação aos demais. Na prática, como a disposição dos grupos é desconhecida, recomenda-se a utilização do *ROCK*, dado que ele é tão bom quanto os melhores métodos do cenário 1 e

é melhor do que eles no cenário 2.

A Tabela 5.1 resume o efeito do aumento do número de grupos, variáveis e categorias no desempenho dos métodos avaliados. Nota-se que não é possível determinar um único efeito do aumento das categorias para todas as metodologias. No modelo Beta e independente do cenário, verificou-se que se aumento ocorresse de forma desigual, desbalanceando os domínios das variáveis categóricas, o *ROCK* sofria um efeito negativo na qualidade dos seus resultados. Por outro lado, se ocorresse de forma igual, eram os métodos k-Modas e *Fuzzy* c-Modas que eram afetados. Em outras circunstâncias e para as demais técnicas não é possível determinar um comportamento único, podendo os desempenhos melhorar ou piorarem. No modelo Uniforme observou-se que o aumento igual ou desigual das categorias não faz com que as taxas do *Fuzzy* c-Modas diminuam. Desconsiderando o caso em que há duas variáveis, o mesmo ocorre com a Ligação Média.

Tabela 5.1: Resumo do efeito do aumento do número de grupos, de variáveis e de categorias nas taxas de alocação correta dos métodos comparados

Método	**# Grupos**	**# Variáveis**		**# Categorias (igual)**		**# Categorias (desigual)**	
		Cenário 1	**Cenário 2**	**Cenário 1**	**Cenário 2**	**Cenário 1**	**Cenário 2**
Simulação de observações de variáveis categóricas							
Modelo Beta							
Ligação Média	d	d	a	a	d	a	d
ROCK	d (a k=3)	d	a	a	a (k=3;5) d (k=2)	d	d
k-Modas	d	d	m	m	d	dam	d
Fuzzy c-Modas	d	d	a	m	d	dam	d
k-Populações	d	d	a	a	d	a	d
Modelo Uniforme							
Ligação Média	d	d	a	a	dm	a	ma
ROCK	d	d	a	a	a	a	dam
k-Modas	d	d	d	a	da	a	dam
Fuzzy c-Modas	d	d	d	a	dm	a	dam
k-Populações	d	d	a	a	da	a	dam
Simulação de observações de variáveis categóricas e contínuas							
Modelo Beta							
Ligação Média	d	*	a	*	*	*	*
ROCK	d	*	a	*	*	*	*
k-Protótipos	d	*	a	*	*	*	*
Modelo Uniforme							
Ligação Média	d	*	a	*	*	*	*
ROCK	d	*	a	*	*	*	*
k-Protótipos	d	*	a	*	*	*	*

Nota: * indica que não se aplica

Legenda: d = diminui; a = aumenta; m = mantém; da = diminui e aumenta; dam = diminui, aumenta e mantém

5.1.2 Simulação de Variáveis Categóricas e Contínuas

Baseando-se na análise do grau de sobreposição em relação aos dois tipos de atributos (seção 4.1.4.4.1 na página 85) , notou-se que as taxas de alocação correta dos métodos Ligação Média e k-Protótipos diminuem com o aumento da sobreposição. As taxas do primeiro método mais do que as do segundo. Enquanto isso, o desempenho do *ROCK* praticamente não se altera. Isso demonstra que essa metodologia é mais estável a mudanças nas estruturas dos grupos.

Nesse contexto, observou-se que o número de grupos e o de variáveis (seção 4.1.4.4.2 na página 86) também afetam o desempenho dos métodos comparados. Enquanto o aumento do primeiro faz com que as taxas sejam menores, o aumento do segundo faz com que elas sejam maiores. O primeiro efeito é esperado uma vez que a existência de mais grupos dificulta a identificação dos mesmos pelas metodologias. Acredita-se que o segundo efeito reflete o que já havia sido constatado na simulação para variáveis categóricas. A estrutura de sobreposição imposta aos grupos por esse tipo de atributo é minimizada com o acréscimo de outras variáveis.

A correlação (seção 4.1.4.4.3 na página 88) entre as variáveis contínuas e a atribuição de pesos para a medida de proximidade combinada foram também objetos de estudos desta dissertação. O fato das variáveis contínuas serem ou não correlacionadas não alterou o desempenho dos métodos analisados e assim conclui-se que a correlação não é um fator determinante no processo de agrupamento de observações de variáveis categóricas e contínuas. Para alguns métodos de agrupamentos, Mingoti e Lima (2006) já haviam observado isso no caso de variáveis contínuas.

Quando as observações são caracterizadas por atributos dos dois tipos, recomenda-se que seja dado maior peso às variáveis contínuas (seção 4.1.4.4.4 na página 91). Todos os métodos, sem exceção, têm melhor desempenho quando essa estratégia é adotada. Por outro lado, têm piores resultados quando é dado maior peso às categóricas e isso, portanto, deve ser evitado. Isso pode ser devido ao fato das contínuas serem mais informativas que as categóricas.

5.1.3 Comentários Finais

Nos estudos de aplicação de variáveis categóricas (seção 4.2 na página 95) e na simulação de situações gerais, denominadas 64 e 65, (seção 4.1.3.4.5 na página 76), em que nada se especificou quanto à estrutura dos grupos, o *ROCK* se destacou, assim como Ligação

Média. No exemplo em que variáveis contínuas também estavam presentes, no entanto, o desempenho do primeiro método não foi interessante. Se somente essa aplicação particular tivesse servido como base para a comparação dos desempenhos, como acontece em muitos trabalhos encontrados na literatura, as propostas de extensão do *ROCK* poderiam ser abandonadas. Porém, no estudo de simulação o algoritmo proposto tem resultados muito superiores que a Ligação Média e k-Protótipos. Em algumas situações essas últimas metodologias alcançam o patamar das taxas do *ROCK* e em poucas delas os seus desempenhos são maiores. Quando isso ocorre, no entanto, o *ROCK* não deixa de possuir bons resultados. Lembra-se ainda que, no estudo sobre a hepatite, o que parece estar acontecendo é que o número de grupos especificado pela aplicação ($k = 2$) não existe de fato. Estimando-se esse número por meio do Dendograma apresentado na Figura 4.10 (página 107) e também com o auxílio de outros métodos, acredita-se que hajam de 2 a 4 grupos de pacientes com hepatite.

O *ROCK* destaca-se entre os métodos de agrupamentos de observações de variáveis categóricas, exclusivamente, e entre aqueles que permitem que variáveis contínuas também ocorram. É um método eficiente, estável e o seu conceito de robustez, relacionado ao uso de *links* ao invés de medidas de proximidade usuais para realizar o agrupamento, parece ser o diferencial dessa metodologia. Uma dificuldade que se apresenta é a escolha de θ. Um breve estudo foi feito nesta dissertação porém, é necessário verificar mais detalhadamente a sensibilidade desse método frente a diferentes valores desse parâmetro.

A Tabela 5.2 apresenta os postos dos métodos de agrupamentos segundo disposição dos grupos e modelo para a simulação de observações de variáveis categóricas. Os seguintes passos foram seguidos para se obter esses postos:

1. ordenar os métodos do melhor (1) para o pior (5) de acordo com a taxa de alocação correta obtida em cada situação simulada;

2. tirar uma média dos postos. Nesse passo, consideram-se para o cálculo da média o conjunto de situações que determinam a disposição dos grupos e número de variáveis categóricas (m);

3. fazer os arredondamentos necessários.

Pretende-se com essa tabela resumir os resultados obtidos nesta dissertação. Para o modelo Beta os métodos mais eficientes são o *ROCK* (posto= 1) e o k-Populações (posto= 2) e para o modelo Uniforme, o k-Modas e novamente o *ROCK*. Para ambos os

modelos a metodologia menos eficiente é o *Fuzzy* c-Modas (posto= 5). A Ligação Média possui desempenhos intermediários.

Da simulação de observações de variáveis categóricas e contínuas, obtém-se a Tabela 5.3. Os postos foram determinados de forma análoga, porém agora agregaram-se as informações quanto à estrutura de correlação (2) e pesos (3). Assim, a taxa de alocação média para a situação 1, por exemplo, e modelo Beta é a média das 6 taxas dessa situação e modelo. Na Tabela 5.3 apresentam-se os postos dos métodos comparados. Mais uma vez a eficiência do *ROCK* é destaque. Nota-se que a Ligação Média e o k-Protótipos têm resultados muito parecidos e estão empatados nos desempenhos médios de ambos os modelos.

Tabela 5.2: Postos dos métodos de agrupamentos comparados segundo disposição dos grupos e modelo, para simulação de dados categóricos

Disposição dos grupos		Ligação Média	*ROCK*	k-Modas	*Fuzzy* c-Modas	k-Populações
Modelo Beta						
Grau de separação	1(m=2)	1	3	5	4	2
	1(m=4)	3	2	5	4	1
	2	1	5	3	4	2
Grau de sobreposição	1(m=2)	5	2	1	3	4
	1(m=4)	3	1	2	5	4
	2	5	1	2	3	4
	3	4	1	3	2	5
Não há controle	64	3	1	4	5	2
	65	3	2	4	5	1
Média Modelo Beta		4	1	3	5	2
Modelo Uniforme						
Grau de separação	1(m=2)	1	3	4	5	2
	1(m=4)	5	1	3	4	2
	2	1	3	4	5	2
Grau de sobreposição	1(m=2)	5	4	1	2,5	2,5
	1(m=4)	4	1	2	5	3
	2	5	3	1	2	4
	3	4	1	3	2	5
Não há controle	64	3	2	4	5	1
	65	2	1	3	5	4
Média Modelo Uniforme		4	2	1	5	3

Legenda: 1 para melhor e 5 para pior

Tabela 5.3: Postos dos métodos de agrupamentos comparados segundo disposição dos grupos e modelo, para simulação de dados categóricos e contínuos

Sobreposição nas contínuas	Sobreposição nas categóricas	Ligação Média	*ROCK*	*k*-Protótipos
Modelo Beta				
20%	1	3	1	2
	2	2	1	3
	3	3	1	2
40%	1	3	1	2
	2	2	1	3
	3	3	1	2
80%	1	2	1	3
	2	2	1	3
	3	3	1	2
	Média Modelo Beta	2,5	1	2,5
Modelo Uniforme				
20%	1	3	1	2
	2	2	1	3
	3	3	1	2
40%	1	3	1	2
	2	2	1	3
	3	3	1	2
80%	1	2	1	3
	2	2	1	3
	3	2,5	1	2,5
	Média Modelo Uniforme	2,5	1	2,5

Legenda: 1 para melhor e 3 para pior

Outra forma de resumir a comparação das técnicas estudadas é fazendo-se uma análise descritiva das taxas de alocação correta. A Figura 5.1 apresenta os *boxplots* dos resultados obtidos pelos métodos Ligação Média, *ROCK*, k-Modas, *Fuzzy* c-Modas e k-Populações, nessa ordem, segundo o modelo (1 para Beta e 2 para Uniforme), o cenário e o grau (11 e 12 para grupos separados na primeira e nas duas primeiras variáveis categóricas, 21, 22 e 23 para grupos sobrepostos na primeira, nas duas primeiras e nas três primeiras variáveis categóricas, respectivamente).

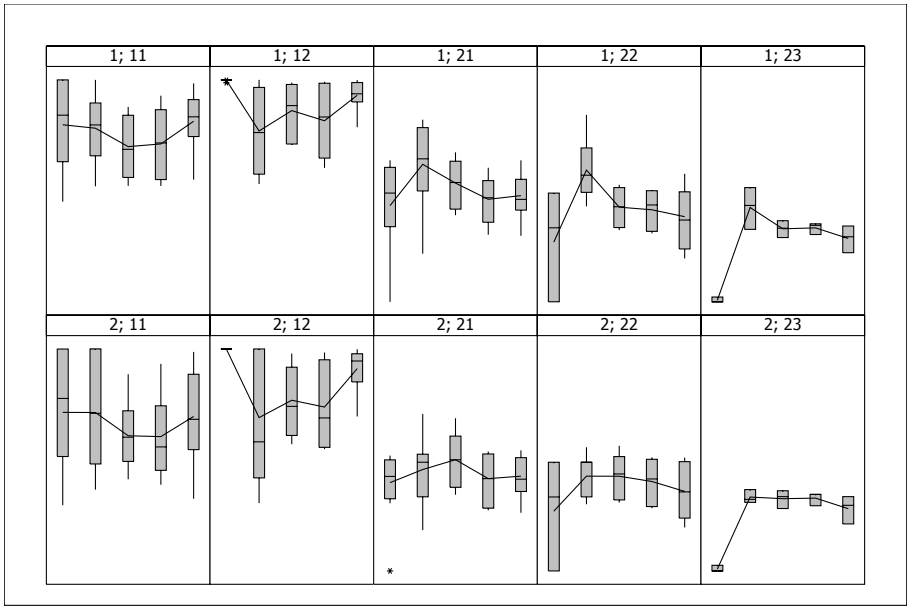

Figura 5.1: *Boxplots* das taxas de alocação correta dos métodos Ligação Média, *ROCK*, k-Modas, *Fuzzy* c-Modas e k-Populações segundo modelo, cenário e grau

Observa-se que a distribuição das taxas de alocação é muito dispersa, para todos os métodos. Os resultados são mais concentrados no cenário de sobreposição e no grau 3 (23). Chama a atenção o ótimo desempenho da Ligação Média no cenário de separação e no grau 2 12, para ambos os modelos. Por outro lado, esse método não tem bons resultados quando os grupos são sobrepostos pois, em média, as suas taxas são as menores sempre.

Dentre as 10 combinações de modelo, cenário e grau, os métodos *ROCK* e k-Modas aparecem em oito delas entre as três melhores técnicas. O k-Populações aparece em seis

e, finalmente, a Ligação Média e o *Fuzzy c*-Modas, aparecem em quatro. O *ROCK* é mais eficiente que o k-Modas, pois quando ele tem taxas superiores as desse último método, elas realmente são maiores. Enquanto que quando o k-Modas tem taxas superiores as do *ROCK*, elas não são tão maiores. A Ligação Média é menos eficiente que o *Fuzzy c*-Modas devido ao seu desempenho no cenário de sobreposição.

Assim, a lista das metodologias de análise de agrupamentos para varáveis categóricas, da mais eficiente para a menos eficiente seria: *ROCK*, k-Modas, k-Populações, *Fuzzy c*-Modas e Ligação Média.

A Figura 5.2 apresenta, para os modelos Beta e Uniforme, respectivamente, os *boxplots* das taxas de alocação correta dos métodos Ligação Média, *ROCK* e k-Protótipos, nessa ordem, segundo grau de sobreposição nas variáveis contínuas e nas categóricas.

O grau de sobreposição nas categóricas 212, 214, 224 e 234 referem-se aos casos em que os grupos são sobrepostos na primeira de duas variáveis (212) e na primeira, nas duas primeiras e nas três primeiras de quatro variáveis, respectivamente. E o grau de sobreposição nas variáveis contínuas 2, 4 e 8 referem-se aos casos $int = 20\%$, $int = 40\%$ e $int = 80\%$.

Nota-se que a extensão do *ROCK* apresentada nesta dissertação tem desempenho bastante superior em relação às demais metodologias estudadas. E isso acontece independente do modelo e do grau de sobreposição. Em relação à Ligação Média e ao k-Protótipos, a primeira técnica pode ser classificada como a menos eficiente. Porém, cabe ressaltar que os seus resultados são bastante próximos aos da segunda técnica.

Assim, para a análise de agrupamentos de observações caracterizadas pelos dois tipos de variáveis a lista das metodologias da mais para a menos eficiente seria: *ROCK*, k-Protótipos e Ligação Média.

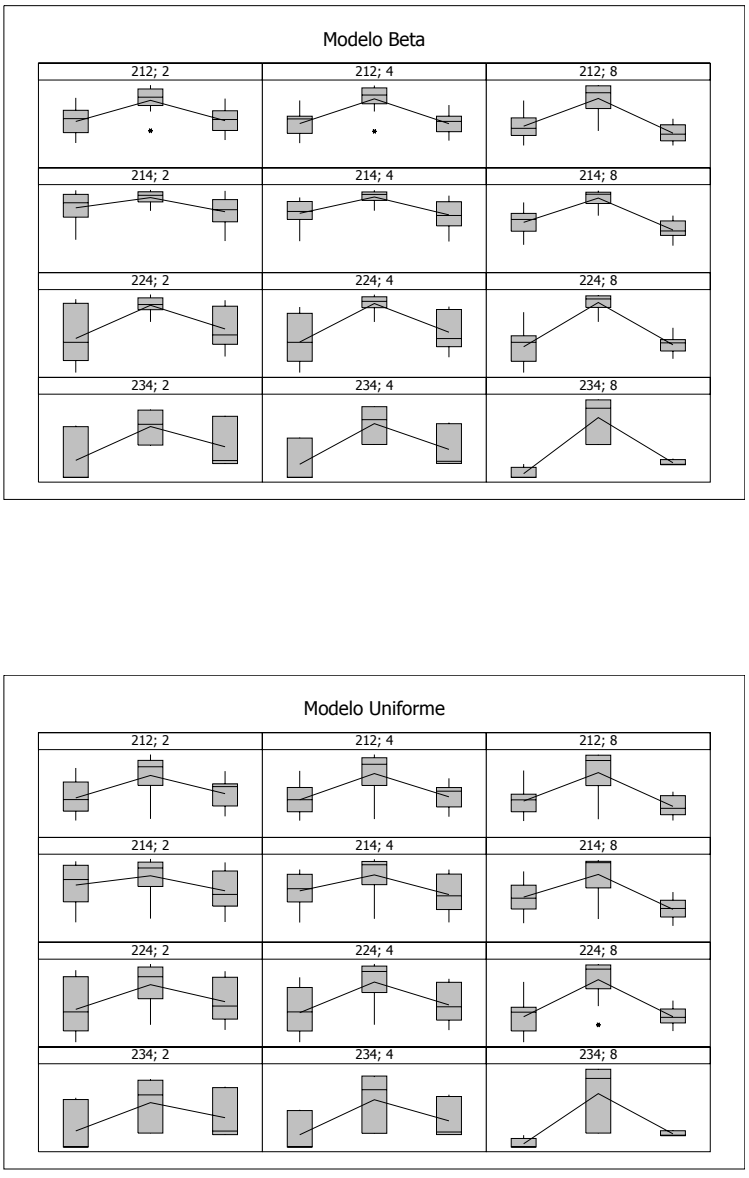

Figura 5.2: *Boxplots* das taxas de alocação correta dos métodos Ligação Média, *ROCK* e k-Protótipos segundo grau de sobreposição nas variáveis categóricas e contínuas, para os modelos Beta e Uniforme

5.2 Considerações Finais

Nesta dissertação elaborou-se uma primeira comparação entre as metodologias de análise de agrupamentos e pelo fato dessa área englobar vários aspectos, alguns não puderam ser desenvolvidos. Os seguintes objetivos poderão fazer parte dos trabalhos que darão prosseguimento a este:

- propor novas estratégias de simulação de dados categóricos ou, talvez, somente alterar os parâmetros da distribuição Beta (página 58). Pretende-se com isso testar a sensibilidade dos métodos de agrupamentos dadas as diferentes probabilidades de ocorrência das categorias das variáveis. Incorporar informações a respeito de possíveis correlações existentes entre si e entre elas e os atributos contínuos;

- implementar o k-Representativos e o algoritmo de refinamento dos pontos iniciais. Apesar dos artigos-referência se basearem somente no método k-Modas, acredita-se que essas modificações possam ser incorporadas também ao k-Protótipos e talvez melhorar o desempenho dessas metodologias em relação às demais;

- estender o *Fuzzy* c-Modas e construir o *Fuzzy* c-Protótipos, o qual permitiria agrupar observações tanto de variáveis categóricas quanto contínuas. No algoritmo k-Protótipos, que integra o k-Médias e o k-Modas, utiliza-se a medida de proximidade combinada, que considera tanto o coeficiente de dissimilaridade das variáveis categóricas quanto a distância euclidiana, das contínuas, ambos transformados em similaridades. Essa idéia foi inicialmente testada, sem sucesso, no *Fuzzy* c-Modas. Entende-se que seja possível então ao invés de fazer a combinação dos coeficientes de proximidade, combinar as matrizes de partição (W) obtidas a partir deles.

- propor novas expressões para a função $f(\theta)$, além da originalmente proposta pelos autores do *ROCK* em Guha *et al.* (2000);

- estudar mais extensivamente a influência dos parâmetros θ, do método *ROCK*, e α, do *Fuzzy* c-Modas e do k-Populações. Pelo fato dessas quantidades fazerem parte desses algoritmos acredita-se que o desempenho dessas técnicas está relacionado com os valores atribuídos a elas. Caso esses parâmetros tenham grande influência nos agrupamentos finais, é interessante identificar ponto(s) do intervalo de valores testado o(s) qual(is) fornece(m) os melhores resultados para esses métodos;

- alterar nas simulações para variáveis categóricas e contínuas a medida de proximidade combinada pela proposta de *Gower*. Com isso, será possível verificar o impacto

da ocorrência de valores faltantes. Uma sugestão para esse estudo é fazer a análise de agrupamentos no banco de dados original, retirar aleatoriamente algumas observações desse banco e, novamente, fazer a análise. Ao final, comparar os resultados das duas análises realizadas, para cada objeto agrupado;

- verificar o impacto dos erros de classificação. Uma idéia preliminar é análoga ao estudo dos valores faltantes. Dado um banco de dados, aplicar pequenas, médias e grandes distorções no dados e comparar os agrupamentos originais e finais. Outra idéia é por em prática o método proposto no artigo Andreopoulos *et al.* (2005), que considera graus de confiabilidade dos dados durante o processo de agrupamento.

Um primeiro estudo comparativo dos métodos de análise de agrupamentos de observações caracterizadas por variáveis categóricas foi realizado nesta dissertação. As conclusões foram tiradas a partir de uma comparação adequada das técnicas e os resultados norteiam a aplicação delas. Além disso, foram desenvolvidas duas novas metodologias. O *ROCK* é agora aplicável a situações em que ocorram variáveis contínuas, exclusivamente, e a situações em que essas ocorram conjuntamente com as categóricas. Os estudos realizados indicam a potencialidade dessas novas técnicas, apesar de análises mais profundas com respeito ao parâmetro θ precisarem ser feitas.

Referências Bibliográficas

ANDERBERG, M. R. *Cluster Analysis for Applications*. New York: Academic Press, Inc., 1973.

ANDREOPOULOS, B.; AN, A.; WANG, X. Clustering mixed numerical and low quality categorical data: significance metrics on a yeast example. In: *IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems*. New York: ACM Press, 2005. p. 87–98.

ANDREOPOULOS, B.; AN, A.; WANG, X. Bi-level clustering of mixed categorical and numerical biomedical data. *International Journal of Data Mining and Bioinformatics*, v. 1, n. 1, p. 19–56, 2006.

ANDRITSOS, P. *Scalable Clustering of Categorical Data and Applications*. 2004. Tese (Doutorado) — Universidade de Toronto. Disponível em: <http:// www.dit.unitn.it/ p̃eriklis/ publications.html>.

BEZDEK, J. C. *Fuzzy Mathematics in Pattern Classification*. 1973. Tese (Doutorado) — Center for Applied Mathematics, Cornell University.

BRADLEY, P. S.; FAYYAD, U. M. Refining initial points for k-means clustering. In: *Proceedings of the 15th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1998. p. 91–99.

CHIU, T. *et al.* A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM Press, 2001. p. 263–268.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons Inc, 1973.

DUTTA, M.; MAHANTA, A. K.; PUJARI, A. K. QROCK: A quick version of the rock algorithm for clustering of categorical data. *Pattern Recognition*, v. 26, p. 2364–2373, 2005.

EVERITT, B. S. *Cluster Analysis*. New York: John Wiley & Sons, Inc., 1993.

FELIX, F. do N. *Aplicando Bootstrap para Determinação de Intervalos de Confiança para o Número de Grupos no Procedimento Hierárquico Aglomerativo de Ward*. 2004. Dissertação (Mestrado) — Departamento de Estatística, Universidade Federal de Minas Gerais.

GANTI, V.; GEHRKE, J.; RAMAKRISHNAN, R. CACTUS: Clustering categorical data using summaries. In: *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: ACM Press, 1999. p. 73–83.

GIBSON, D.; KLEINBERG, J. M.; RAGHAVA, P. Clustering categorical data: An approach based in dynamical systems. *The VLDB Journal*, Springer-Verlag New York, Inc., v. 8, n. 3-4, p. 222–236, 2000.

GOWER, J. C. A general coefficient of similarity and some of its properties. *BioMetrics*, v. 27, p. 857–872, 1971.

GREENACRE, M. J. *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1984.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, v. 25, p. 345–366, 2000.

HARTIGAN, J. A. *Clustering Algorithm*. New York: John Wiley & Sons, Inc., 1975.

HUANG, Z. Clustering large data sets with mixed numeric and categorical values. In: *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*. Singapore: World Scientific, 1997. p. 21–34.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, v. 2, p. 283–304, 1998.

HUANG, Z.; NG, M. K. A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, v. 7, p. 446–452, 1999.

JOBSON, J. D. *Applied Multivariate Data Analysis*. New York: Springer, 1991–92.

JOHNSON, D. E. *Applied Multivariate Methods for Data Analysts*. Pacific GroveLondon: Duxbury Press, 1998.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Englewood Clifs: Prentice-Hall, 2002.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc., 1990.

KIM, D.-W. *et al.* A k-populations algorithm for clustering categorical data. *Pattern Recognition*, v. 38, p. 1131–1134, 2005.

KIM, D.-W.; LEE, K. H.; LEE, D. Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters*, v. 25, p. 1263–1271, 2004.

LIMA, J. de O. *Uma Comparação dos Métodos Fuzzy e Redes Neurais Artificiais com os Procedimentos de Agrupamentos Hierárquicos e Não Hierárquicos Tradicionais*. 2001. Dissertação (Mestrado) — Departamento de Estatística, Universidade Federal de Minas Gerais.

MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, v. 45, n. 3, p. 325–342, 1980.

MILLIGAN, G. W. An algorithm for generating artificial test clusters. *Psychometrika*, v. 50, n. 1, p. 123–127, 1985.

MINGOTI, S. A. *Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada*. Belo Horizonte: Editora UFMG, 2005.

MINGOTI, S. A.; LIMA, J. O. Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, v. 174, n. 3, p. 1742–1759, 2006.

OH, C.-H.; HONDA, K.; ICHIHASHI, H. Fuzzy clustering categorical multivariate data. In: *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*. Vancouver: [s.n.], 2001. p. 2154–2159.

PANTUZZO, A. E. *Comparação de Métodos de Análise de Cluster na Presença de Dados Categóricos e a Aplicação no Contexto de Data Mining: Estudo de Casos*. 2002. Dissertação (Mestrado) — Departamento de Engenharia de Produção, Universidade Federal de Minas Gerais.

RALAMBONDRAINY, H. A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, v. 16, p. 1147–1157, 1995.

REIS, E. *Estatística Multivariada Aplicada*. Lisboa: Edições Sílabo, 1997.

RENCHER, A. C. *Methods of Multivariate Analysis*. New York: John Wiley & Sons, 1995.

SAN, O. M.; HUYNH, V.-N.; NAKAMORI, Y. An alternative extension of the k-means algorithm for clustering categorical data. *Int. J. Appl. Math. Comput. Sci.*, v. 14, p. 241–247, 2004.

SHARMA, S. *Applied Multivariate Techniques*. New York: John Wiley & Sons, 1996.

SOKAL, R. R.; SNEATH, P. H. A. *Principle of Numerical Taxonomy*. San Francisco: Freeman, 1963.

SUN, Y.; ZHU, Q.; CHEN, Z. An interative initial-points refinement algorithm for categorical data clustering. *Pattern Recognition Letters*, v. 23, p. 875–884, 2002.

TIMM, N. H. *Applied Multivariate Analysis*. New York: Springer, 2002.

UMAYAHARA, K.; MIYAMOTO, S.; NAKAMORI, Y. Formulations of fuzzy clustering for categorical data. *International Journal of Innovative Computing, Information and Control*, v. 1, n. 1, p. 83–94, March 2005.

ANEXO A – Taxa de alocação correta das situações da simulação de variáveis categóricas

Tabela A.1: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para modelo Beta$(1; 0, 1)$

Situação	Ligação Média	*ROCK*	k-Modas	*Fuzzy* c-Modas	k-Populações
Grau de separação 1					
1	84,31	79,63	87,13	87,20	89,22
2	100,00	96,90	85,44	89,74	96,86
3	100,00	100,00	88,04	92,98	98,41
4	72,58	66,56	70,03	72,26	75,99
5	100,00	85,01	70,17	71,99	85,81
6	100,00	99,89	75,40	72,52	92,11
7	59,79	53,01	57,18	56,16	62,41
8	100,00	93,78	60,81	55,76	82,92
9	84,42	87,92	85,38	86,41	83,67
10	90,94	80,36	83,91	86,47	88,81
11	91,98	80,09	84,84	87,12	90,40
12	99,96	91,59	83,70	87,92	97,79
14	65,45	80,59	69,29	71,75	77,42
13	70,34	84,74	68,69	72,13	79,59
15	67,40	70,38	68,46	71,32	83,95
16	98,54	77,17	68,86	72,15	93,88
20	47,01	64,67	56,62	56,30	55,94
18	46,22	62,72	54,90	54,84	63,07

continua na próxima página

Tabela A.1 – continuação da página anterior

Situação	Ligação Média	*ROCK*	*k*-Modas	*Fuzzy c*-Modas	*k*-Populações
19	46,68	59,27	54,53	53,60	66,05
17	62,31	66,36	53,55	53,47	78,30
21	95,35	70,92	53,19	53,14	73,94
Grau de separação 2					
22	100,00	90,53	97,89	98,48	93,82
23	100,00	99,38	97,94	98,50	98,80
25	100,00	100,00	98,25	98,83	99,90
24	100,00	96,65	98,78	99,12	99,82
27	100,00	75,03	89,77	86,17	90,24
26	100,00	76,72	87,50	83,61	95,69
28	100,00	86,19	88,59	77,39	97,50
32	98,70	56,10	77,20	70,30	79,09
30	99,78	58,53	71,67	65,47	85,09
29	100,00	58,37	71,21	62,50	93,75
31	100,00	54,01	71,47	61,16	91,20
Grau de sobreposição 1					
33	64,48	81,08	64,42	55,74	60,35
34	50,00	51,83	63,71	55,26	53,02
35	50,00	50,00	61,10	53,53	55,73
36	51,73	71,60	55,43	46,68	47,24
37	34,69	45,46	53,83	45,53	40,79
38	34,64	34,77	52,35	44,84	44,74
39	37,60	51,91	42,62	33,88	33,84
40	20,00	23,26	40,25	31,62	31,13
41	63,03	81,85	67,96	61,20	64,48
42	63,60	82,41	66,57	58,70	59,68
43	50,07	65,14	66,24	59,29	56,42

continua na próxima página

Tabela A.1 – continuação da página anterior

Situação	Ligação Média	*ROCK*	*k*-Modas	*Fuzzy c*-Modas	*k*-Populações
44	62,29	81,78	55,12	50,89	54,11
45	60,43	70,50	54,66	48,40	50,21
46	35,65	73,33	53,61	47,97	46,74
47	45,95	63,34	43,24	38,42	42,70
48	45,30	61,09	42,53	35,84	42,27
49	40,68	76,73	41,11	35,67	47,02
Grau de sobreposição 2					
50	50,00	57,88	53,70	51,32	58,51
51	50,00	50,30	53,08	51,16	54,15
52	50,00	52,26	51,64	50,75	50,39
53	50,00	50,00	52,41	51,04	50,54
54	34,66	84,54	44,70	44,79	38,10
55	34,64	77,00	43,96	44,26	44,42
56	34,66	44,15	43,33	44,91	36,46
57	20,00	69,91	34,79	32,42	21,16
58	20,04	61,69	35,69	33,12	24,47
59	20,00	56,11	33,61	32,24	25,35
60	20,03	58,10	34,68	33,11	31,88
Grau de sobreposição 3					
61	40,11	44,54	37,80	36,45	35,46
62	20,00	33,96	34,50	35,75	30,72
63	20,00	52,40	30,35	31,68	23,69
Não há controle sobre a separação e a sobreposição dos grupos					
64	50,43	72,83	46,83	42,01	54,85
65	88,74	89,34	81,18	77,75	90,12

Tabela A.2: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para modelo Uniforme

Situação	Ligação Média	*ROCK*	k-Modas	*Fuzzy* c-Modas	k-Populações
Grau de separação 1					
1	78,46	70,93	78,85	79,90	78,53
2	100,00	100,00	81,83	88,82	97,62
3	100,00	100,00	88,89	93,42	98,50
4	66,78	56,21	61,95	61,92	67,40
5	100,00	100,00	64,51	63,99	91,38
6	100,00	100,00	74,59	66,87	86,35
7	54,26	45,04	53,28	49,17	59,94
8	100,00	100,00	61,05	52,88	78,07
9	66,49	73,01	71,89	72,91	68,95
10	78,15	71,62	69,91	74,76	75,33
11	78,60	71,68	70,45	75,07	76,06
12	99,98	100,00	73,47	80,02	98,71
14	51,59	53,46	54,78	54,29	53,23
13	54,17	59,21	54,69	54,48	56,42
15	53,53	54,55	57,17	56,35	68,81
16	98,02	99,98	56,45	56,76	93,55
20	31,00	39,39	43,56	40,05	33,89
18	35,89	38,11	42,48	40,20	34,26
19	35,17	37,91	43,88	41,13	42,38
17	42,16	40,94	43,63	40,98	54,68
21	87,58	99,24	47,49	43,65	60,71
Grau de separação 2					
22	100,00	100,00	88,29	92,23	85,49
23	100,00	100,00	91,93	95,22	96,44
25	100,00	100,00	95,96	97,54	99,77

continua na próxima página

Tabela A.2 – continuação da página anterior

Situação	Ligação Média	*ROCK*	*k*-Modas	*Fuzzy c*-Modas	*k*-Populações
24	100,00	100,00	98,00	98,46	99,90
27	100,00	50,36	74,09	69,56	82,33
26	100,00	58,99	74,64	68,50	94,68
28	100,00	100,00	83,46	70,48	97,87
32	100,00	31,96	59,53	57,13	70,21
30	100,00	37,88	57,94	55,75	90,87
29	100,00	43,09	61,83	56,44	97,41
31	100,00	44,83	65,52	56,56	89,42
Grau de sobreposição 1					
33	52,93	71,35	69,43	53,92	55,01
34	50,00	50,00	61,82	53,33	51,00
35	50,00	50,00	58,79	52,62	55,22
36	43,73	52,97	54,08	42,39	39,70
37	34,66	34,74	52,30	42,85	36,73
38	34,65	34,67	51,10	44,12	48,58
39	31,87	33,52	39,31	29,68	27,70
40	20,00	20,00	38,50	29,72	34,14
41	52,89	71,29	61,26	54,73	53,77
42	52,81	71,01	62,83	54,47	53,09
43	50,03	50,00	62,74	54,32	50,47
44	50,93	53,58	47,84	42,06	40,90
45	51,15	53,45	50,12	42,76	41,50
46	35,49	34,66	49,91	41,48	42,49
47	33,12	35,34	35,67	29,68	28,30
48	37,27	40,81	37,12	29,76	37,44
49	33,08	38,13	36,43	28,78	49,14

continua na próxima página

Tabela A.2 – continuação da página anterior

Situação	Ligação Média	*ROCK*	*k*-Modas	*Fuzzy c*-Modas	*k*-Populações
Grau de sobreposição 2					
50	50,00	50,00	57,24	52,15	52,07
51	50,00	50,00	54,53	51,88	50,60
52	50,00	50,00	51,49	50,93	50,25
53	50,00	50,00	52,52	51,31	50,06
54	34,69	56,59	44,84	42,63	35,53
55	34,65	50,26	45,93	42,52	37,41
56	34,65	34,67	43,09	43,89	36,90
57	20,00	37,20	32,81	30,45	21,19
58	20,00	38,17	33,78	30,01	22,13
59	20,00	34,10	32,22	30,90	25,31
60	20,00	31,45	33,36	29,68	26,83
Grau de sobreposição 3					
61	40,48	37,91	37,40	35,78	34,79
62	20,00	32,28	34,78	35,80	30,99
63	20,00	33,63	29,55	30,79	22,75
Não há controle sobre a separação e a sobreposição dos grupos					
64	37,42	41,87	36,65	31,30	48,08
65	66,05	68,14	58,37	55,19	55,22

ANEXO B – Taxa de alocação correta das situações da simulação de variáveis categóricas e contínuas

Tabela B.1: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para pesos iguais, variáveis contínuas não-correlacionadas e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	64,82	92,92	63,10	63,88	94,83	61,26	62,16	96,37	57,58
4	53,40	86,86	56,29	52,61	91,04	54,34	50,24	96,74	48,64
7	43,16	78,89	45,28	41,54	83,73	43,41	38,60	96,70	37,73
Média	53,79	86,23	54,89	52,67	89,86	53,00	50,33	96,60	47,98
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	84,71	95,60	82,93	79,35	94,79	77,50	76,92	93,31	67,25
12	85,71	91,36	77,66	76,55	92,39	72,08	67,59	94,41	56,06
17	81,20	79,25	67,10	71,78	82,21	63,27	50,88	94,62	44,76
Média	83,87	88,74	75,90	75,90	89,80	70,95	65,13	94,11	56,02
Grau de sobreposição 2 nas categóricas									
18	50,11	93,27	61,15	50,01	95,10	56,91	50,08	96,24	52,37
22	36,03	86,56	57,37	34,88	90,03	52,22	34,80	96,08	44,25
25	23,94	75,94	48,93	20,92	79,69	44,88	20,60	95,90	34,70
Média	36,70	85,26	55,82	35,27	88,27	51,34	35,16	96,07	43,77
Grau de sobreposição 3 nas categóricas									

Tabela B.2: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis categóricas, variáveis contínuas não-correlacionadas e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	63,48	80,93	63,62	64,71	81,18	62,13	63,24	81,25	58,36
4	52,46	71,09	55,64	52,03	71,01	54,30	51,06	71,10	48,50
7	39,88	51,72	42,80	39,75	51,55	42,27	37,45	52,22	37,47
Média	51,94	67,91	54,02	52,17	67,91	52,90	50,59	68,19	48,11
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	73,19	94,98	70,61	73,11	95,24	70,07	71,48	96,15	66,59
12	60,56	88,54	60,99	59,76	88,29	59,61	57,64	87,88	55,02
17	48,08	76,29	46,75	46,50	76,41	46,20	42,73	76,32	42,07
Média	60,61	86,60	59,45	59,79	86,64	58,63	57,28	86,78	54,56
Grau de sobreposição 2 nas categóricas									
18	50,00	88,25	54,13	50,00	90,64	54,01	50,00	91,41	52,39
22	34,63	84,20	46,50	34,63	85,19	45,90	34,65	84,38	43,63
25	20,00	70,24	35,98	20,01	70,18	35,77	20,00	70,26	33,71
Média	34,88	80,90	45,54	34,88	82,00	45,23	34,88	82,02	43,24
Grau de sobreposição 3 nas categóricas									
31	20,00	51,49	33,74	20,00	52,22	33,84	20,00	52,44	32,56

Tabela B.3: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis contínuas, não-correlacionadas, e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	84,56	97,10	83,81	81,99	96,71	77,45	81,93	95,92	63,82
4	77,97	93,51	78,95	71,69	94,18	72,28	70,38	90,41	52,67
7	70,32	85,26	69,51	64,44	87,57	64,71	54,59	79,60	43,74
Média	77,62	91,96	77,42	72,71	92,82	71,48	68,96	88,64	53,41
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	94,76	96,91	95,96	89,43	96,67	91,24	84,62	94,76	71,53
12	94,26	93,87	94,03	85,53	94,80	90,55	72,70	89,73	61,76
17	96,49	87,11	85,19	89,32	89,82	84,07	59,94	82,83	56,77
Média	95,17	92,63	91,73	88,09	93,76	88,62	72,42	89,11	63,35
Grau de sobreposição 2 nas categóricas									
18	92,62	97,31	90,32	84,38	97,16	84,53	79,51	95,68	64,38
22	91,31	94,29	91,64	80,01	95,00	85,39	68,00	93,83	55,04
25	88,31	86,92	84,46	79,34	90,41	82,72	52,22	87,60	50,27
Média	90,75	92,84	88,80	81,24	94,19	84,21	66,58	92,37	56,56
Grau de sobreposição 3 nas categóricas									
31	71,12	86,28	80,82	59,18	89,80	74,06	28,81	92,00	38,55

Tabela B.4: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para pesos iguais, variáveis contínuas correlacionadas e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	64,81	93,27	63,46	64,30	95,08	61,23	63,05	96,03	57,82
4	53,40	87,31	55,68	52,47	91,90	53,61	50,73	96,75	48,62
7	43,05	79,23	44,96	41,42	84,01	43,29	39,04	96,63	37,49
Média	53,75	86,61	54,70	52,73	90,33	52,71	50,94	96,47	47,98
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	84,27	94,81	81,30	79,75	93,96	76,67	78,78	92,20	66,23
12	84,24	91,78	77,17	74,89	92,50	71,32	67,97	93,97	55,72
17	80,29	79,23	66,96	70,35	82,93	62,03	51,66	95,24	44,23
Média	82,93	88,61	75,14	74,99	89,80	70,01	66,14	93,80	55,39
Grau de sobreposição 2 nas categóricas									
18	50,21	93,82	61,21	50,07	95,15	57,32	50,10	95,94	52,23
22	35,97	87,41	57,26	34,80	91,57	52,26	34,94	96,30	44,05
25	23,58	76,48	48,65	20,96	81,63	44,97	20,87	96,45	34,12
Média	36,59	85,90	55,71	35,28	89,45	51,52	35,30	96,23	43,47
Grau de sobreposição 3 nas categóricas									
31	20,00	72,73	36,92	20,01	77,89	36,00	20,00	96,80	32,79

Tabela B.5: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis categóricas, variáveis contínuas correlacionadas e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	64,44	81,45	63,10	63,46	81,74	61,77	62,78	81,24	57,93
4	52,51	71,02	54,64	52,28	71,16	54,25	50,11	70,43	48,03
7	40,35	52,27	42,98	39,79	51,26	42,20	37,88	51,83	37,44
Média	52,43	68,25	53,58	51,84	68,05	52,74	50,25	67,84	47,80
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	73,62	95,24	71,07	72,92	95,45	69,47	71,39	95,49	65,64
12	60,49	88,39	60,04	60,12	87,84	59,58	57,67	88,01	55,00
17	47,72	76,62	46,50	46,33	76,64	46,00	43,27	76,19	41,91
Média	60,61	86,75	59,20	59,79	86,65	58,35	57,44	86,56	54,18
Grau de sobreposição 2 nas categóricas									
18	50,00	89,53	53,97	50,00	91,12	53,44	50,00	92,03	52,08
22	34,65	85,42	46,25	34,64	86,15	46,12	34,64	84,62	44,11
25	20,00	70,23	36,04	20,01	70,16	35,23	20,00	70,37	33,90
Média	34,88	81,73	45,42	34,88	82,48	44,93	34,88	82,34	43,37
Grau de sobreposição 3 nas categóricas									
31	20,00	52,02	33,35	20,00	52,57	33,60	20,00	52,28	32,41

Tabela B.6: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis contínuas, correlacionadas, e modelo Beta$(1; 0, 1)$

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	84,51	96,92	83,28	81,09	96,38	76,98	80,39	94,91	63,14
4	77,68	93,36	78,57	71,34	94,09	71,38	68,94	88,78	51,95
7	70,40	85,19	69,14	64,48	87,20	64,52	54,08	74,96	43,47
Média	77,53	91,83	77,00	72,30	92,56	70,96	67,80	86,21	52,85
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	91,91	96,77	94,54	86,56	95,96	89,23	81,85	93,06	68,15
12	91,33	94,03	93,73	80,97	93,49	88,98	71,66	83,81	59,29
17	94,49	87,49	84,86	85,51	88,87	82,93	60,02	71,52	54,50
Média	92,58	92,76	91,05	84,35	92,77	87,05	71,18	82,79	60,64
Grau de sobreposição 2 nas categóricas									
18	90,99	97,23	88,85	84,97	96,77	82,44	79,80	95,13	62,61
22	89,41	94,59	90,21	78,46	94,58	83,36	68,13	88,60	53,26
25	87,66	87,26	84,72	77,16	89,93	81,30	52,59	77,21	48,68
Média	89,35	93,03	87,93	80,20	93,76	82,37	66,84	86,98	54,85
Grau de sobreposição 3 nas categóricas									
31	69,76	86,88	80,18	58,64	89,89	72,64	33,21	84,46	37,22

Tabela B.7: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para pesos iguais, variáveis contínuas não-correlacionadas e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	52,88	90,35	66,32	52,79	94,05	63,66	52,52	96,13	56,72
4	43,71	84,99	50,07	43,16	91,45	48,90	43,15	96,88	44,05
7	36,63	78,50	37,59	34,48	83,09	36,45	33,40	96,99	32,16
Média	44,41	84,62	51,33	43,48	89,53	49,67	43,02	96,67	44,31
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	76,58	94,38	69,32	69,33	95,25	64,77	70,28	95,67	57,14
12	78,82	90,42	60,54	68,08	92,02	54,59	65,06	94,90	43,38
17	73,31	71,72	53,46	63,61	75,91	48,79	46,84	94,88	33,22
Média	76,24	85,51	61,11	67,01	87,73	56,05	60,73	95,15	44,58
Grau de sobreposição 2 nas categóricas									
18	50,05	93,45	58,40	50,00	95,72	56,33	50,01	96,41	53,28
22	34,73	83,52	50,23	34,64	90,50	46,79	34,76	96,14	42,30
25	20,81	71,97	39,92	20,34	76,69	37,47	20,48	96,27	31,45
Média	35,20	82,98	49,52	34,99	87,64	46,86	35,08	96,27	42,34
Grau de sobreposição 3 nas categóricas									
31	20,00	70,81	35,67	20,00	75,35	34,65	20,00	96,57	31,69

Tabela B.8: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis categóricas, variáveis contínuas não-correlacionadas e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	*k*-Protótipos	Ligação Média	*ROCK*	*k*-Protótipos	Ligação Média	*ROCK*	*k*-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	52,66	71,65	65,89	52,51	71,05	63,38	52,44	71,01	57,16
4	42,75	53,42	49,30	43,27	53,51	48,95	43,78	53,24	44,09
7	31,99	33,64	36,36	31,87	33,40	35,80	31,47	33,28	32,02
Média	42,46	52,90	50,52	42,55	52,65	49,38	42,56	52,51	44,42
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	57,32	86,83	62,28	57,56	91,65	61,33	56,91	94,26	56,84
12	48,48	65,96	46,55	47,91	60,76	46,02	48,54	59,63	42,66
17	35,14	38,68	35,22	34,80	38,36	35,15	34,12	38,19	31,79
Média	46,98	63,82	48,02	46,75	63,59	47,50	46,53	64,03	43,76
Grau de sobreposição 2 nas categóricas									
18	50,00	65,23	56,13	50,00	73,27	55,05	50,00	81,27	52,80
22	34,68	56,19	44,13	34,63	56,90	43,34	34,68	56,51	41,79
25	20,01	37,20	32,27	20,00	37,31	31,96	20,00	36,89	31,25
Média	34,90	52,87	44,18	34,88	55,83	43,45	34,89	58,22	41,95
Grau de sobreposição 3 nas categóricas									
31	20,00	33,59	32,08	20,00	33,55	32,19	20,00	33,52	31,05

Tabela B.9: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis contínuas não-correlacionadas e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	*k*-Protótipos	Ligação Média	*ROCK*	*k*-Protótipos	Ligação Média	*ROCK*	*k*-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	84,01	97,32	80,91	81,16	97,27	73,76	81,48	96,59	60,51
4	76,20	93,95	75,33	71,71	94,27	68,40	69,83	92,15	49,16
7	68,08	85,44	66,18	62,46	88,13	61,20	54,23	81,01	40,34
Média	76,10	92,24	74,14	71,78	93,22	67,79	68,51	89,92	50,00
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	93,39	97,49	94,03	87,05	97,17	87,00	85,17	96,26	64,81
12	92,35	94,33	92,17	82,66	94,90	86,12	72,37	91,50	54,48
17	95,10	87,63	83,89	86,35	90,30	82,10	58,66	84,58	49,66
Média	93,61	93,15	90,03	85,35	94,12	85,07	72,07	90,78	56,32
Grau de sobreposição 2 nas categóricas									
18	91,21	97,36	90,24	83,19	97,42	82,74	79,34	97,00	61,18
22	88,78	94,44	89,22	77,58	95,15	81,84	67,64	94,01	50,90
25	84,14	87,78	82,98	73,53	90,06	79,94	49,58	88,11	45,57
Média	88,05	93,19	87,48	78,10	94,21	81,50	65,52	93,04	52,55
Grau de sobreposição 3 nas categóricas									
31	68,60	85,58	79,57	56,64	90,02	71,66	27,32	91,78	36,34

Tabela B.10: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para pesos iguais, variáveis contínuas correlacionadas e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	52,91	90,73	66,22	52,74	94,11	63,04	52,58	96,54	56,51
4	43,69	85,97	50,24	43,20	91,73	48,54	43,42	97,02	43,82
7	37,08	78,88	37,91	34,89	84,34	36,27	33,44	97,02	32,18
Média	44,56	85,20	51,46	43,61	90,06	49,28	43,14	96,86	44,17
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	78,02	94,70	68,55	73,16	95,12	64,01	73,42	95,18	56,64
12	77,68	90,53	59,76	68,13	91,74	53,86	66,48	94,53	42,99
17	71,55	73,12	52,46	62,12	77,58	48,07	47,62	95,49	32,90
Média	75,75	86,12	60,26	67,80	88,15	55,31	62,51	95,07	44,18
Grau de sobreposição 2 nas categóricas									
18	50,05	94,70	58,29	50,06	96,09	56,01	50,06	96,55	52,98
22	34,68	86,19	49,84	34,69	91,91	46,37	34,78	96,53	41,79
25	20,85	73,15	39,65	20,48	78,80	37,21	20,60	96,60	31,47
Média	35,20	84,68	49,26	35,08	88,93	46,53	35,15	96,56	42,08
Grau de sobreposição 3 nas categóricas									
31	20,00	72,10	35,69	20,00	77,72	34,86	20,00	97,06	31,57

Tabela B.11: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis categóricas, variáveis contínuas correlacionadas e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	52,88	71,23	65,18	52,82	71,42	62,50	52,59	71,60	56,45
4	42,80	52,99	49,48	42,65	52,87	48,71	43,14	53,05	43,74
7	32,06	33,62	36,15	32,03	33,43	35,78	31,93	33,55	32,12
Média	42,58	52,61	50,27	42,50	52,57	48,99	42,55	52,73	44,10
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	57,36	89,36	62,65	57,20	91,88	60,74	57,38	93,33	56,50
12	48,09	63,15	46,65	48,35	60,51	46,00	48,23	59,87	42,52
17	34,94	38,60	35,24	34,74	38,35	34,87	34,04	38,48	31,40
Média	46,80	63,70	48,18	46,76	63,58	47,20	46,55	63,89	43,47
Grau de sobreposição 2 nas categóricas									
18	50,00	70,17	55,40	50,00	76,23	55,02	50,00	83,54	52,64
22	34,66	56,36	43,85	34,64	56,27	43,73	34,63	55,68	41,62
25	20,00	37,35	32,15	20,00	37,65	32,11	20,00	37,50	31,00
Média	34,89	54,63	43,80	34,88	56,72	43,62	34,88	58,91	41,75
Grau de sobreposição 3 nas categóricas									
31	20,00	33,72	32,19	20,00	33,36	32,29	20,00	33,78	31,11

Tabela B.12: Taxa de alocação correta dos métodos segundo situações e disposição dos grupos, para peso maior para variáveis contínuas, correlacionadas, e modelo Uniforme

	Sobreposição de 20% nas contínuas			Sobreposição de 40% nas contínuas			Sobreposição de 80% nas contínuas		
Situação	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos	Ligação Média	*ROCK*	k-Protótipos
Grau de sobreposição 1 nas categóricas e 4 variáveis, 2 de cada tipo									
1	83,92	97,30	80,76	80,56	97,14	73,66	79,88	96,34	60,22
4	76,24	93,83	74,08	71,01	94,14	67,45	69,30	90,63	48,11
7	67,85	85,81	66,17	62,61	87,27	60,87	54,12	76,15	39,65
Média	76,00	92,31	73,67	71,39	92,85	67,33	67,77	87,70	49,32
Grau de sobreposição 1 nas categóricas e 8 variáveis, 4 de cada tipo									
9	90,76	97,39	92,13	84,33	96,81	84,07	81,13	95,08	62,65
12	88,51	94,25	90,68	79,63	93,93	83,97	71,29	86,50	52,11
17	92,62	87,72	83,60	82,19	89,31	80,75	58,77	71,93	47,56
Média	90,63	93,12	88,80	82,05	93,35	82,93	70,40	84,50	54,11
Grau de sobreposição 2 nas categóricas									
18	89,98	97,31	88,42	84,26	97,08	80,44	79,68	96,00	60,21
22	86,96	94,37	88,11	76,06	95,15	79,10	67,93	90,58	49,24
25	82,93	88,26	83,05	72,48	89,85	78,80	50,23	78,36	43,70
Média	86,62	93,31	86,53	77,60	94,03	79,45	65,95	88,31	51,05
Grau de sobreposição 3 nas categóricas									
31	66,28	86,77	78,28	55,68	89,97	69,26	31,63	83,49	35,93

Livros Grátis

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

Milhares de Livros para Download:

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