Comparação e escolha de agrupamentos: uma proposta utilizando a entropia

Estêvão Freitas de Souza

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS.

Área de Concentração: **Estatística**
Orientadora: **Profa. Dra. Viviana Giampaoli**

Durante a elaboração deste trabalho o autor recebeu apoio financeiro parcial da CAPES e CNPq

– São Paulo, maio de 2007 –

Comparação e escolha de agrupamentos: uma proposta utilizando a entropia

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Estêvão Freitas de Souza e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Profa. Dra. Viviana Giampaoli (Orientadora) - IME-USP
- Prof. Dr. Júnior Barrera - IME/USP
- Prof. Dr. Alejandro César Frery Orgambide - UFAL

"Ter não é ser."

domínio público

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

"Algorithms for Clustering Data", Jain & Dubes

"O seu tempo é limitado, então não o gaste vivendo a vida de outra pessoa. Não seja amarrado pelo dogma, que é viver com os resultados do pensamento de outra pessoa. Não deixe que o barulho da opinião dos outros emudeça sua voz interior. Tenha a coragem de seguir seu coração e sua intuição."

Steve Jobs

Agradecimentos

Gostaria de expressar meus agradecimentos:

- aos meus pais pela ajuda e incentivo sempre dados aos meus estudos e a minha vida;

- aos meus irmãos que também estão sempre aí, prontos para tudo;

- à minha inigualável namorada Hideko, tão compreensiva e paciente que merece a co-autoria espiritual deste trabalho;

- a minha orientadora Profa. Viviana Giampaoli, uma privilegiada que possui em seu DNA os genes da paciência e dedicação orientais no trabalho educacional, tão úteis na formação das novas gerações;

- ao Prof. Armando Infante (UNICAMP), Mestre-Guia dos meus primeiros passos acadêmicos, sempre com incentivos e ótimos conselhos, mesmo depois de 10 anos !

- a nossa "tchurma" do mestrado: Germán, Rafael, Alexandre, Dalton, Démerson, Luz Marina, Lauren, Camila etc (vocês não acharam que eu iria botar aqui os nomes da galera toda, né ?!);

- aos outros inúmeros novos amigos que pude fazer no IME e na USP: Núbia, Juvêncio, Caio, Alexandre, Gustavo, Diana etc etc (vide item anterior);

- um profundo OBRIGADO a Amy V. Kapp (http://www.stanford.edu/akapp/), aluna de doutorado do Prof. Tibshirani, em Stanford, que muito gentilmente cedeu os códigos em R para implantação da estatística Gap;

- aos funcionários da biblioteca do IME e da secretaria da CPG: esse pessoal é 10 !

- um agradecimento especial ao amigo, professor e torcedor caldense Edson Martinez (ϵ) que desde o início deste projeto sempre se mostrou pronto a me ajudar em tudo;

- a CAPES e CNPq, pelo apoio financeiro;

- a Eumir Deodato, Tom Jobim e Moacir Santos: a quem não tenho palavras, só ouvidos!

- a Deus, pela paz, por tudo isso aí em cima e muito, MUITO mais !

Resumo

A análise de agrupamentos (*cluster analysis*) é o conjunto de ferramentas estatísticas de análise multivariada para encontrar ou revelar a existência de grupos em uma amostra. A literatura apresenta muitos métodos para particionar um conjunto de dados. Porém, ao utilizá-los, o pesquisador muitas vezes se depara com o problema de decidir em quantos grupos deverá ser feita essa divisão, bem como comparar agrupamentos obtidos por diferentes métodos estabelecendo quão semelhantes eles são.

Neste trabalho é feita uma revisão dos principais métodos de comparação de agrupamentos e é apresentada uma nova técnica para a escolha do número ideal de grupos, baseada na diferença de entropias. Afim de avaliá-la, estudos de simulação foram realizados comparando-a com outras técnicas conhecidas: a estatística Gap e a silhueta média. Os resultados indicaram que a nova proposta é tão ou mais eficiente que as demais, no sentido de encontrar o número correto de grupos. Além disso, ela também é computacionalmente mais rápida e de simples implementação. Duas aplicações a dados reais são apresentadas, ambas na área de genética.

Palavras-chave: análise de agrupamentos, comparação de agrupamentos, análise multivariada

Abstract

Cluster analysis is the set of multivariate statistical techniques to uncover or discover groups in a sample. There's plenty of methods in the literature to partition a dataset. But, when doing so, the user is frequently faced with the problem of choosing the appropriate number of groups and, also, how to compare clusterings obtained through different methods and establish how similar they are.

In the present work, it is presented a revision of methods to compare clusterings and proposed a new technique to choose the appropriate number of groups, based on the difference of entropies. To evaluate it, a simulation study was made comparing it with other already known techniques: the Gap statistic and the silhouette. The results indicated that the new approach is more or as efficient as the others, in the sense of finding the correct number of clusters. Moreover, it is computationally faster and simple to implement. Two application are shown, both in genetics.

Keywords: cluster analysis, comparing clusterings, multivariate analysis

Índice

Página

Capítulo 1

Introdução

A análise de agrupamentos (*cluster analysis*) compreende os métodos estatísticos utilizados para separar os indivíduos de uma amostra em classes sem que essas sejam pré-estabelecidas, e de tal modo que os indivíduos em uma mesma classe sejam similares entre si e diferentes daqueles pertencentes às outras classes, de acodo com as variáveis observadas. Esta técnica é amplamente usada em várias áreas de estudo, sendo uma ferramenta muito útil na construção de sistemas de classificação. O presente trabalho discute a questão da comparação dos agrupamentos e da escolha do número mais apropriado de grupos que se pode construir com uma amostra.

A dissertação possui outros seis capítulos, além deste. No capítulo 2 é introduzido o conceito da análise de agrupamentos e uma breve visão dos métodos estatísticos disponíveis para sua utilização. Além disso, coloca-se o problema da comparação de agrupamentos e escolha do melhor número de grupos. No capítulo 3 é apresentada uma revisão bibliográfica das propostas para resolução dos problemas citados. No capítulo 4 são introduzidos alguns conceitos da Teoria da Informação, e a partir dos mesmos discute-se o método de comparação denominado *Variação da Informação* (Meilă 2002). Neste capítulo também é proposto um novo critério para escolha do melhor número de grupos utilizando-se a diferença das entropias condicionais de dois agrupamentos. No capítulo 5, estão os resultados dos estudos de simulação realizados com o intuito de avaliar e comparar as técnicas de escolha do número ótimo de grupos descritas nos capítulos anteriores: a diferença de entropias condicionais, estatística Gap (Tibshirani, Walther & Hastie 2001) e silhueta (Rousseeuw 1987). Duas aplicações são mostradas no capítulo 6 e, para encerrar o trabalho, apresenta-se as considerações finais no capítulo 7.

Capítulo 2

A análise de agrupamentos

2.1 Considerações iniciais

Uma das habilidades mais básicas do ser humano envolve agrupar objetos similares para produzir classificações. Desde pequena, uma pessoa se vê envolvida num contínuo processo de aprendizagem através de esquemas de classificação, por exemplo, quando aprende a distinguir cães de gatos, cadeiras de mesas, homens de mulheres.

No dia-a-dia, é possível encontrar exemplos de classificação por toda parte. Um cardápio de restaurante geralmente é dividido nos grupos: entradas, saladas, pratos principais, sobremesas e bebidas, um padrão lógico para nossos costumes. Um cobrador de ônibus organiza o dinheiro do caixa em divisórias para notas de 1, 2, 5, 10 e as moedas. Isso o ajuda no trabalho de receber e dar o troco.

Na ciência, Aristóteles (384-322 A.C.) promoveu um elaborado sistema para classificar as espécies animais, dividindo-os inicialmente segundo a presença ou não de sangue vermelho, depois conforme a forma de reprodução etc. Aristóteles posteriormente estendeu o sistema à botânica e contribuiu para o desenvolvimento desta ciência até ser suplantado no século XVIII pelos trabalhos do sueco Lineu (1707-1778) e do francês Jussieu (1748-1836), ambos botanistas. Seus novos sistemas de classificação perduram até os dias de hoje e permitem acomodar o contínuo aumento do conhecimento das espécies vegetais, promovendo uma visão organizada do saber.

Outro exemplo bastante contundente de classificação científica é a tabela periódica, proposta por Mendeleyev (1834-1907). Nela, os elementos aparecem agrupados segundo suas propriedades químicas e físicas. O uso desse sistema facilitou a descoberta de elementos até então desconhecidos pelo homem.

Assim, no processo de construção do conhecimento, observar características dos objetos e agrupá-los segundo as mesmas parece ser uma atividade inata do ser humano.

Como explicam Everitt, Landau & Leese (2001), dependendo da área que se estuda, o processo de classificação recebe diferentes nomes. Na biologia o termo mais usado é "taxonomia numérica". Na psicologia tem-se a "análise Q". Os profissionais de publicidade aplicam o termo "segmentação". Do ponto de vista estatístico, obter agrupamentos das unidades amostrais segundo seus atributos observados é considerado um método multivariado denominado análise de agrupamentos. Nesse ponto, como coloca Gordon (1981), é importante estabelecer a distinção entre a análise de agrupamentos e outras situações parecidas, porém distintas.

Por vezes, deseja-se distribuir os indivíduos de uma amostra em um conjunto conhecido de categorias. Assim, supõe-se que eles foram retirados de uma população onde todos os elementos podem ser classificados nessas categorias. A esse processo dá-se o nome de 'identificação'. Um exemplo dessa situação ocorre quando se coleta um espécime de algum ser vivo e determina-se seu reino como vegetal ou animal.

Em outras situações, pode-se dividir uma amostra em grupos sem que seus elementos precisem ser semelhantes segundo os atributos observados. Ou seja, não há preocupação em buscar alguma estrutura presente na amostra. Por exemplo, os imóveis de uma cidade são divididos em bairros para facilitar a localização geográfica. Nesse caso a estrutura é imposta à amostra para, por exemplo, entregar a correspondência. Esse processo é denominado 'dissecação'. A análise de agrupamentos pode ser empregada como parte do processo de dissecação, mas não necessariamente isso sempre ocorre.

É preciso acrescentar ainda que na análise de agrupamentos não há a preocupação em derivar, a partir de uma amostra inicial, regras que permitam alocar novas amostras aos grupos então construídos. Essa situação é mais característica de outra técnica multivariada denominada 'análise discriminante'.

Para se construir agrupamentos com bancos de dados de uma ou duas dimensões, geralmente emprega-se gráficos como o de dispersão e o histograma. Um exemplo é mostrado na figura 2.1, que utiliza os dados originalmente publicados por Ruspini (1970), com 75 observações e duas variáveis, disponíveis no pacote *cluster* do programa R (http://www.r-project.org/). Com apenas uma inspeção visual do gráfico de dispersão (figura 2.1(a)), pode-se propor a divisão dos dados em quatro grupos, segundo as variáveis X e Y. Entretanto, nos histogramas onde se apresentam as distribuições univariadas de X e Y, respectivamente figuras 2.1(b) e 2.1(c), não é possível identificar os quatro grupos, sendo imprescindível utilizar simultaneamente a informação de ambas variáveis para se obter a divisão citada.

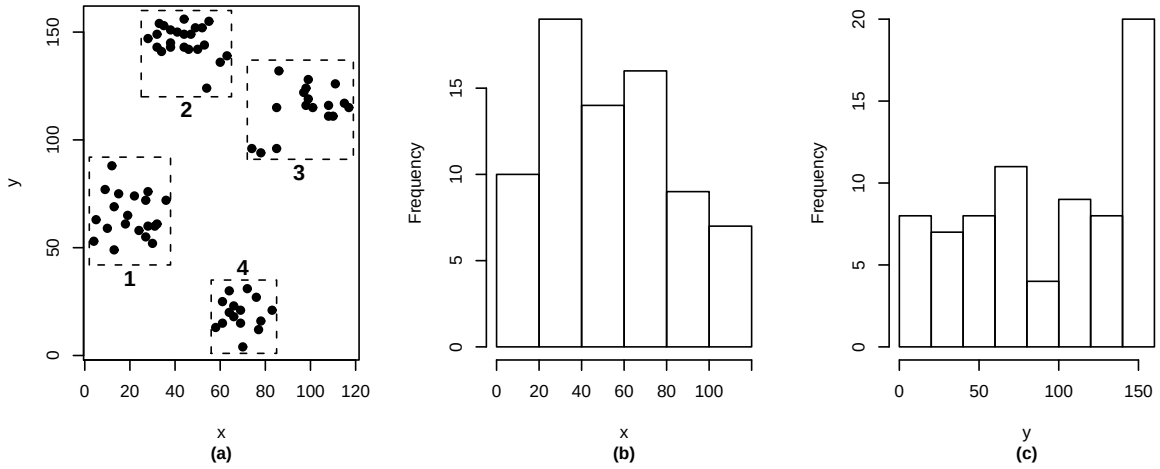

Figura 2.1: (a) Gráfico de dispersão, com quatro grupos identificados visualmente, (b) histograma da variável X, e (c) histograma da variável Y.

Entretanto, com a crescente facilidade em mensurar e armazenar informações, bancos de dados com mais observações e variáveis tornaram-se comuns. Nesses casos torna-se difícil o emprego de estratégias puramente visuais, como as acima citadas. Faz-se necessário, então, o uso de métodos analítico-computacionais que levem em conta essa estrutura multivariada. Nesse novo contexto, para o desenvolvimento de algoritmos que sirvam à construção dos agrupamentos, é imperativo definir-se claramente o que é um *grupo*. Mas, como colocam Kaufman & Rousseeuw (1990), não existe uma definição única e clara, e o que se encontra na literatura são 'tipos de grupos': esféricos, alongados, lineares etc, que acompanham as diferentes necessidades encontradas na aplicação da análise dos dados. Devido a isso, nos últimos 40 anos houve um abundante desenvolvimento de técnicas para realizar a análise de agrupamentos, afim de acomodar a diversidade de estruturas que se pode definir com a expressão 'tipos de grupos'. Além disso, diversos podem ser os critérios de distinção numérica entre os elementos da amostra, ou grupos destes, ou seja, a medida de distância entre eles, fator que também deve ser levado em conta no agrupamento. Vale ainda citar dois desideratos básicos para um agrupamento, enunciados por Cormack (1971): coesão interna e isolação externa, ou seja, elementos dentro de um mesmo grupo devem estar tão próximos quanto possível entre si, e os que estão em grupos diferentes devem ter sua distância maximizada.

Dentre essa abundância de técnicas analíticas para construir agrupamentos, a maioria pode ser classificada em três categorias: hierárquicas, de partição e *clumping*. Não faz parte dos objetivos do presente trabalho descrever detalhadamente cada uma delas, apenas uma

visão geral será oferecida. A literatura dispõe de muitas opções para consulta sobre o assunto, sendo duas excelentes referências Hartigan (1975) e Kaufman & Rousseeuw (1990).

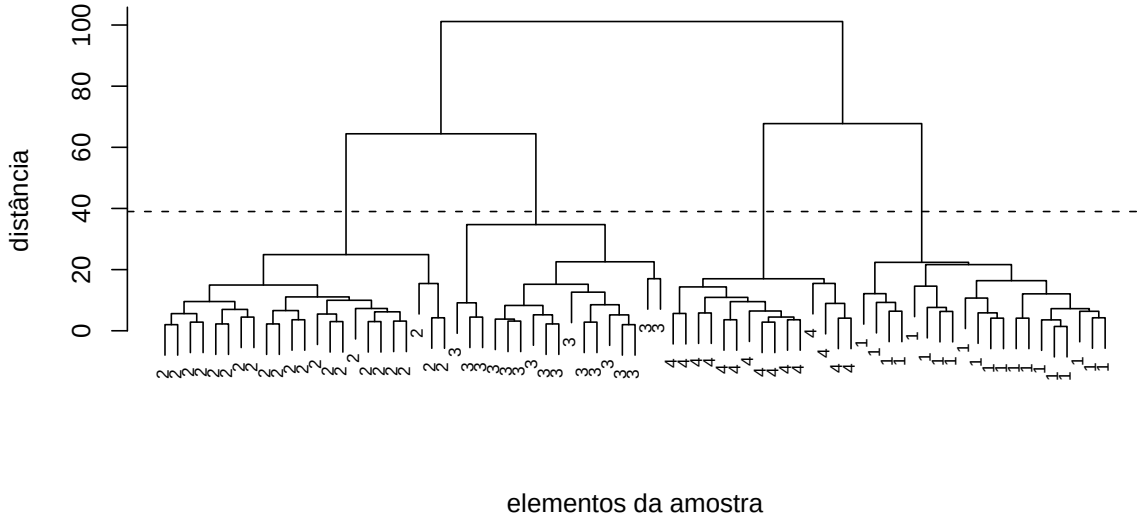

Figura 2.2: Dendrograma para dados de Ruspini.

Nos métodos hierárquicos, também chamados de não-supervisionados, há basicamente dois grupos de algoritmos. O primeiro deles é chamado de *aglomerativo*, em que têm-se um processo iterativo iniciado com um agrupamento onde cada unidade amostral é um grupo isolado composto, portanto, de um único elemento. O número de grupos é reduzido a cada passo com a união de elementos da amostra, ou grupos dos mesmos formados nas iterações anteriores, segundo as distâncias calculadas entre eles. Como consequência, ao final há apenas um único grupo que contém toda a amostra. Deste modo, a cada passo, esses algoritmos criam grupos dentro de grupos, o que impõe uma estrutura hierárquica, advindo daí a denominação desses métodos. Essa característica permite que se faça um gráfico denominado "dendrograma" (do grego: *dendron*=árvore e *grama*=lista ou registro) onde no eixo das ordenadas lê-se a distância entre os grupos e no eixo das abscissas os elementos pertencentes a cada grupo. Um exemplo de dendrograma é mostrado na figura 2.2, no qual foram utilizados novamente os dados de Ruspini (1970). Os 75 elementos da amostra aparecem representados pelos número de 1 a 4, correspondentes à identificação dos grupos proposta na figura 2.1(a). À medida em que cresce o valor no eixo das ordenadas, esses elementos são unidos com linhas representando os pares com a menor distância entre si, reduzindo o número de grupos. A linha pontilhada indica o local de 'corte' para se obter os quatro grupos citados. O segundo grupo de algoritmos hierárquicos é chamado de *divisivo*. A diferença é que se faz o caminho inverso, ou seja, inicia-se o processo iterativo considerando

a amostra como um único grupo e daí realizam sucessivas subdivisões a cada passo.

Um dos pontos fracos dos métodos hierárquicos é seu alto custo computacional, crescente à medida que aumenta-se o tamanho da amostra. De fato, como é explicado no manual disponível na internet para o programa SAS, versão 8 (http://v8doc.sas.com/sashtml/), a maioria dos algoritmos implementados nesse programa para a construção de agrupamentos hierárquicos não é adequada para bancos de dados muito grandes, pois o tempo de uso da CPU varia com o quadrado ou o cubo do número de observações.

Na categoria dos métodos de partição uma grande diferença em relação aos hierárquicos é que o número de grupos do agrupamento deve ser informado ao computador antes da execução do algoritmo. Este, ao iniciar-se, escolhe as coordenadas de um 'centróide' inicial para cada grupo, e em torno deles são distribuídas as unidades amostrais, segundo algum critério de distância escolhido. Posteriormente, através de sucessivas iterações, os elementos são realocados para outros grupos num processo que busca minimizar ou maximizar alguma medida de qualidade do agrupamento. Esse processo de realocação, a escolha dos centróides e a medida de qualidade adotada caracterizam o algoritmo de partição. O critério de parada do processo iterativo pode ser um número máximo de iterações ou um limite mínimo de variação na medida de qualidade adotada. O principal representante dessa categoria é o método das k-médias. Pelo fato de haver a escolha antecipada do número de grupos estes métodos também são chamados de 'supervisionados'

Poderia-se argumentar que, para um determinado número de grupos, basta construir todos os agrupamentos possíveis com a amostra e escolher aquele com melhor medida de qualidade. Entretanto, como colocam Bussab, Andrade & Miazaki (1990), o número de combinações é assintoticamente igual a k^{n-1}, com k igual ao número de grupos e n o tamanho da amostra. Por exemplo, se $n = 20$ e $k = 3$, o número possível de agrupamentos ultrapassa 1 bilhão (3^{19}), inviabilizando completamente essa possibilidade! Segundo Everitt et al. (2001), o agrupamento final obtido é altamente dependente dos centróides iniciais escolhidos e, como não se pode testar todas as combinações, essa escolha é um dos aspectos críticos dos métodos supervisionados.

Por último tem-se os métodos "clumping", cuja principal característica está em permitir a um elemento da amostra pertencer a dois ou mais grupos simultaneamente. Nos algoritmos hierárquicos e de partição essa condição não é possível. Mas devido a questões de desenvolvimento teórico, como será mostrado na seção 3.1, o presente trabalho imporá a restrição de que os grupos formados sejam conjuntos mutuamente exclusivos entre si. Assim, não serão dados maiores detalhes sobre os métodos "clumping". Duas referências para os mesmos são Everitt et al. (2001) e Gordon (1981).

Como se pode ver, a análise de agrupamentos é uma metodologia cujas aplicações podem ser em diversas áreas. Mais do que um fim em si mesma, seu uso deve ser considerado dentro do escopo da análise explorátória de dados. Com tantas nuances nas suas características e possibilidades de aplicação, encerra-se o capítulo com a definição enunciadada por Kaufman & Rousseeuw (1990): "análise de agrupamentos é a arte de encontrar grupos em um banco de dados."

2.2 Número de grupos e a comparação de agrupamentos

Como já citado, o número de métodos disponíveis para a obtenção de um agrupamento é vasto. Porém, a simples escolha de um deles não esgota as decisões a serem tomadas. Ainda resta a escolha do número de grupos para se construir. Ou, posto de outra forma, qual o número ótimo de grupos que deverá conter o agrupamento construído ao final do processo ?

Na comunidade científica, segundo Everitt (1979), uma das primeiras tentativas de solucionar esse problema foi proposta por Thorndike (1953). Ele sugeriu observar descritivamente, em um gráfico de dispersão, o descrescimento da média das somas de distâncias quadráticas intra-grupos versus o número de grupos, e procurar o ponto onde ocorre uma queda brusca, o que seria um indicativo do número ótimo de grupos. Thorndike fez algumas tentativas de demonstrar sua proposição com algumas simulações mas não obteve sucesso. Entretanto, suas idéias, incluindo a sugestão de se desenvolver um teste de significância estatística para a 'queda brusca', abriram caminho para o surgimento de outros métodos, como será visto na seção 3.5. Além disso, outros pesquisadores desenvolveram soluções para o problema da escolha do número ótimo de grupos utilizando abordagens diferentes de Thorndike. Uma delas é a silhueta, exposta na seção 3.4. Várias outras existem e sugere-se a consulta de Milligan & Cooper (1985) e Kapp & Tibshirani (2007), onde os autores, através de simulações, avaliaram vários desses métodos e construíram uma classificação ordenada dos mesmos. Entretanto, como explica Milligan (1996), resultados dessa natureza devem ser usados com cautela pois tratam-se de estudos únicos de comparação, sem nenhum tipo de validação independente. Portanto, ele recomenda o uso simultâneo de dois ou três dos métodos melhor cotados. No caso de coincidência dos seus resultados, deve-se concluir que há evidência forte para o uso do número de grupos indicado. Em caso de concordância parcial, deve-se optar pelo maior número de grupos indicado e aguardar estudos futuros para decidir se alguns deles podem ser unidos, reduzindo-se sua quantidade e simplificando o agrupamento. Já se não houver nenhuma consistência nos resultados, nenhuma das opções deverá fornecer uma interpretação clara da amostra ou senão, o pesquisador deverá conside-

rar fortemente a hipótese de não haver um agrupamento razoável para os dados observados. Idêntica recomendação é feita por Everitt et al. (2001), que ainda dedica uma seção do último capítulo de seu livro para discutir a situação da ausência de estrutura nos dados e dar várias referências sobre o assunto.

É sempre recomendável ter em mente o conselho de Thorndike (1953): considerações práticas devem ser levadas em conta na decisão do número ótimo de grupos. Por exemplo, uma empresa que decide segmentar uma carteira de clientes pode, por limitações de ordem física e de pessoal, impor um limite máximo de cinco grupos. Nessas condições, torna-se inútil concluir pela recomendação de um número superior de classes pois a aplicação não será possível, ainda que algum índice mostre que isso é "estatisticamente melhor".

Outra questão que se coloca diz respeito à comparação entre dois agrupamentos ou duas classificações diferentes, sem que haja a necessidade de escolher entre uma delas. Nesses casos a intenção é obter algum índice que indique a semelhança (ou diferença) entre duas formas de segmentar uma amostra. É claro que quando se define o número ótimo de grupos há também um processo de comparação envolvido, porém não existe necessariamente a preocupação de dizer o quão parecidos são os agrupamentos. Daí a diferença entre os dois problemas colocados. Fowlkes & Mallows (1983) citam que situações assim podem ocorrer quando se deseja estudar, por exemplo, o efeito de dois algoritmos diferentes na construção dos agrupamentos. Outra possibilidade é lembrada por Everitt et al. (2001): em estudos de robustez, o pesquisador pode estar interessado em comparar entre si agrupamentos obtidos pelo mesmo método através de sucessivas reamostragens do mesmo banco de dados. Outra situação é avaliar esses agrupamentos contra uma classificação pré-existente e conhecida, dentro de um processo de validação. Para este problema de comparação a literatura também oferece várias opções e Meilă (2005) coloca que não há claramente um critério superior aos demais, sendo o melhor aquele mais adequado às características do problema sob investigação. Além disso, ela também discute propriedades desejáveis para um critério dessa natureza e avalia alguns dos métodos existentes sob tais condições.

Por último, é importante ressaltar que a revisão de métodos apresentada nos capítulos 3 e 4 não esgota a extensa literatura disponível para a solução dessas duas questões, sendo fontes de consultas adicionais as referências já citadas acima.

Capítulo 3

Comparação e escolha de agrupamentos

3.1 Introdução

Antes de iniciar a revisão dos métodos propriamente dita, a notação será apresentada afim de facilitar as discussões posteriores.

Considere A o conjunto dos índices de uma amostra de n observações e r variáveis. Seja o agrupamento $\boldsymbol{C}$ uma partição dos indivíduos de A em K grupos, C_1,...,C_K, cada um com n_k elementos, $n_k \geq 1$, $k = 1, \ldots, K$, tal que $C_i \cap C_j = \emptyset, i, j = 1, \ldots, K, i \neq j$ e $\bigcup_{k=1}^{K} C_k = A$. Sob as mesmas restrições, considere ainda uma segunda partição $\boldsymbol{C'}$ de A, com K' grupos, cada um com $n'_{k'}$ elementos, $n'_{k'} \geq 1$, $k' = 1, \ldots, K'$, e K e K' podem ser diferentes.

Seja $\mathbf{D}$ uma matriz quadrada nxn onde o d_{ab}-ésimo elemento representa a parecença entre duas unidades amostrais a e b de A. A parecença é um critério de distinção numérico entre elas e pode ser uma medida de similaridade (o quão parecidos são) ou de dissimilaridade (quão diferentes são), esse último muitas vezes referido genericamente como 'distância'. Um exemplo do primeiro caso é o coeficiente de correlação e do segundo a distância euclidiana. Neste texto, quando necessário, far-se-á a distinção entre eles, caso contrário será utilizado o termo parecença. A $\mathbf{D}$ dá-se o nome de matriz de parecenças. Uma boa revisão dessas medidas encontra-se em Bussab et al. (1990).

Considere $\mathbf{M}$ uma tabela de contingência $K \times K'$, associada aos agrupamentos $\boldsymbol{C}$ e $\boldsymbol{C'}$, em que a $m_{kk'}$-ésima casela é a quantidade de pontos de A que está simultaneamente nos grupos C_k e $C'_{k'}$. As marginais de $\mathbf{M}$ são os valores de n_k e $n'_{k'}$. Na literatura, $\mathbf{M}$ é chamada de matriz de confusão (*confusion matrix*) ou de proximidades (*proximity*).

3.2 Comparação com o uso de pares de pontos da amostra

Vários dos métodos de comparação de agrupamentos encontrados na literatura baseiam-se nos pares de pontos de A. Considere a e b dois indivíduos de A, o total de pares de pontos (a,b), assumindo $(a,b)=(b,a)$, que se pode obter é $\frac{n(n-1)}{2}$. Eles podem divididos nos conjuntos abaixo, que não possuem intersecção e cuja soma é o próprio total de pares:

N_{11}: n^o de pares de pontos que estão no mesmo grupo em $\boldsymbol{C}$ e $\boldsymbol{C'}$;
N_{00}: n^o de pares de pontos que estão em grupos diferentes em $\boldsymbol{C}$ e $\boldsymbol{C'}$;
N_{10}: n^o de pares de pontos que estão no mesmo grupo em $\boldsymbol{C}$ e diferentes em $\boldsymbol{C'}$;
N_{01}: n^o de pares de pontos que estão no mesmo grupo em $\boldsymbol{C'}$ e diferentes em $\boldsymbol{C}$.

A obtenção dessas quantidades pode ser feita através de uma contagem simples no banco de dados, uma vez disponíveis as alocações dos indivíduos de A em $\boldsymbol{C}$ e $\boldsymbol{C'}$. Nos casos onde somente a matriz $\mathbf{M}$ está disponível, Hubert & Arabie (1985) oferecem as fórmulas abaixo, baseadas somente nas suas marginais:

$$N_{11} = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}\left(n_{ij}-1\right), \quad N_{00} = \frac{1}{2}\left(n^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right)\right),$$

$$N_{10} = \frac{1}{2}\left(\sum_{i=1}^{K} n_{i\cdot}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \text{ e } N_{01} = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right).$$

Alguns índices derivados destas quantidades são apresentados na Tabela 3.1.

Em seu artigo original, Rand (1971) descreveu a idéia intuitiva para criar o índice (1), simétrico, e diz que o mesmo varia de 0 (agrupamentos totalmente diferentes) a 1 (idênticos). Ele o utilizou para comparar dois métodos de construção de agrupamentos através de simulações de Monte Carlo. Para tanto, ele calculou as médias e desvios-padrões de $R\left(\boldsymbol{C},\boldsymbol{C'}\right)$ e, com eles, avaliou se os dois métodos eram ou não semelhantes. Entretanto, essa comparação foi apenas descritiva.

Nos anos posteriores à publicação do índice de Rand, outros autores sugeriram modificações para aperfeiçoá-lo. Uma delas foi feita por Fowlkes & Mallows (1983), que propuseram o índice (4). Ele também varia de 0 (agrupamentos diferentes) a 1 (idênticos), e pressupõe duas hipóteses:(a) os dois agrupamentos foram obtidos de forma independente, e (b) a respectiva tabela $\mathbf{M}$ observada é uma das possibilidades dentre todas as distribuições possíveis

Tabela 3.1: Índices baseados na classificação de pares de pontos de A

Índice	**Autor**	**Fórmula**
1	Rand	$R\left(\boldsymbol{C},\boldsymbol{C'}\right)=\frac{N_{11}}{N_{11}+N_{10}+N_{01}+N_{00}}$
2	Wallace	$W_I\left(\boldsymbol{C},\boldsymbol{C'}\right)=\frac{N_{11}}{\frac{\sum_{k=1}^{K} n_k\left(n_k-1\right)}{2}}$
3	Wallace	$W_{II}\left(\boldsymbol{C},\boldsymbol{C'}\right)=\frac{N_{11}}{\frac{\sum_{k=1}^{K'} n'_{k'}\left(n'_{k'}-1\right)}{2}}$
4	Fowlkes e Mallows	$F\left(\boldsymbol{C},\boldsymbol{C'}\right)=\sqrt{W_I\left(\boldsymbol{C},\boldsymbol{C'}\right)\times W_{II}\left(\boldsymbol{C},\boldsymbol{C'}\right)}$
5	Jacard	$J\left(\boldsymbol{C},\boldsymbol{C'}\right)=\frac{N_{11}}{N_{11}+N_{10}+N_{01}}$
6	Mirkin	$M\left(\boldsymbol{C},\boldsymbol{C'}\right)=2\left(N_{10}+N_{01}\right)$

dos elementos de A, considerando-se as marginais fixas e $K = K'$. A partir disso, pode-se calcular a esperança e variância de $F\left(\boldsymbol{C},\boldsymbol{C'}\right)$. Fowlkes & Mallows sugeriram utilizar o índice para comparar dois dendrogramas obtidos via diferentes métodos, para uma mesma amostra A. Assim, se $\boldsymbol{C_2},\boldsymbol{C_3},\ldots,\boldsymbol{C_K},\boldsymbol{C'_2},\boldsymbol{C'_3},\ldots,\boldsymbol{C'_K}$ são, respectivamente, os agrupamentos com 2, 3,..., K grupos realizando-se os cortes apropriados nos dois dendrogramas, deve-se calcular os valores de $F\left(\boldsymbol{C_2},\boldsymbol{C'_2}\right),\ldots,F\left(\boldsymbol{C_K},\boldsymbol{C'_K}\right)$ e colocá-los no eixo das ordenadas de um gráfico de dispersão, contra os valores de K nas abscissas. Além disso, deve-se acrescentar as bandas $E\left(F\right)\pm 2\times\sqrt{Var\left(F\right)}$ e, segundo os autores, valores de $F\left(\boldsymbol{C_i},\boldsymbol{C'_i}\right)$, i=2,...,K, fora desses limites indica que a similaridade dos dois agrupamentos para aquele i pode ser significante. Eles não explicaram a origem destas bandas mas justificam seu uso através de várias simulações descritas no artigo, baseadas em dados oriundos da distribuição Normal. Na prática, as hipóteses (a) e (b) dificilmente são respeitadas. Se $\boldsymbol{C}$ e $\boldsymbol{C'}$ foram construídos com a mesma amostra, a independência entre eles é questionável. Além disso, conforme Meilă (2002), pode-se encontrar métodos de construção de agrupamentos que usem K como informação inicial, mas não se tem notícia de algum que utilize como parâmetro a quantidade de pontos em cada grupo, tornando difícil aceitar a suposição de marginais fixas em $\mathbf{M}$.

Apesar de Fowlkes & Mallows (1983) terem restringido sua proposta a comparações entre agrupamentos com $K = K'$, Wallace (1983) mostrou que o desenvolvimento matemático por eles apresentado é facilmente adaptável para $K \neq K'$. Ele também mostrou que, na prática, o valor mínimo de $F\left(\boldsymbol{C},\boldsymbol{C'}\right)$ é sensível a K e n. Isso é facilmente percebido quando n é maior do que o número de caselas de $\mathbf{M}$, pois será impossível obter N_{11}=0 e, assim, o mínimo teórico $F\left(\boldsymbol{C},\boldsymbol{C'}\right)$=0 jamais será atingido. Meilă (2002) ainda acrescenta que se K for menor do que $n/3$, o mínimo varia bruscamente com o número de grupos, tornando comparações entre diferentes pares de $(\boldsymbol{C},\boldsymbol{C'})$ pouco úteis.

Wallace (1983) mostrou que o índice (4) pode ser visto como a média geométrica dos índices (2) e (3), e sugere utilizá-los conjuntamente com $F(\boldsymbol{C}, \boldsymbol{C'})$ em um gráfico de dispersão de seus valores contra K para melhor avaliar a semelhança dos agrupamentos construídos. Note que (2) e (3) representam, respectivamente, a proporção dos pares de pontos que estão em um mesmo grupo de $\boldsymbol{C}(\boldsymbol{C'})$, dado que estão no mesmo grupo de $\boldsymbol{C'}(\boldsymbol{C})$.

O índice de Jacard, (Ben-Hur, Elisseeff & Guyon 2002), e a métrica de Mirkin (Mirkin 1996) são outras formas modificadas de $R(\boldsymbol{C},\boldsymbol{C'})$. A eles também se aplica a discussão acima sobre valor mínimo e a sensibilidade a K e n. Meilă (2005) propôs dividir $M(\boldsymbol{C},\boldsymbol{C'})$ por n^2 para torná-lo invariante a n.

Hubert & Arabie (1985) derivaram fórmulas para cálculo da esperança de N_{11}, N_{10}, N_{01} e N_{00}, uma vez assumida a hipótese de marginais fixas em $\mathbf{M}$. Consequentemente, qualquer combinação linear delas pode ter sua esperança calculada. A partir daí eles discutem estratégias para corrigir o problema da variação do mínimo dos índices, sem apresentar, no entanto, uma solução definitiva.

3.3 Comparação pelo "melhor emparelhamento"

Outra classe de métodos de comparação baseia-se na idéia de cardinalidade dos conjuntos $\boldsymbol{C}$ e $\boldsymbol{C'}$. Na Tabela 3.2 são apresentados alguns desses índices.

Tabela 3.2: Índices baseados no melhor emparelhamento entre $\boldsymbol{C}$ e $\boldsymbol{C'}$

Índice	**Autor**	**Expressão**
7	Larsen	$L(\boldsymbol{C},\boldsymbol{C'}) = \frac{1}{K}\sum_{k=1}^{K} \max_{k'} \frac{2m_{kk'}}{n_k + n'_{k'}}$
8	Meilă e Heckerman	$H(\boldsymbol{C},\boldsymbol{C'}) = \frac{1}{n}\sum_{k'=match(k)} m_{kk'}$
9	van Dongen	$D(\boldsymbol{C},\boldsymbol{C'}) = 2n - \sum_{k=1}^{K} \max_{k'}\left(m_{kk'}\right) - \sum_{k'=1}^{K'} \max_{k}\left(m_{kk'}\right)$
10	Meilă	$CE(\boldsymbol{C},\boldsymbol{C'}) = 1 - \frac{1}{n}\max_{\sigma}\sum_{k=1}^{K} n_{k,\sigma(k)}\ , \quad K \leq K'$

Para todos eles, a cada grupo C_i de $\boldsymbol{C}$, i=1,...,K, é escolhido um par C'_j, j=1,...,K', como o melhor emparelhamento de $\boldsymbol{C'}$, e utilizada a correspondente m_{ij}-ésima casela de $\mathbf{M}$ nos cálculos dos índices (7), (8), (9) e (10).

Para (7) e (9), em cada grupo C_i de $\boldsymbol{C}$ o melhor emparelhamento C'_j será aquele com $\max_{k'}(m_{kk'})$. Fazendo-se o inverso obtém-se $\max_{k}(m_{kk'})$. No índice (8), $match(k)$ é encontrado tomando-se, em primeiro lugar, a maior casela de $\mathbf{M}$. Seja ela m_{ab}, então isso implica no par (C_a,C'_b). Em seguida, dentre os elementos restantes, e excluídos aqueles que não

pertençam nem à a-ésima linha nem à b-ésima coluna, escolhe-se o segundo maior, e assim sucessivamente até que min(K,K') emparelhamentos sejam feitos.

O índice (7), de Larsen & Aone (1999), representa uma generalização da medida F (*F-measure*)[1] , de van Rijsbergen (1980), aplicada sobre $\boldsymbol{C}$ e $\boldsymbol{C'}$. $L(\boldsymbol{C}, \boldsymbol{C'})$ foi proposta para uma aplicação no contexto de busca informatizada de documentos em uma base eletrônica de textos, utilizando-se palavras-chaves definidas pelo usuário como variáveis. Para avaliar o índice, Larsen & Aone (1999) utilizaram uma base de textos cujas classificações por assunto eram previamente conhecidas e realizaram sucessivas reamostragens. A cada uma delas eles construíram diversos agrupamentos, variando-se o método de obtenção dos mesmos e calcularam $L(\boldsymbol{C}, \boldsymbol{C'})$ em relação à classificação conhecida. Com uma análise apenas descritiva dos resultados, eles concluíram qual seria o melhor método de busca de documentos. Larsen & Aone (1999) não discutem a possibilidade de se utilizar $L(\boldsymbol{C}, \boldsymbol{C'})$ em outros contextos, nem derivam quaisquer propriedades matemáticas do mesmo.

Já o índice (8), de Meilă & Heckerman (2001), foi proposto para avaliar a eficácia de três algoritmos para construir agrupamentos que utilizam somente variáveis discretas. Os autores fizeram algumas simulações para testá-los variando-se a maneira de escolher seus centróides iniciais. Em seguida, apresentaram uma aplicação utilizando amostras de números escritos à mão em correspondências, digitalizados. Os métodos foram, nesse caso, comparados quanto a sua capacidade de separar em grupos as unidades amostrais com algarismos de interesse. Tanto nas simulações quanto na aplicação, utilizou-se apenas a análise descritiva das médias e desvios-padrões dos valores de $H(\boldsymbol{C,C'})$ para obter-se as conclusões.

Os índices $L(\boldsymbol{C,C'})$ e $H(\boldsymbol{C,C'})$ são assimétricos, com máximo em 1 (igualdade de agrupamentos). Seus autores não citam quais seriam seus valores mínimos.

$D(\boldsymbol{C,C'})$ foi proposto por van Dongen (2000) que o utilizou para avaliar um método de construção de agrupamentos para grafos. Um grafo é um conjunto de pontos, denominados *vértices*, conectados por linhas, chamadas de *arestas*. Dependendo da aplicação os vértices e arestas podem receber um peso numérico. Os grupos obtidos são subconjuntos de vértices do grafo. Problemas de interesse prático podem ser formulados como questões sobre certos grafos. Por exemplo, no artigo de van Dongen ele mostrou uma aplicação na área de análise de imagens. Uma fotografia em preto & branco foi digitalizada e cada

[1] Seja B um subgrupo de A, onde todos n_B elementos têm alguma característica de interesse W em comum. Após construir-se $\boldsymbol{C}$, calcula-se para cada um de seus K grupos as quantidades P e R como abaixo. P e R são chamadas precisão (*precision*) e recuperação (*recall*), e normalmente empregadas para avaliar a qualidade individual de um único grupo com respeito a W. Considerando N_i o número de elementos de C_i que possue W, a medida F é a média harmônica das quantidades P e R: $F = \frac{2PR}{P+R}$, com $P(W,i) = \frac{N_i}{|C_i|}$ e $R(W,i) = \frac{N_i}{n_B}$.

pixel representado por um vértice com peso igual ao seu tom de cinza, numa escala de 0 a 255. Cada um deles foi conectado aos vizinhos horizontais, verticais e diagonais por arestas cujo peso era inversamente proporcional à diferença nos tons de cinza dos dois *pixels* em questão. Os agrupamentos foram obtidos variando-se os parâmetros do método de construção e comparando-os com $D(\boldsymbol{C,C'})$. van Dongen mostrou que essa medida é uma métrica no espaço dos agrupamentos[2], com valor 0 se $\boldsymbol{C}=\boldsymbol{C'}$ e máximo estritamente menor que $2n$.

O índice (10) (Meilă 2005) chama-se *erro de classificação* (CE). O máximo da somatória é tomado sobre as possíveis combinações do mapa injetivo $\sigma(k)$ de $\{1,\ldots,K\}$ em $\{1,\ldots,K'\}$. Por exemplo, para k=1 este mapa será formado pelas correspondências entre C_1 e cada um dos $C'_{k'}$ que receberem uma ou mais de suas n_1 unidades amostrais. CE representa a menor porcentagem de indivíduos de A originalmente agrupados conforme $\boldsymbol{C}$ e que mudaram de classificação em $\boldsymbol{C'}$. Quando $\boldsymbol{C}=\boldsymbol{C'}$, ele atinge o valor mínimo igual a 0.

Os índices da Tabela 3.2 são todos insensíveis a certas mudanças na configuração de $\mathbf{M}$. Um exemplo é mostrado na Tabela 3.3(a). Os emparelhamentos para os grupos de $\boldsymbol{C}$ (caselas realçadas em negrito na tabela) pela regra dos índices $L(\boldsymbol{C,C'})$ e $D(\boldsymbol{C,C'})$ seriam (C_1, C'_3)=9, (C_2, C'_2)=10, (C_3, C'_2)=12 e (C_4, C'_5)=15, e para $H(\boldsymbol{C,C'})$ e $CE(\boldsymbol{C,C'})$ seriam (C_4, C'_5)=15, (C_3, C'_2)=12, (C_1, C'_3)=9 e (C_2, C'_1)=3. Alterando-se a configuração da tabela para a forma mostrada na Tabela 3.3(b), os melhores emparelhamentos permanecem os mesmos e, consequentemente, os valores dos índices (7), (8), (9) e (10) não se alteram. Meilă (2002) chama tal aspecto da falta de sensibilidade desta classe de métodos de "problema do emparelhamento" (*matching problem*).

Todos os índices desta seção foram propostos e aplicados pelos seus respectivos autores de maneira *ad hoc*, o que pode comprometer seu uso em outros contextos.

Tabela 3.3: Exemplo de melhores emparelhamentos iguais em tabelas diferentes.

	C'_1	C'_2	C'_3	C'_4	C'_5	total
C_1	6	5	**9**	1	4	25
C_2	**3**	**10**	4	0	1	18
C_3	0	**12**	6	3	5	26
C_4	2	3	1	1	**15**	22
total	11	30	20	5	25	78

(a)

	C'_1	C'_2	C'_3	C'_4	C'_5	total
C_1	3	8	**9**	4	1	25
C_2	**3**	**10**	5	0	0	18
C_3	0	**12**	5	0	9	26
C_4	5	0	1	1	**15**	22
total	11	30	20	5	25	78

(b)

[2]O espaço dos agrupamentos é formado por todas as possíveis combinações entre os n elementos de A para formar 1, 2, 3,..., n grupos.

3.4 Silhueta

O método das silhuetas foi proposto por Rousseeuw (1987) para uso em métodos de obtenção de agrupamentos por partição. A idéia é auxiliar o pesquisador a escolher o número ótimo de grupos e, ao mesmo tempo, permitir que se construa uma representação gráfica do agrupamento encontrado. A silhueta é composta por um gráfico ilustrativo de $\boldsymbol{C}$ contendo um índice $s(i)$, i=1,...,n, que reflete a qualidade da alocação da i-ésima unidade amostral ao seu grupo, permitindo uma visualização global da estrutura encontrada.

Para obter as silhuetas é necessário conhecer a distribuição de A nos grupos de $\boldsymbol{C}$ e a matriz $\mathbf{D}$, com as parecenças necessariamente numa razão de escala.

Inicialmente deve-se calcular o índice $s(i)$, i=1,...,n, para todos os elementos da amostra. No caso da parecença ser uma dissimilaridade, e assumindo-se $K \geq 2$, o procedimento é o seguinte:

- para cada elemento i de A calcula-se $a(i)$, a dissimilaridade média de i em relação aos indivíduos do mesmo grupo C_ℓ ao qual ele pertence:

$$a\left(i\right) = \sum_{j \in C_\ell} \frac{d\left(i,j\right)}{n_\ell};$$

- para cada grupo C_k ao qual i **não** pertença, calcula-se $b(i)$, sendo $d(i,C_k)$ a dissimilaridade média entre i e os elementos de C_k:

$$b(i) = \min_{i \notin C_k} \left[d(i, C_k)\right].$$

  Pode-se interpretar $b(i)$ como a distância entre i e o grupo vizinho mais próximo a ele em termos do critério de dissimilaridade utilizado;

- obtém-se os índices $s(i)$ como abaixo:

$$s\left(i\right) = \begin{cases} 1 - a\left(i\right)/b\left(i\right) & \text{se } a\left(i\right) < b\left(i\right), \\ 0 & \text{se } a\left(i\right) = b\left(i\right), \\ b\left(i\right)/a\left(i\right) - 1 & \text{se } a\left(i\right) > b\left(i\right), \end{cases} \Leftrightarrow \; s\left(i\right) = \frac{b\left(i\right) - a\left(i\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}} \; . \qquad (3.1)$$

A figura 3.1 ilustra um exemplo das distâncias consideradas no cálculo de $s\left(i\right)$ num agrupamento hipotético com quatro grupos C_1, C_2, C_3 e C_4. Com referência ao i-ésimo ponto, as linhas pontilhadas representam as distâncias consideradas para encontrar $a\left(i\right)$, e as contínuas para $b\left(i\right)$.

Para os casos onde i constitui-se por si só num grupo de elemento único, tornando-se impossível calcular $a(i)$, Rousseeuw (1987) recomenda assumir-se $s(i)$=0.

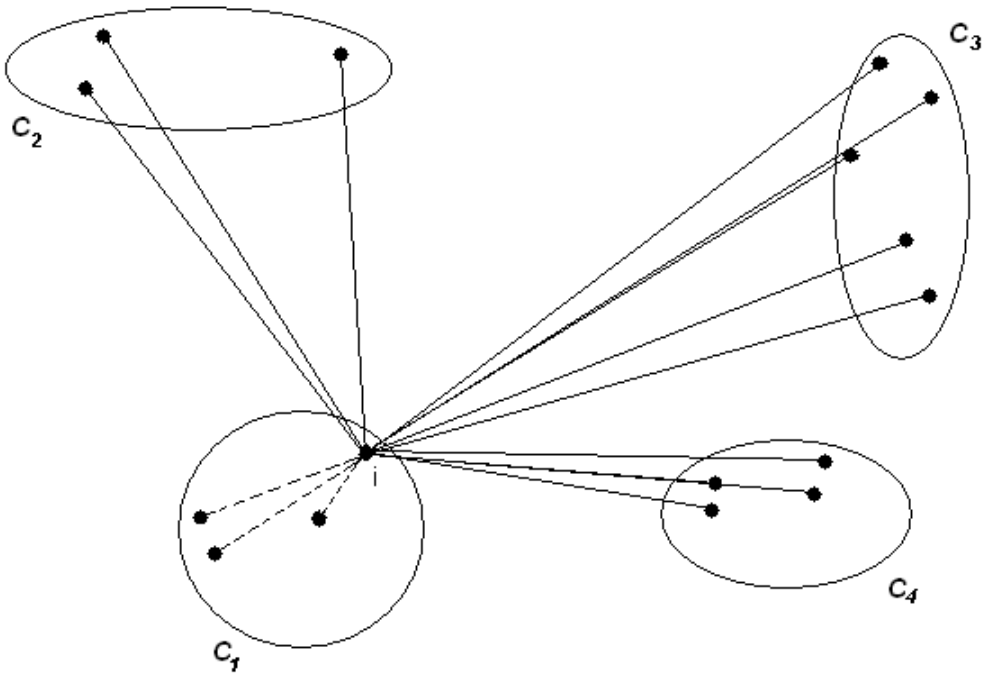

Figura 3.1: Agrupamento com linhas representando as distâncias do i-ésimo ponto aos demais elementos da amostra.

Observando-se (3.1), é direto concluir que $s(i)$ varia no intervalo [-1;1], além de ser adimensional. Um valor de $s(i)$ mais próximo de -1 é um indício de que o i-ésimo elemento foi provavelmente mal-classificado, pois neste caso $b\,(i) << a\,(i)$, ou seja, i, em média, está mais distante dos elementos do seu próprio grupo do que aquele usado no cálculo de $b\,(i)$. Por outro lado, se $s(i)$ é mais próximo de 1, pode-se supor que sua classificação foi adequada, pois $b\,(i) >> a\,(i)$. Um valor de $s(i)$ próximo ao zero ocorre quando $a(i)$ e $b(i)$ são semelhantes, indicando que o i-ésimo indivíduo está num ponto intermediário entre os dois grupos.

A silhueta média (SM) do j-ésimo grupo é dada por:

$$SM_j = \frac{\sum_{i=1}^{n_j} s(i)}{n_j}\ .$$

O coeficiente de silhueta médio (CSM) é um índice de qualidade para todo o agrupamento $\boldsymbol{C}$, dado pela média de $s\,(i)$:

$$CSM = \frac{\sum_{i=1}^{n} s(i)}{n} \tag{3.2}$$

A interpretação deste coeficiente é dada na Tabela 3.4. Kaufman & Rousseeuw (1990) explicam que os intervalos nos quais CSM foram então divididos são oriundos de sua experiência, não tendo sido aplicado nenhum tipo de critério de validação sobre os mesmos.

Tabela 3.4: Avaliação do agrupamento segundo *CSM*

CSM	**Interpretação sugerida**
0.71-1.00	Grupos encontrados possuem estrutura muito robusta
0.51-0.70	Grupos razoalmente unidos
0.26-0.50	A estrutura encontrada é fraca, tente outros métodos de agrupamento
≤ 0.25	Nenhuma estrutura encontrada

O cálculo de $s(i)$ quando as parecenças usadas forem similaridades requer algumas adaptações, afim de não modificar o limite de varição [-1;1] e nem a interpretação descrita no parágrafo anterior. Para tanto, define-se *a'*(i) e *d'*(i,C_k) como similaridades médias, calculadas de maneira análoga à $a(i)$ e $d(i,C_k)$, e faz-se:

$$b'(i) = \max_{i \notin C_k} \left[d'(i, C_k) \right].$$

Em seguida calcula-se $s(i)$ como abaixo:

$$s(i) = \begin{cases} 1 - b'(i)/a'(i) & \text{se } a'(i) > b'(i), \\ 0 & \text{se } a'(i) = b'(i), \\ a'(i)/b'(i) - 1 & \text{se } a'(i) < b'(i), \end{cases} \Leftrightarrow \; s(i) = \frac{a'(i) - b'(i)}{\max\{a'(i), b'(i)\}} \tag{3.3}$$

Finalmente, é mostrado como se constrói a silhueta com um exemplo.

A partir dos dados de Ruspini (1970), já usados no capítulo 2, construiu-se um agrupamento de quatro grupos utilizando-se o método de partição PAM (Kaufman & Rousseeuw 1990), o qual encontra-se implementado no pacote *cluster* do R. Como parecença foi utilizada a distância euclidiana.

Uma vez obtida a divisão da amostra nos quatro grupos, calculou-se $s(i)$, SM_j e CSM, i=1,. . . ,75, j=1,. . . ,4, e construiu-se o gráfico mostrado na figura 3.2. O pacote *cluster* possui implementada uma função para se construir o gráfico da silhueta.

Representa-se as unidades amostrais no eixo das ordenadas por barras cujo comprimento é igual aos seus respectivos valores de $s(i)$, que podem ser lidos no eixo das abscissas. Essas barras foram classificadas de acordo com os quatro grupos formados e, em cada um deles, estão em ordem descrescente de $s(i)$. À sua direita vê-se a identificação do grupo (letra j), a quantidade de elementos de cada um (n_j) e suas silhuetas médias ($\text{ave}_{i \in Cj} s_i$). No rodapé do gráfico aparece indicado o valor de $CSM = 0.74$. Os menores valores de $s(i)$ encontram-se no terceiro grupo, que também tem o mais baixo SM, igual a 0.67.

Para a escolha do número ótimo de grupos, Kaufman & Rousseeuw (1990) sugerem construir-se todos os respectivos agrupamentos com k=2, 3,. . . , n-1 grupos e escolher aquele

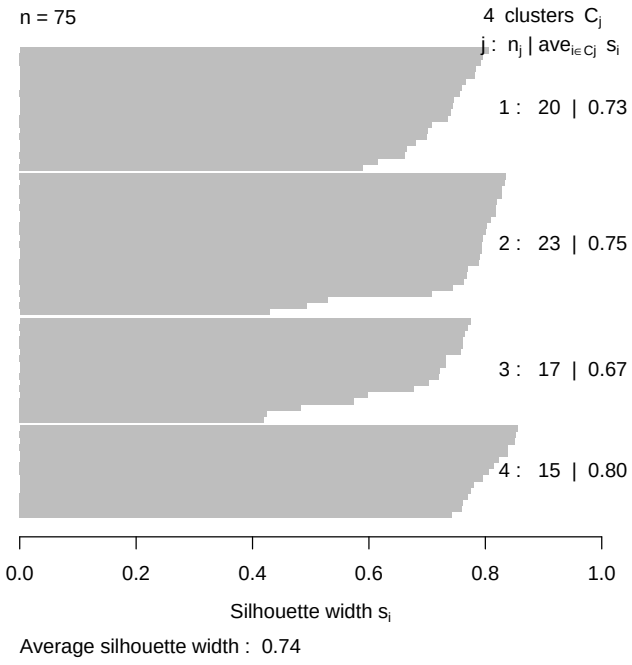

Figura 3.2: Silhueta para os dados de Ruspini, com K=4.

com o maior CSM. Em seguida, eles também recomendam avaliar o agrupamento eleito de acordo com a Tabela 3.4.

Apesar de Rousseeuw (1987) ter proposto a silhueta para aplicação a métodos supervisionados, Kaufman & Rousseeuw (1990) afirmam que a mesma pode ser estendida a qualquer método de obtenção de agrupamentos.

3.5 Método da Estatística Gap

Nesta seção será feita uma breve apresentação da estatística Gap, proposta por Tibshirani et al. (2001) para solucionar o problema da escolha do número ótimo de grupos.

Seja $W(\boldsymbol{C_k})$ uma medida de erro[3] do agrupamento $\boldsymbol{C_k}$ quando este possui k grupos, por exemplo como em (3.4). Segundo Tibshirani et al. (2001), há uma heurística estatística sob a qual $W(\boldsymbol{C_k})$ seria monótona decrescente à medida em que k aumentasse, e a partir de um

[3]O termo 'medida de erro' é uma tradução direta de *error measure*, utilizada no próprio artigo de Tibshirani et al. (2001). Entretanto, em Kapp & Tibshirani (2007), a mesma função foi chamada de 'medida de homogeneidade intra-grupo' (*within-cluster-homogeneity*). Teixeira (2003), que também trabalhou com a estatística Gap, preferiu a tradução 'medida de qualidade'.

certo ponto o decaimento reduziria-se bruscamente e $W(\boldsymbol{C_k})$ formaria um "cotovelo", como mostrado para k=4 na figura 3.3, construída com os dados de Ruspini citados no capítulo 2. Segundo essa heurística, a localização do cotovelo indicaria o número ótimo de grupos. Note-se que essa proposição remete diretamente às idéias expostas por Thorndike (1953), já citadas na seção 2.2. Tibshirani et al. (2001) propuseram o método da estatística Gap afim de formalizar um procedimento estatístico para esta heurística. Importante dizer que os agrupamentos observados nessa situação devem ser todos construídos com o mesmo método.

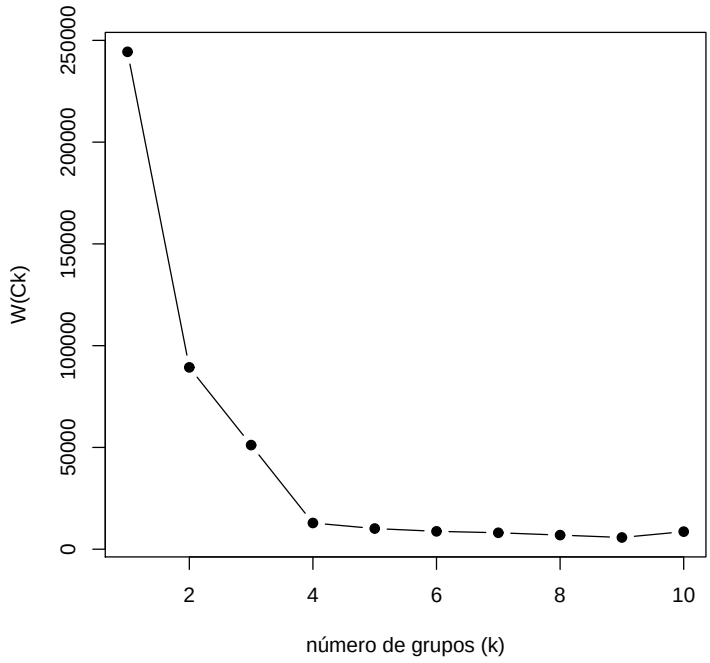

Figura 3.3: o número de grupos (k) versus a medida de erro $W(\boldsymbol{C_k})$.

A idéia básica da estatística Gap pode ser assim descrita: se as unidades amostrais não formassem grupos, sua distribuição seria aproximadamente uniforme dentro das amplitudes das r variáveis observadas. Com uma $W(\boldsymbol{C_k})$ apropriada e realizando-se uma padronização adequada, calcula-se a estatística Gap que permite então avaliar se a amostra obedece a esse padrão ou, em caso negativo, onde o cotovelo ocorreria. A seguir, é apresentado sucintamente como obtê-la. Doravante, considere $W(\boldsymbol{C_k})$ denominada apenas de W_k.

Como medida de erro do agrupamento $\boldsymbol{C}$ quando este possui k grupos Tibshirani et al. (2001) propõem utilizar:

$$W_k = \sum_{r=1}^{K} \frac{1}{2n_r} D_r, \tag{3.4}$$

com

$$D_r = \sum_{i,j \in C_r} d_{ij}$$

e d_{ij} a dissimilaridade entre $i, j \in A$. Note-se que W_k é soma das médias ponderadas das distâncias entre as observações dentro de cada grupo k , k =1,. . . ,K.

Afim de avaliar se a distribuição das variáveis observadas aproximam-se de uma Uniforme, como dito acima, os autores da estatística Gap introduziram o uso de populações de referência como padrão de comparação. Assim, para cada k de interesse e com o uso de um gerador de números aleatórios, obtém-se B dessas populações. Cada uma delas é representada por uma matriz de dados com dimensões equivalentes às de A, ou seja, n observações e r variáveis. Estas últimas devem possuir distribuições Uniformes com amplitudes iguais àquelas originalmente observadas na amostra original. Em seguida, a partir das populações de referência, constrói-se agrupamentos com k=1,. . . ,K grupos utilizando o mesmo método aplicado a A e calcula-se as respectivas medidas de erro W^*_{kb}, b=1,. . . , B. A quantidade B de populações de referência é livremente fixada pelo usuário. Kapp & Tibshirani (2007) realizaram diversos estudos de simulação com B=5 e obtiveram bons resultados. Finalmente, para cada agrupamento $\boldsymbol{C_k}$ de A, a estatística Gap é obtida com a fórmula:

$$Gap_n(k) = E_n[\log(W^*_k)] - \log(W_k) = \frac{1}{B}\sum_b \log(W^*_{kb}) - \log(W_k). \tag{3.5}$$

Para utilizar (3.5) na escolha do número ótimo $\hat{k}$ de grupos, calcula-se $\bar{\ell}$ e sd_k segundo:

$$\bar{\ell} = \frac{1}{B}\sum_b \log(W^*_{kb}) \qquad \text{e} \qquad sd_k = \sqrt{\frac{1}{B}\sum_b \left[\log(W^*_{kb}) - \bar{\ell}\,\right]^2}$$

e aplica-se a seguinte regra de decisão:

$$\hat{k} = \text{menor } k \text{ tal que } Gap(k) \geq Gap(k+1) - sd_{k+1}. \tag{3.6}$$

No artigo original, os autores apresentam justificativas para o uso da W_k mostrada em (3.4) e da distribuição Uniforme na geração dos dados de referência. Eles também afirmam que o uso de $sd_k = 1$ em (3.6) funcionou bem em estudos de simulação e análises de certos conjuntos de dados reais.

Tibshirani et al. (2001) sugerem outra alternativa para construir as distribuições de referência: se $\mathbf{X}$ é a matriz de dimensão $n \times r$ com os dados originais de A, obtém-se $\mathbf{V}$, a matriz cujas colunas são os autovetores de $\mathbf{X^T X}$. Em seguida, calcula-se $\mathbf{X'} = \mathbf{XV}$ e, sobre as amplitudes observadas nas colunas de $\mathbf{X'}$ e utilizando a distribuição Uniforme, como da maneira anteriormente feita, geram-se as populações de referência $\mathbf{Z'_b}$, de dimensões $n \times r$. Finalmente, aplica-se a transformação $\mathbf{Z_b} = \mathbf{Z'_b V^T}$ para se obter $\mathbf{Z_b}$ contendo os dados das populações de referência na base original de $\mathbf{X}$. Com essas matrizes constrói-se os agrupamentos com o mesmo método utilizado em $\mathbf{X}$, obtém-se os respectivos W^*_{kb} e aplica-se (3.5) e (3.6) para a escolha do melhor k. Em outras palavras, trata-se de calcular as populações de referência sobre as componentes principais (CP) de A. Segundo Tibshirani et al. (2001), este método procura tirar proveito do contorno das nuvens de dados aos obter os dados de referência. De fato, nas simulações realizadas por esses autores e por Kapp & Tibshirani (2007), essa estratégia mostrou-se superior nas situações onde os grupos artificialmente gerados para os testes tinham um perfil alongado ao invés de esférico.

Para facilitar a compreensão, um exemplo de construção de uma matriz de referência pelo método das componentes principais é apresentado a seguir. Seja neste exemplo B=1, considere a amostra A com a seguinte matriz de dados $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} 12 & 11 & 7 \\ 10 & 12 & 10 \\ 7 & 15 & 11 \\ 8 & 7 & 11 \\ 5 & 2 & 5 \\ 3 & 4 & 8 \end{pmatrix}, \text{ com } \mathbf{X^T X} = \begin{pmatrix} 391 & 435 & 398 \\ 435 & 559 & 481 \\ 398 & 481 & 480 \end{pmatrix}.$$

A matriz de autovetores ($\mathbf{V}$) associados a $\mathbf{X^T X}$, bem como $\mathbf{X'}$ acima citada, são respectivamente:

$$\mathbf{V} = \begin{pmatrix} -0,52 & 0,28 & 0,81 \\ -0,63 & 0,52 & -0,58 \\ -0,58 & -0,81 & -0,09 \end{pmatrix} \quad \text{e} \quad \mathbf{X'} = \mathbf{XV} = \begin{pmatrix} -17,20 & 3,34 & 2,64 \\ -18,53 & 0,86 & 0,17 \\ -19,43 & 0,76 & -4,10 \\ -14,92 & -3,08 & 1,37 \\ -6,75 & -1,63 & 2,41 \\ -8,70 & -3,59 & -0,65 \end{pmatrix}.$$

Os pares de mínimos/máximos para cada coluna de $\mathbf{X'}$ são dados por (-19,43;-6,75), (-3,59;3,34) e (-4,10;2,64). Com o comando *runif*(n=6, *min*=, *max*=) do pacote R e atribuindo-

se aos parâmetros *min* e *max* os valores acima, pode-se gerar os três vetores correspondentes às colunas de $\mathbf{Z'_1}$ que, com a devida transformação, leva à obtenção de $\mathbf{Z_1}$, a distribuição de referência para $\mathbf{X}$. Assim:

$$\mathbf{Z'_1} = \begin{pmatrix} -16,98 & -2,09 & 0,13 \\ -18,20 & -1,99 & -1,12 \\ -12,52 & -3,50 & -1,20 \\ -9,66 & 0,97 & 2,39 \\ -9,18 & 2,14 & 1,06 \\ -6,84 & -0,15 & 0,23 \end{pmatrix} \quad \text{e} \quad \mathbf{Z_1} = \mathbf{Z'_1 V^T} = \begin{pmatrix} 8,34 & 9,53 & 11,50 \\ 8,00 & 11,07 & 12,24 \\ 4,56 & 6,77 & 10,19 \\ 7,22 & 5,18 & 4,58 \\ 6,22 & 6,26 & 3,48 \\ 3,70 & 4,09 & 4,06 \end{pmatrix}.$$

No presente estudo, quando se desejar fazer referência à estatística Gap calculada pelas maneiras acima, usará-se as siglas Gap UNI e Gap CP.

Tibshirani et al. (2001) recomendam sempre observar o gráfico de linhas Gap(k) versus k, pois seu comportamento pode indicar outras escolhas interessantes de $\hat{k}$, diferentes daquela indicada por (3.6). No capítulo 6 será abordado um exemplo onde isso ocorre.

Teixeira (2003) comenta que a função W_k é sempre monótona quando o algoritmo de construção do agrupamento é do tipo hierárquico. Para a classe dos algoritmos supervisionados, entretanto, pode haver incrementos em W_k quando se passa de k para k+1, eliminando o comportamento monótono. Isso pode perturbar um pouco a estatística Gap , uma vez que seu cálculo em (3.5) envolve uma álgebra de W_k e W^*_{kb}. Nas simulações de Kapp & Tibshirani (2007), com o método das k-médias a estatística Gap não teve bom desempenho no sentido de não detectar a quantidade correta de grupos nos agrupamentos gerados um número elevado de vezes, quando comparada com outros métodos avaliados no mesmo artigo.

Ao contrário de outros métodos, por exemplo a silhueta, a estatística Gap permite identificar casos onde $k = 1$ representa a escolha mais conveniente.

Capítulo 4

Comparação baseada na entropia

A teoria da informação é um campo de estudos originalmente surgido na área de engenharia das comunicações há pouco mais de 50 anos e tem servido para a criação de novos métodos de comparação e avaliação de agrupamentos. Neste capítulo será apresentado um recente índice de comparação descrito por Meilă (2002). Outros exemplos envolvendo entropia são: o método *Jump* (Sugar & James 2003), o método da pureza (Tan, Steinbach & Kumar 2006) e a coesividade (Loganantharaj, Cheepala & Clifford 2006), estes dois últimos úteis tanto para se avaliar um agrupamento quanto um único grupo individualmente. Uma proposta inédita para a escolha do número ótimo de grupos é apresentada na última seção.

Afim de facilitar a compreensão das seções posteriores, nesta primeira estão algumas definições e propriedades ligadas ao conceito de entropia.

4.1 Entropia e informação mútua

Considere o experimento de sortear, ao acaso e sob a suposição de eqüiprobabilidade, um indivíduo qualquer de A, e observar a qual grupo de $\boldsymbol{C}$ ou $\boldsymbol{C'}$ ele pertence. Duas medidas de probabilidade associadas a esse experimento são:

$$P(k) = \frac{n_k}{n} \text{ , e} \tag{4.1}$$

$$P(k, k') = \frac{|C_k \cap C'_{k'}|}{n} = \frac{m_{kk'}}{n}. \tag{4.2}$$

A equação (4.1) representa a probabilidade do ponto sorteado pertencer ao grupo C_k. Desta maneira, o experimento descrito acaba por definir uma variável aleatória discreta com K valores possíveis e associada ao agrupamento $\boldsymbol{C}$. Raciocínio idêntico é válido para o agrupamento $\boldsymbol{C'}$, com n_k sendo substituído por $n'_{k'}$ em (4.1). Também pode-se obter uma outra

distribuição de probabilidade, $P(k,k')$, que representa a chance do indivíduo sorteado pertencer simultaneamente a C_k e $C'_{k'}$, e cuja fórmula é dada por (4.2).

A entropia do agrupamento $\boldsymbol{C}$ é definida com a expressão:

$$H(\boldsymbol{C}) = \sum_{k=1}^{K} P(k) \log \frac{1}{[P(k)]} = -\sum_{k=1}^{K} P(k) \log [P(k)]. \tag{4.3}$$

A expressão (4.3) foi originalmente proposta por Shannon & Weaver (1949). Ela mede a quantidade média de informação da variável aleatória acima descrita com relação a $\boldsymbol{C}$. Quando uma realização desta v.a. é expressa em um código binário, ou seja utilizando-se apenas uma sequência de zeros e uns, diz-se que a entropia é medida em *bits*[1], palavra criada por J. W. Tukey. Neste caso, utiliza-se *log* na base 2. Manning & Schütze (2003) dizem que a entropia pode ser pensada como o comprimento médio da sequência de caracteres necessária para expressar uma realização do experimento e, neste sentido, ela é também uma medida da incerteza média da v.a.

Se há apenas um grupo, ou seja, $C_1 = A$, $H(\boldsymbol{C})$=0, pois $P(1)$=1. Seu máximo é $\log(n)$, e ocorre quando $P(k) = \frac{1}{n}, \forall\, k \in \{1, 2, \ldots, K\}$, ou seja, $K = n$ (cada ponto é um grupo). Se, por qualquer razão, $K < n$, o máximo é atingido com a distribuição uniforme das unidades amostrais dentre os grupos, ou o mais próximo dela caso n não seja múltiplo de K. Assim, conclui-se que ela será sempre maior ou igual a zero.

Também é possível definir a entropia conjunta de $\boldsymbol{C}$ e $\boldsymbol{C'}$:

$$H(\boldsymbol{C}, \boldsymbol{C'}) = -\sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log [P(k,k')]. \tag{4.4}$$

Neste caso lê-se $H(\boldsymbol{C},\boldsymbol{C'})$ como o comprimento médio da sequência necessária para expressar simultaneamente o resultado das v.as. implícitas em $\boldsymbol{C}$ e $\boldsymbol{C'}$ quando realizado o experimento descrito.

Define-se a entropia condicional $H(\boldsymbol{C'}|\boldsymbol{C})$ como:

$$\begin{aligned} H(\boldsymbol{C}|\boldsymbol{C'}) &= \sum_{k'=1}^{K'} P(k')\, H(\boldsymbol{C}|\, k') \\ &= \sum_{k'=1}^{K'} P(k') \left\{ -\sum_{k=1}^{K} P(k|k') \log [P(k|k')] \right\} \\ &= -\sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log [P(k|k')], \end{aligned} \tag{4.5}$$

[1]Contração da expressão *binary digit*

sendo (4.5) interpretada como a quantidade média de informação que ainda resta para comunicar $\boldsymbol{C'}$ dado que $\boldsymbol{C}$ já é conhecido. É direto ver que se $\boldsymbol{C}$ e $\boldsymbol{C'}$ forem independentes, (4.5) se reduzirá a (4.3).

Com a entropia condicional (4.5), pode-se obter a *regra da cadeia para entropia*, assim chamada por Manning & Schütze (2003):

$$H\left(\boldsymbol{C}, \boldsymbol{C'}\right) = H\left(\boldsymbol{C}\right) + H\left(\boldsymbol{C'}|\boldsymbol{C}\right) \tag{4.6}$$

Prova de (4.6):

$$\begin{aligned}
H\left(\boldsymbol{C}, \boldsymbol{C'}\right) &= -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k,k'\right)\right] \\
&= -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k\right)P\left(k'|k\right)\right] \\
&= -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k\right)\right] - \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k'|k\right)\right] \\
&= -\sum_{k=1}^{K} P\left(k\right)\log\left[P\left(k\right)\right] - \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k'|k\right)\right] \\
&= H\left(\boldsymbol{C}\right) + H\left(\boldsymbol{C'}|\boldsymbol{C}\right).
\end{aligned}$$

■

Uma relação bastante útil envolvendo a entropia condicional é:

$$H\left(\boldsymbol{C}\right) \geq H\left(\boldsymbol{C}|\boldsymbol{C'}\right), \tag{4.7}$$

com igualdade quando $\boldsymbol{C}$ e $\boldsymbol{C'}$ forem independentes. Assim, conclui-se que a informação adicional de $\boldsymbol{C'}$ jamais aumentará a incerteza original $H(\boldsymbol{C})$.

Prova de (4.7):

$$\begin{aligned}
H\left(\boldsymbol{C}|\boldsymbol{C'}\right) - H\left(\boldsymbol{C}\right) &= -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k|k'\right)\right] + \sum_{k=1}^{K} P\left(k\right)\log\left[P\left(k\right)\right] \\
&= \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[\frac{1}{P\left(k|k'\right)}\right] + \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k\right)\right] \\
&= \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[\frac{P\left(k\right)}{P\left(k|k'\right)}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\left[\frac{P(k)}{P(k|k')}-1\right]\log e \qquad (*)\\
&\leq \sum_{k=1}^{K}\sum_{k'=1}^{K'}\left[\frac{P(k,k')\,P(k)}{P(k|k')}-P(k,k')\right]\log e\\
&\leq \sum_{k=1}^{K}\sum_{k'=1}^{K'}\left[P(k)\,P(k')-P(k,k')\right]\log e\\
&\leq \sum_{k=1}^{K}\left[P(k)-P(k)\right]\log e\\
&\leq 0\\
H(\boldsymbol{C}) &\geq H(\boldsymbol{C}|\boldsymbol{C'}).
\end{aligned}
$$

Note-se que a desigualdade em $(*)$ foi introduzida utilizando-se

$$\ln x \leq x-1 \;\Rightarrow \ln x \log e \leq (x-1)\log e \;\Rightarrow \log x \leq (x-1)\log e,$$

na qual se fez uso da mudança de base no logaritmo, ou seja:

$$\log_b^a = \frac{\log_e^a}{\log_e^b} \quad \Leftrightarrow \quad \log_e^a = \log_b^a \log_e^b .$$

■

Já que de (4.7) têm-se

$$H(\boldsymbol{C}) + H(\boldsymbol{C'}) \geq H(\boldsymbol{C}|\boldsymbol{C'}) + H(\boldsymbol{C'}),$$

com a regra da cadeia (4.6) obtém-se outra relação útil:

$$H(\boldsymbol{C}) + H(\boldsymbol{C'}) \geq H(\boldsymbol{C},\boldsymbol{C'}). \qquad (4.8)$$

Também com a regra da cadeia, e dado o fato da entropia conjunta ser simétrica, tem-se:

$$
\begin{aligned}
H(\boldsymbol{C},\boldsymbol{C'}) &= H(\boldsymbol{C'},\boldsymbol{C})\\
H(\boldsymbol{C}) + H(\boldsymbol{C'}|\boldsymbol{C}) &= H(\boldsymbol{C'}) + H(\boldsymbol{C}|\boldsymbol{C'})\\
H(\boldsymbol{C}) - H(\boldsymbol{C}|\boldsymbol{C'}) &= H(\boldsymbol{C'}) - H(\boldsymbol{C'}|\boldsymbol{C}). \qquad (4.9)
\end{aligned}
$$

A diferença expressa em ambos os lados da igualdade (4.9) é chamada de *informação mútua* entre $\boldsymbol{C}$ e $\boldsymbol{C'}$. Ela é a redução obtida na incerteza da variável aleatória associada a $\boldsymbol{C}(\boldsymbol{C'})$, uma vez conhecido a que grupo de $\boldsymbol{C'}(\boldsymbol{C})$ pertence o indivíduo sorteado, ou seja, quanta informação contém um agrupamento a respeito de outro.

Ela é denotada por $I(\boldsymbol{C},\boldsymbol{C'})$ e com o auxílio de (4.6) pode-se deduzir uma expressão formal para a mesma:

$$
\begin{aligned}
I\left(\boldsymbol{C},\boldsymbol{C'}\right) &= H\left(\boldsymbol{C}\right) - H\left(\boldsymbol{C}|\boldsymbol{C'}\right) \qquad (4.10)\\
&= H\left(\boldsymbol{C}\right) + H\left(\boldsymbol{C'}\right) - H\left(\boldsymbol{C},\boldsymbol{C'}\right)\\
&= \underbrace{\sum_{k=1}^{K} P\left(k\right)\log\frac{1}{\left[P\left(k\right)\right]}}_{\mathbf{I}} + \underbrace{\sum_{k'=1}^{K'} P\left(k'\right)\log\frac{1}{\left[P\left(k'\right)\right]}}_{\mathbf{II}} + \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k,k'\right)\right].
\end{aligned}
$$

Dado que **I** e **II** podem ser escritos respectivamente como

$$
\begin{aligned}
\mathbf{I}: \qquad & \sum_{k=1}^{K} P\left(k\right)\log\frac{1}{\left[P\left(k\right)\right]} = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\frac{1}{\left[P\left(k\right)\right]} \quad \text{e}\\
\mathbf{II}: \qquad & \sum_{k'=1}^{K'} P\left(k'\right)\log\frac{1}{\left[P\left(k'\right)\right]} = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\frac{1}{\left[P\left(k\right)\right]},
\end{aligned}
$$

a informação mútua entre $\boldsymbol{C}$ e $\boldsymbol{C'}$ é igual a:

$$
I\left(\boldsymbol{C},\boldsymbol{C'}\right) = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[\frac{P\left(k,k'\right)}{P\left(k\right)P\left(k'\right)}\right]. \qquad (4.11)
$$

Observando-se (4.11) é direto ver que a informação mútua é simétrica e, além disso, com (4.7) e (4.10) pode-se concluir que ela é maior que zero. Isto é:

$$
I\left(\boldsymbol{C},\boldsymbol{C'}\right) = I\left(\boldsymbol{C'},\boldsymbol{C}\right) \geq 0, \qquad (4.12)
$$

em que se satisfaz a igualdade se $\boldsymbol{C}$ e $\boldsymbol{C'}$ forem independentes.

Supondo $H\left(\boldsymbol{C}\right) > H\left(\boldsymbol{C'}\right)$, a partir de (4.9) pode-se dizer que $I\left(\boldsymbol{C},\boldsymbol{C'}\right) \leq H\left(\boldsymbol{C'}\right)$. Ao contrário, se $H\left(\boldsymbol{C'}\right) > H\left(\boldsymbol{C}\right) \Rightarrow I\left(\boldsymbol{C},\boldsymbol{C'}\right) \leq H\left(\boldsymbol{C}\right)$. Então, pode-se escrever:

$$
I\left(\boldsymbol{C},\boldsymbol{C'}\right) \leq \min\Big(H\left(\boldsymbol{C}\right), H\left(\boldsymbol{C'}\right)\Big). \qquad (4.13)
$$

Um caso onde a igualdade ocorre em (4.13) é o seguinte: suponha que $K' = K - 1$ e que C'_a foi obtido tomando-se um $C'_a = C_i + C_j,\ a \in \{1,\ldots,K'\},\ i,j \in \{1,\ldots,K\},\ i \neq j$. Para qualquer outro $C'_b \in \boldsymbol{C'},\ b \in \{1,\ldots,K'\},\ b \neq a$, faça $C'_b = C_b$. Então, $H\left(\boldsymbol{C'}|\boldsymbol{C}\right) = 0$ pois se é conhecido o grupo ao qual o indivíduo sorteado pertence em $\boldsymbol{C}$, nenhuma informação extra é necessária para se conhecer sua respectiva alocação em $\boldsymbol{C'}$. Assim, de (4.9) tem-se que $H\left(\boldsymbol{C}\right) - H\left(\boldsymbol{C}|\boldsymbol{C'}\right) = H\left(\boldsymbol{C'}\right)$, e como $H\left(\boldsymbol{C}|\boldsymbol{C'}\right) > 0$, conclui-se que $H\left(\boldsymbol{C}\right) > H\left(\boldsymbol{C'}\right)$. Com (4.10) escreve-se:

$$
I\left(\boldsymbol{C},\boldsymbol{C'}\right) = H\left(\boldsymbol{C'}\right) - H\left(\boldsymbol{C'}|\boldsymbol{C}\right) = H\left(\boldsymbol{C'}\right),
$$

e está mostrada a situação em que o limite inferior de (4.13) é atingido. É possível estendê-la dizendo que a igualdade em (4.13) será observada toda vez que um dos agrupamentos for completamente determinável a partir de outro, pois uma das entropias condicionais de (4.9) assumirá valor nulo. E se $\boldsymbol{C}$=$\boldsymbol{C'}$ é direto ver que:

$$I\left(\boldsymbol{C},\boldsymbol{C'}\right) = H\left(\boldsymbol{C}\right) = H\left(\boldsymbol{C'}\right).$$

Para auxiliar a visualização dos conceitos relacionados às entropias das partições, um diagrama baseado na teoria dos conjuntos é apresentado na figura 4.1.

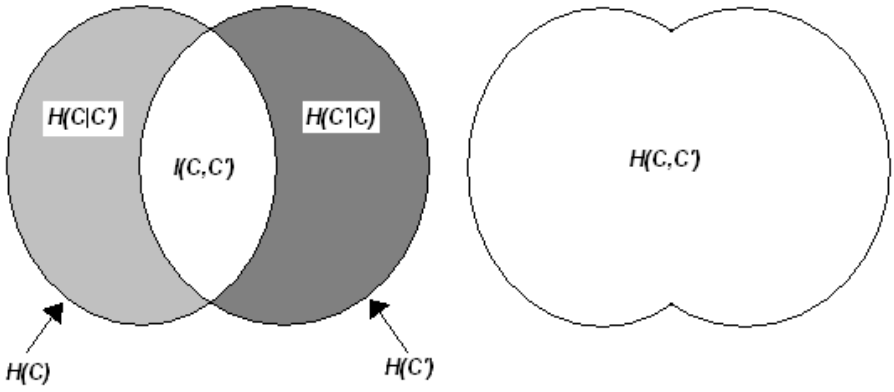

Figura 4.1: Relações entre as entropias das partições $\boldsymbol{C}$ e $\boldsymbol{C'}$ e sua informação mútua.

4.2 Comparação utilizando a variação da informação

Utilizando os conceitos apresentados na seção anterior, Meilă (2002) propôs um novo índice de comparação:a variação da informação (VI) entre dois agrupamentos. Sua expressão é a seguinte:

$$VI\left(\boldsymbol{C}, \boldsymbol{C'}\right) = H\left(\boldsymbol{C}|\boldsymbol{C'}\right) + H\left(\boldsymbol{C'}|\boldsymbol{C}\right) = H\left(\boldsymbol{C}\right) + H\left(\boldsymbol{C'}\right) - 2\,I\left(\boldsymbol{C}, \boldsymbol{C'}\right). \tag{4.14}$$

Assim, VI dá uma medida da variação total da informação induzida pelos agrupamentos $\boldsymbol{C}$ e $\boldsymbol{C'}$ com respeito a A. Além disso, em uma mudança do primeiro para o segundo agrupamento, ela também permite que se quantifique quanta informação sobre $\boldsymbol{C}$ será perdida e quanta informação sobre $\boldsymbol{C'}$ será ganha, a partir das suas respectivas entropias condicionais. Na figura 4.1 ela seria representada pela soma das duas áreas cinzas.

É direto ver que no caso de independência, VI se reduz à soma das entropias. E se, por exemplo, $\boldsymbol{C'}$ for completamente determinável a partir de $\boldsymbol{C}$, a variação da informação será a diferença das entropias, pois, $I(\boldsymbol{C},\boldsymbol{C'})$=$H(\boldsymbol{C'})$.

4.3 Propriedades da variação da informação

Meilă (2002) apresenta várias propriedades da VI para compreender sua estrutura no espaço dos agrupamentos e checar sua adequação nas comparações entre os mesmos. As propriedades que não estão demonstradas no artigo original são verificadas nesta seção.

Propriedade 1: VI é uma métrica no espaço dos agrupamentos. Isso significa que ela satisfaz:

(a) positividade: $VI(C,C')\geq 0$, com igualdade somente se $C=C'$

(b) simetria: $VI(C,C')=VI(C',C)$

(c) desigualdade triangular: com C_1, C_2 e C_3 três agrupamentos quaisquer de A

$$VI(C_1,C_2) + VI(C_2,C_3) \geq VI(C_1,C_3) \tag{4.15}$$

As provas de (a) e (b) são diretas. Para (c), a demonstração segue abaixo.

Prova de (4.15):

$$\textbf{i:} \quad VI(\mathbf{C_1},\mathbf{C_2}) = H(\mathbf{C_1}|\mathbf{C_2}) + H(\mathbf{C_2}|\mathbf{C_1}) = 2\,H(\mathbf{C_1},\mathbf{C_2}) - H(\mathbf{C_1}) - H(\mathbf{C_2})$$

$$\textbf{ii:} \quad VI(\mathbf{C_2},\mathbf{C_3}) = H(\mathbf{C_2}|\mathbf{C_3}) + H(\mathbf{C_3}|\mathbf{C_2}) = 2\,H(\mathbf{C_2},\mathbf{C_3}) - H(\mathbf{C_2}) - H(\mathbf{C_3})$$

$$\textbf{iii:} \quad VI(\mathbf{C_1},\mathbf{C_3}) = H(\mathbf{C_1}|\mathbf{C_3}) + H(\mathbf{C_3}|\mathbf{C_1}) = 2\,H(\mathbf{C_1},\mathbf{C_3}) - H(\mathbf{C_1}) - H(\mathbf{C_3})$$

Reescrevendo-se a desigualdade triangular usando **i**, **ii** e **iii** obtém-se:

$$2\Big[H(\mathbf{C_1},\mathbf{C_2}) + H(\mathbf{C_2},\mathbf{C_3}) - H(\mathbf{C_1},\mathbf{C_3}) - H(\mathbf{C_2})\Big] \geq 0$$

E com a relação expressa em (4.8), tem-se:

$$\begin{aligned} 2\Big[H(\mathbf{C_1}) + H(\mathbf{C_2}) + H(\mathbf{C_2}) + H(\mathbf{C_3}) - H(\mathbf{C_1}) - H(\mathbf{C_3}) - H(\mathbf{C_2})\Big] &\geq 0 \Rightarrow \\ \Rightarrow H(\mathbf{C_2}) &\geq 0, \end{aligned}$$

o que demonstra a desigualdade triangular para VI . ∎

O fato da VI ser uma métrica a torna facilmente utilizável e de assimilação simples. As pessoas estão acostumadas a utilizar este tipo de estrutura, por exemplo, para escolher o melhor caminho entre dois pontos da cidade, em termos de distância, empregando, sobretudo, a simetria e a desigualdade triangular.

Como o espaço dos agrupamentos é finito, pode-se afirmar que VI também é limitada. As próximas propriedades tratam desses limites.

Propriedade 2: n-invariância. O valor de $VI(\boldsymbol{C},\boldsymbol{C'})$ depende somente dos tamanhos relativos dos grupos, não importando quantos pontos contenham cada um deles.

Propriedade 3: VI tem limite superior conhecido, para qualquer n:

$$VI(\mathbf{C},\mathbf{C'}) \leq \log(n). \tag{4.16}$$

Como a entropia é máxima quando cada ponto é um grupo diferente ($K = n$) e mínima quando há um único grupo igual ao próprio A, é direto ver que o limite acima sempre poderá ser atingido. Por exemplo, fazendo-se $\boldsymbol{C} = \{\{1\},\{2\},\ldots,\{n\}\}$ e $\boldsymbol{C'} = \{A\}$. Isso não contradiz a propriedade 2, pois nela a dependência de VI é em relação às proporções de pontos contidos em cada um dos grupos. E padronizando-se VI pelo limite dado em (4.16), ela se torna uma medida que varia em [0,1].

Propriedade 4: se $K, K' \leq K^*$, com $K^* \leq \sqrt{n}$, então:

$$VI(\boldsymbol{C},\boldsymbol{C'}) \leq 2\log(K^*). \tag{4.17}$$

Prova:

$$\begin{aligned}
VI(\boldsymbol{C},\boldsymbol{C'}) &= H(\boldsymbol{C}) + H(\boldsymbol{C'}) - 2\,I(\boldsymbol{C},\boldsymbol{C'}) \\
&\leq H(\boldsymbol{C}) + H(\boldsymbol{C}) && \text{(vide 4.12)} \\
&\leq -\sum_{k=1}^{K} P(k)\log[P(k)] - \sum_{k'=1}^{K'} P(k')\log[P(k')] \\
&\leq -\sum_{k=1}^{K} \frac{n_k}{n}\log\left(\frac{n_k}{n}\right) - \sum_{k'=1}^{K'} \frac{n_{k'}}{n}\log\left(\frac{n_{k'}}{n}\right) \\
&\leq -\sum_{k=1}^{K} \frac{1}{K}\log\left(\frac{1}{K}\right) - \sum_{k'=1}^{K'} \frac{1}{K'}\log\left(\frac{1}{K'}\right) \quad \left(\text{entropia máxima} \Rightarrow n_k = \frac{n}{K}\right) \\
&\leq \log K + \log K' \\
&\leq 2\log K^*. && (\max(K,K') = K^*) \qquad (4.18)
\end{aligned}$$

■

Então, o máximo em (4.18) será atingido toda vez que $K{=}K'{=}K^*$ e $I(\boldsymbol{C},\boldsymbol{C'}) = 0, \forall k,k' \in \{1,2,\ldots,K^*\}$, ou seja, de (4.11):

$$\frac{P(k,k')}{P(k)\,P(k')} = 1 \Rightarrow \frac{m_{kk'}}{n} = \frac{1}{n}\times\frac{1}{n} \Rightarrow \frac{m_{kk'}}{n} = \frac{1}{n^2} \Rightarrow m_{kk'} = \frac{1}{n} \Rightarrow n \text{ múltiplo de } (K^*)^2.$$

Com o número de grupos em $\boldsymbol{C}$ e $\boldsymbol{C'}$ menor ou igual a um K^*, VI passa a ter um limite superior que depende somente dessa constante. Portanto, agrupamentos obtidos de diferentes amostras, cada qual com suas próprias dimensões, porém sujeitos ao mesmo limite K^*, estão numa mesma escala da métrica da variação da informação.

Esta propriedade traz uma interessante possibilidade de aplicação. Suponha que A_1, A_2 e A_3 sejam três amostras de pacientes com suspeita de uma certa doença, obtidas em três hospitais distintos. Para o diagnóstico definitivo é feito um exame tipo biópsia, que distingue os pacientes em três categorias: (1) sem a doença, (2) doença de leve intensidade e (3) doença grave. Então, respectivamente, obtém-se $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ e $\boldsymbol{C_3}$, com K_1=K_2=K_3=3. Suponha agora que é proposto um teste clínico menos invasivo do que a biópsia e cujo resultado também se espera que possa discriminar os doentes. Então, para as amostras sob estudo, aplica-se um mesmo método de agrupamento assumindo-se K'_1=K'_2=K'_3=K^*=3 e constrói-se $\boldsymbol{C'_1}$, $\boldsymbol{C'_2}$ e $\boldsymbol{C'_3}$. Uma vez calculados os valores de $VI(\boldsymbol{C_1},\boldsymbol{C'_1})$, $VI(\boldsymbol{C_2},\boldsymbol{C'_2})$ e $VI(\boldsymbol{C_3},\boldsymbol{C'_3})$, pode-se compará-los entre si pois a propriedade 4 garante que eles se encontram na mesma escala métrica de VI. O desvio-padrão das mesmas pode, por exemplo, servir como medida de dispersão para avaliar o quanto o novo teste clínico varia em relação ao padrão-ouro.

Em outras palavras, se há sentido em se considerar dois estudos de agrupamentos obtidos com duas diferentes amostras como equivalentes, dentro dos parâmetros acima discutidos, então pode-se comparar diretamente as medidas de VI obtidas nos dois espaços, somando-as e/ou subtraindo-as, pois há garantias de equivalência em suas métricas.

Também é possível encontrar um limite inferior de variação de VI em relação a $\boldsymbol{C}$ ou, em outras palavras, menor valor possível para VI, supondo $\boldsymbol{C} \neq \boldsymbol{C'}$. As próximas propriedades tratam exatamente disso.

Propriedade 5: suponha que $\boldsymbol{C'}$ foi obtido de $\boldsymbol{C}$ dividindo-se algum C_a, $a \in \{1,\ldots,K\}$, nos grupos $C'_{k'_1},\ldots,C'_{k'_m}$. As probabilidades dos grupos de $\boldsymbol{C'}$ são:

$$P\left(k'\right) = \begin{cases} P\left(k\right) & \text{se } C'_{k'} \in \boldsymbol{C}, \text{e} \\ P\left(k'|a\right) P\left(a\right) & \text{se } C'_{k'} \subseteq C_a \in \boldsymbol{C}. \end{cases} \tag{4.19}$$

Em (4.19), para $k' \in \{k'_1,\ldots,k'_m\}$ e $\ell = 1,\ldots,m$, $P\left(k'|k\right)$ acima é:

$$P\left(k'|a\right) = \frac{\left|C'_{k'_\ell}\right|}{\left|C_a\right|} \tag{4.20}$$

e sua entropia, representando a incerteza associada à subdivisão de C_k, é dada por:

$$H_{|C_a} = -\sum_{\ell=1}^{m} P\left(k'_\ell|a\right) \log\left[P\left(k'_\ell|a\right)\right].$$

Então, nesse caso tem-se:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) = P(a)\, H_{|C_a}. \tag{4.21}$$

Prova:

Para quaisquer K e K', tem-se as seguintes fórmulas gerais para cálculos de probabilidades no contexto de agrupamentos:

$$P(k,k') = \frac{m_{kk'}}{n} \qquad P(k|k') = \frac{P(k,k')}{P(k')} = \frac{\frac{m_{kk'}}{n}}{\frac{n'_{k'}}{n}} = \frac{m_{kk'}}{n'_{k'}} \qquad P(k'|k) = \frac{m_{kk'}}{n_k}. \tag{4.22}$$

E na situação particular desta propriedade, pode-se calcular essas quantidades como abaixo é mostrado:

- se $k \neq a$ e $k' \neq k \Rightarrow m_{kk'} = 0 \Rightarrow P(k,k') = 0$;
- se $k \neq a$ e $k' = k \Rightarrow m_{kk'} = n_k = n'_{k'} \Rightarrow \log[P(k|k')] = \log[P(k'|k)] = \log 1 = 0$;
- se $k = a$ e $k' \notin \{k'_1, k'_2, \ldots, k'_m\} \Rightarrow m_{kk'} = 0 \Rightarrow P(k,k') = 0$;
- se $k = a$ e $k' \in \{k'_1, k'_2, \ldots, k'_m\} \Rightarrow m_{kk'} = n'_{k'_\ell} \Rightarrow P(k,k') = P(a)\,.\,P(k'|a)$, $\log[P(k|k')] = 0$ e $P(k'|k) = \frac{n'_{k'_\ell}}{n_a} \neq 0$.

Então tem-se:

$$H(\boldsymbol{C}|\boldsymbol{C'}) = -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log[P(k|k')] = 0, \text{ pois } \log[P(k|k')] = 0, \forall\, k, k'.$$

E com esses resultados finalmente se calcula VI, como mostrado abaixo:

$$\begin{aligned} VI(\boldsymbol{C}, \boldsymbol{C'}) &= H(\boldsymbol{C}|\boldsymbol{C'}) + H(\boldsymbol{C'}|\boldsymbol{C}) \\ &= 0 - \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log[P(k'|k)] \\ &= -\sum_{\ell=1}^{m} P(a)\, P(k'_\ell|a)\log[P(k'_\ell|a)] \\ &= P(a)\, H_{|C_a}. \end{aligned}$$

■

O mesmo valor de VI será obtido se a operação contrária for realizada, ou seja, se dois ou mais grupos de $\boldsymbol{C}$ forem unidos em um só $C'_{k'}$.

A relação apresentada em (4.21) mostra que subdivisões (ou uniões) de grupos proporcionalmente menores afetam menos o valor de VI do que aquelas feitas nos maiores. Além disso, ela mostra uma característica interessante da variação da informação: quando se obtém

$\boldsymbol{C'}$ através de subdivisões e/ou uniões de grupos de $\boldsymbol{C}$, a medida de VI é independente dos grupos que não participam dessas operações.

Propriedade 6: se $\boldsymbol{C'}$ é obtido de $\boldsymbol{C}$ dividindo-se C_a em q grupos iguais, então:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) = P(a) \log q. \tag{4.23}$$

Essa é uma consequência direta da aplicação da propriedade 5, dividindo-se algum C_a em iguais $C'_{k'_1}, \ldots, C'_{k'_\ell}$. Neste caso, $H_{|C_a} = \log q$ pois as probabilidades $P(k'_\ell|a)$ são todas iguais. Multiplicando-se pela probabilidade de escolher-se o grupo C_a, obtém-se a expressão (4.23). Note-se que $H_{|C_a} = \log q$ é uma aplicação de máximo da entropia sobre C_a. Portanto, se deseja-se obter $\boldsymbol{C'}$ a partir de subdivisões de um único grupo de $\boldsymbol{C}$, o maior valor de VI será obtido tomando-se o grupo mais numeroso de $\boldsymbol{C}$ e dividindo-o em grupos de um único elemento (ou seja, maximizando-se simultaneamente $P(a)$ e $\log q$).

Propriedade 7: se $\boldsymbol{C'}$ é obtido de $\boldsymbol{C}$ separando-se um único ponto de C_k e fazendo-o como um novo grupo, então:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) = \frac{1}{n}\Big[n_k \log(n_k) - (n_k - 1)\log(n_k - 1)\Big] \tag{4.24}$$

Prova:

Suponha que foi escolhido $C_a, a \in \{1, \ldots, K\}$ para realizar a operação acima descrita. Assim, é obtido $\boldsymbol{C'}$ contendo $C'_{k'_1}$ e $C'_{k'_2}$, respectivamente com $n'_{k'_1} = m_{kk'_1} = n_a - 1$ e $n'_{k'_2} = m_{kk'_2} = 1$ elementos, $C_a = C'_{k'_1} \cup C'_{k'_2}$. Neste contexto, é possível calcular os valores das probabilidades citadas em (4.22):

- se $k \neq a \Rightarrow m_{kk'} = n_k = n'_{k'} \Rightarrow \log[P(k|k')] = \log[P(k'|k)] = \log 1 = 0$
- $k = a \Rightarrow P(k, k'_1) = \frac{n_a - 1}{n}$, $P(k, k'_2) = \frac{1}{n}$, $P(k|k'_1) = \frac{n_a - 1}{n_a - 1} = 1, P(k|k'_2) = 1$,
  $P(k'_1|k) = \frac{n_a - 1}{n_a}$ e $P(k'_2|k) = \frac{1}{n_a}$.

Então, tem-se:

$$H(\boldsymbol{C}|\boldsymbol{C'}) = -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k, k') \log[P(k|k')] = 0, \text{ pois } \log[P(k|k')] = 0, \forall\, k, k' \tag{4.25}$$

e

$$
\begin{aligned}
H\left(\boldsymbol{C'}|\boldsymbol{C}\right) &= -\sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\left[P\left(k'|k\right)\right] \\
&= -\left[\frac{n_a-1}{n}\log\left(\frac{n_a-1}{n_a}\right)+\frac{1}{n}\log\left(\frac{1}{n_a}\right)\right] \\
&= -\frac{1}{n}\Big[\left(n_a-1\right)\log\left(n_a-1\right)-\left(n_a-1\right)\log n_a-\log n_a\Big] \\
&= \frac{1}{n}\Big[n_a\log n_a-\left(n_a-1\right)\log\left(n_a-1\right)\Big], \qquad (4.26)
\end{aligned}
$$

pois se $k \neq a \Rightarrow \log\left[P\left(k|k'\right)\right] = 0, \forall k, k'$.
E com (4.25) e (4.26) pode-se calcular VI:

$$
\begin{aligned}
VI\left(\boldsymbol{C},\boldsymbol{C'}\right) &= H\left(\boldsymbol{C}|\boldsymbol{C'}\right)+H\left(\boldsymbol{C'}|\boldsymbol{C}\right) \\
&= \frac{1}{n}\Big[n_a\log n_a-\left(n_a-1\right)\log\left(n_a-1\right)\Big].
\end{aligned}
$$

■

Como separar um ponto, fazendo-o um novo grupo, representa a menor perturbação na entropia de um agrupamento, pode-se supor que, em termos da métrica VI, na vizinhança mais próxima de $\boldsymbol{C}$ encontram-se agrupamentos obtidos dessa maneira. O raciocínio inverso também é aplicável, ou seja, se $\boldsymbol{C'}$ é obtido unindo-se um grupo de ponto único a outro qualquer, isso sugere que ele seja um vizinho próximo de $\boldsymbol{C}$.

Antes de prosseguir com essa idéia, uma nova definição é introduzida. Considere:

$$
\boldsymbol{C}\times\boldsymbol{C'} = \left\{C_k\cap C'_{k'} \mid C_k\in\boldsymbol{C}, C'_{k'}\in\boldsymbol{C'}, C_k\cap C'_{k'}\neq\emptyset\right\}, \qquad (4.27)
$$

como a operação de *produto* entre os agrupamentos $\boldsymbol{C}$ e $\boldsymbol{C'}$. Isto é, o agrupamento formado por todas as intersecções não vazias entre cada C_k e $C'_{k'}$. A matriz $\mathbf{M}$ oferece uma boa idéia de como é o agrupamento resultante do produto, e com ela também pode-se concluir que $\boldsymbol{C}\times\boldsymbol{C'}$ determina completamente $\boldsymbol{C}$ e $\boldsymbol{C'}$, isto é: sabendo-se a que grupo do produto pertence um certo ponto de A, pode-se explicitar unicamente os respectivos C_k e $C'_{k'}$. Assim, $\boldsymbol{C}\times\boldsymbol{C'}$ contém simultaneamente toda a informação dos dois agrupamentos originais. Se $\boldsymbol{C}=\boldsymbol{C'}$, $\mathbf{M}$ é quadrada, $m_{kk'}=n_k$ se $k=k'$ e os elementos fora da diagonal são nulos. E então, $\boldsymbol{C}\times\boldsymbol{C'}=\boldsymbol{C}=\boldsymbol{C'}$.

Define-se também que $\boldsymbol{C'}$ será um *refinamento* de $\boldsymbol{C}$ quando para cada $C'_{k'}\in\boldsymbol{C'}$ houver um, e somente um, grupo $C_k\in\boldsymbol{C}$, tal que $C'_{k'}\subseteq C_k$. Assim, refina-se um agrupamento $\boldsymbol{C}$ subdividindo-se um ou mais de seus grupos, e isso implica que $K'\geq K$, com igualdade só se $\boldsymbol{C}=\boldsymbol{C'}$.

Se $\boldsymbol{C'}$ é um refinamento de $\boldsymbol{C}$, então $\boldsymbol{C} \times \boldsymbol{C'} = \boldsymbol{C'}$, caso contrário, o produto entre eles é refinamento tanto de $\boldsymbol{C}$ quanto de $\boldsymbol{C'}$.

Propriedade 8: colinearidade do produto. A desigualdade triangular mostrada em (4.15) assume igualdade para dois agrupamentos e seu produto:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) = VI(\boldsymbol{C}, \boldsymbol{C} \times \boldsymbol{C'}) + VI(\boldsymbol{C'}, \boldsymbol{C} \times \boldsymbol{C'})\,. \tag{4.28}$$

A prova de (4.28) encontra-se em Meilă (2002).

Então, o produto de dois agrupamentos é colinear com os mesmos, e situa-se numa distância intermediária entre ambos dentro da métrica induzida por *VI*.

Propriedade 9: sejam $\boldsymbol{C}$ e $\boldsymbol{C'}$ quaisquer:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) \geq VI(\boldsymbol{C}, \boldsymbol{C} \times \boldsymbol{C'})\,, \tag{4.29}$$

A prova é direta, fazendo-se $VI(\boldsymbol{C}, \boldsymbol{C'}) - VI(\boldsymbol{C}, \boldsymbol{C} \times \boldsymbol{C'}) \geq 0$, com igualdade apenas se $\boldsymbol{C'} = \boldsymbol{C} \times \boldsymbol{C'}$ devido à simetria de VI . E nesse caso, o último termo da expressão (4.28) é nulo.

A partir de (4.28) e (4.29), dados $\boldsymbol{C}$ e $\boldsymbol{C'}$ quaisquer e diferentes entre si, pode-se concluir que sempre é possível construir pelo menos um terceiro agrupamento $\boldsymbol{C''}$ mais próximo de $\boldsymbol{C}$ do que $\boldsymbol{C'}$, bastando fazer $\boldsymbol{C''} = (\boldsymbol{C} \times \boldsymbol{C'})$. A única exceção ocorre quando $\boldsymbol{C'} = (\boldsymbol{C} \times \boldsymbol{C'})$.

Pela propriedade (4.24), dado um $\boldsymbol{C}$ qualquer, basta separar um único ponto do menor grupo de $\boldsymbol{C}$, fazendo-o um novo grupo, para obter o vizinho mais próximo com $K' = K + 1$. Ao contrário, se $K' = K - 1$, basta unir os dois menores grupos para obtê-lo. Note que não é possível obter um vizinho mais próximo com $K' = K$ pois nesse caso $\boldsymbol{C''} = (\boldsymbol{C} \times \boldsymbol{C'})$ é um agrupamento mais próximo que $\boldsymbol{C'}$, com $K' \neq K$.

Generalizando-se sobre o espaço de todos os agrupamentos possíveis de A, tem-se que a menor distância possível é aquela obtida quando se une dois grupos com um único ponto cada (ou, analogamente, quando se divide um grupo de dois pontos). E, com a aplicação direta de (4.24), isso implica:

$$VI(\boldsymbol{C}, \boldsymbol{C'}) \geq \frac{1}{n}\Big[2\log(2) - (2-1)\log(2-1)\Big] = \frac{2}{n}, \quad \text{para } \boldsymbol{C} \neq \boldsymbol{C'}. \tag{4.30}$$

Assim, o aumento de n diminui o limite inferior de *VI*, dado em (4.30). Além disso, como mostrado em (4.16), cresce também o diâmetro do espaço dos agrupamentos. Isso é esperado pois com amostras maiores pode-se construir agrupamentos impossíveis de serem obtidos com n menor. A esse fato, Meilă (2002) chamou de aumento da "granularidade" do espaço dos agrupamentos.

4.4 Comparação de agrupamentos com a diferença de entropias condicionais

Nesta seção é apresentado um novo índice para comparação de agrupamentos e escolha do número ótimo de grupos. Este índice é baseado na diferença das entropias condicionais.

Considere uma sequência de agrupamentos e a heurística do "cotovelo", como discutida no início da seção 3.5. Uma conjectura sobre as entropias condicionais $H(\boldsymbol{C_2}|\boldsymbol{C_1})$, $H(\boldsymbol{C_3}|\boldsymbol{C_2})$,..., $H(\boldsymbol{C_{k\text{-}1}}|\boldsymbol{C_{k\text{-}2}})$, $H(\boldsymbol{C_k}|\boldsymbol{C_{k\text{-}1}})$ é se as mesmas poderiam ser usadas para construir uma medida de erro do agrupamento e encontrar o "cotovelo".

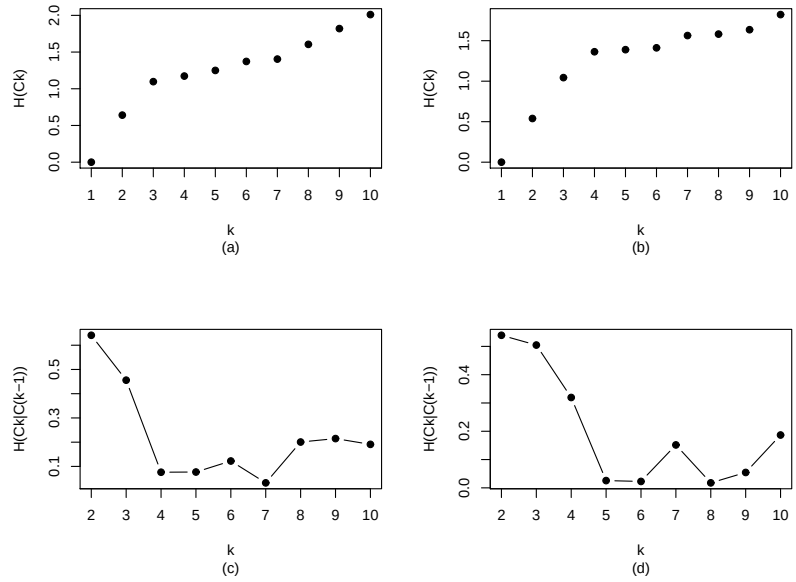

Figura 4.2: Gráficos de $H(\boldsymbol{C_k})$ e $H(\boldsymbol{C_k}|\boldsymbol{C_{k\text{-}1}})$ versus o número de grupos k: (a) e (c) são para um conjunto de dados contendo três grupos, (b) e (d) para o caso com quatro grupos.

As figuras 4.2(a) e 4.2(c), respectivamente, mostram os valores de $H(\boldsymbol{C_k})$ e $H(\boldsymbol{C_k}|\boldsymbol{C_{k\text{-}1}})$ para um conjunto de dados gerado artificialmente contendo três grupos bem distintos no $\Re^2$. Já as figuras 4.2(b) e 4.2(d) mostram os valores para o caso com quatro grupos no $\Re^3$ [2]. Para ambos foi usado o método hierárquico com ligação média e distância euclidiana. As entropias de $H(\boldsymbol{C_k})$ são monótonas crescentes nos dois casos. Entretanto, ao observar-se as entropias condicionais, nota-se a existência do "cotovelo" procurado quando o número verdadeiro de grupos é ultrapassado por k. Então, uma possível medida de erro para o agrupamento $\boldsymbol{C_k}$

[2]Note que $H(\boldsymbol{C}_1|\boldsymbol{C}_0)$ não está nos gráficos 4.2(c) e 4.2(d) pois seu cálculo não tem sentido.

pode ser dada por:

$$dHC\,(k) = \Big[H\,(\boldsymbol{C_{k+1}}|\boldsymbol{C_k}) - H\,(\boldsymbol{C_k}|\boldsymbol{C_{k\text{-}1}})\Big] \times H\,(\boldsymbol{C_k})\,,\ \text{para } k = 2, \ldots, K. \tag{4.31}$$

A proposta de (4.31) é acompanhar o decrescimento dos valores das entropias condicionais à medida em que k aumenta. A determinação do cotovelo é feita encontrando-se o valor $\hat{k}$ para o qual (4.31) é mínimo. Ou seja:

$$dHC\left(\hat{k}\right) = \min_k \Big[dHC(k)\Big], \quad k = 2, \ldots, K. \tag{4.32}$$

Assim como na estatística Gap, ela também privilegia a escolha do menor k. Para tanto, uma penalização foi imposta multiplicando-se as diferenças de entropias condicionais pela entropia de $\boldsymbol{C_k}$, uma vez que esta última é uma medida de incerteza crescente quando se aumenta o número de grupos.

As figuras 4.3(a) e 4.3(b) mostram $dHC\,(k)$ para os conjuntos de dados com três e quatro grupos citados anteriormente. Note como os mínimos encontram-se em k=3 e k=4 grupos. Assim, o método corretamente indicou os números ótimos de grupos.

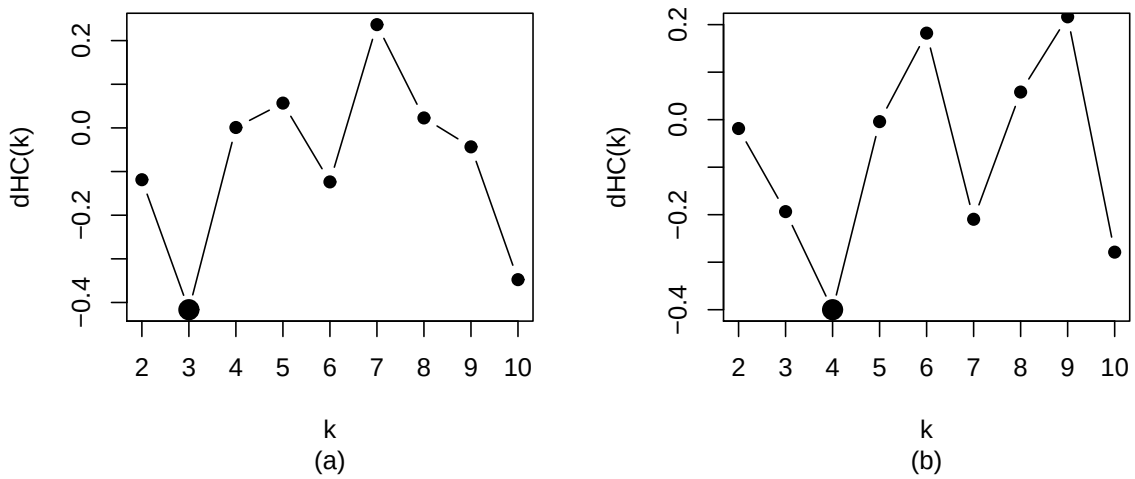

Figura 4.3: (a) apresenta $dHC(k)$ para o caso com três grupos e (b) mostra a mesma medida para quatro, com os pontos maiores indicando os respectivos mínimos.

Como se pôde ver, dHC é uma medida de erro obtida de forma absolutamente empírica. O comportamento crescente observado na entropia foi certamente determinado pelo escolha do algoritmo hierárquico. Uma vez que ele se baseia no príncipio de contínuas uniões de grupos menores para formar novos agrupamentos, reduzindo progressivamente o valor de k até

atingir-se k=1, esse processo acaba por influenciar a evolução dos valores das probabilidades utilizadas para calcular $H(\boldsymbol{C_k})$, $H(\boldsymbol{C_k}|\boldsymbol{C_{k\text{-}1}})$ e $H(\boldsymbol{C_{k+1}}|\boldsymbol{C_k})$.

Em estudos de simulação preliminares, onde o número de grupos existentes na amostra era conhecido, observou-se que dHC às vezes indicava um número ótimo de grupos superior ao real. Nesses casos, ao avaliar-se o comportamento do erro de classificação (CE), definido na Tabela 3.2, percebeu-se que ele poderia ser capaz de detectar a falsa indicação. Assim, decidiu-se propor um outro índice para determinação do k ótimo que utiliza simultaneamente a diferença de entropias e CE.

Para considerar como critério adicional o erro de classificação entre os agrupamentos $\boldsymbol{C_i}$ e $\boldsymbol{C_{i+1}}$, considere o mesmo denotado como $CE(i, i+1)$. Em seguida, tome o algoritmo abaixo:

$j \leftarrow 2$;
Para i variando de 2 a K-1 faça
  INICIO
  se $dHC(i) > dHC(i+1)$ então
    se $dHC(i+1) < dHC(j)$ então
      se $CE(i+1, i+2) < CE(j, j+1)$ então $j \leftarrow i+1$;
  FIM

Ao final do processo, o valor do índice j é justamente o $\hat{k}$ procurado.

Para fins de notação, considere que o algoritmo acima para a escolha do número ótimo de grupos é denominado $dHCE(k)$.

No capítulo 5 serão apresentados alguns estudos de simulação para avaliar estas duas novas propostas de índice para a escolha do k ótimo em comparação com outros métodos.

Capítulo 5

Simulações

Neste capítulo são apresentados estudos de simulação realizados para avaliar o desempenho dos índices: silhueta, Gap UNI, Gap CP, *dHC* e *dHCE* (vide capítulos 3 e 4).

Para fins de comparação, os cenários de simulação aqui utilizados e descritos a seguir, baseiam-se naqueles apresentados nos ítens (a) a (d) do artigo de Tibshirani et al. (2001).

a) *grupo único*[1]: 200 realizações aleatórias de um v.a. com distribuição Uniforme dentro de um cubo de dimensão 10, com aresta de comprimento unitário;

b) *três grupos em duas dimensões*: cada grupo é representado por realizações aleatórias de uma v.a. com distribuição Normal multivariada padrão, centrada em (0,0), (0,5) e (5,-3) e os respectivos tamanhos amostrais são 25, 25 e 50 observações. Um exemplo de banco de dados neste contexto é mostrado na figura 5.1;

c) *quatro grupos em três dimensões*: cada grupo é representado por realizações aleatórias de uma v.a. com distribuição Normal multivariada de variância 0,16$\mathbf{I}$, onde $\mathbf{I}$ é a matriz identidade 3×3. Os centros são aleatoriamente escolhidos como realizações da Normal multivariada (0;5$\mathbf{I}$) e os tamanhos amostrais escolhidos ao acaso entre 25 e 50;

d.1) *quatro grupos em dez dimensões*: cada grupo é representado por realizações aleatórias de uma v.a. com distribuição Normal multivariada de variância 0,05$\mathbf{I}$, onde $\mathbf{I}$ é matriz identidade 10×10. Os centros são aleatoriamente escolhidos como realizações da Normal multivariada (0;1,9$\mathbf{I}$) e os tamanhos amostrais escolhidos ao acaso entre 25 e 50;

d.2) idem cenário (d.1), com a variância das observações modificada para 0,16$\mathbf{I}$;

d.3) idem cenário (d.1), com a variância das observações modificada para 0,25$\mathbf{I}$;

[1]Na literatura, essa situação é também chamada de "modelo nulo".

d.4) idem cenário (d.1), com a variância das observações modificada para 0,50**I**;

d.5) idem cenário (d.1), com a variância das observações modificada para 0,75**I**;

d.6) idem cenário (d.1), com a variância das observações modificada para 1,00**I**;

d.7) idem cenário (d.1), com a variância das observações modificada para 1,50**I**.

Para os cenários (b) e (c), considerou-se como critério de exclusão do conjunto de dados gerado qualquer um onde pelo menos um par de grupos tinha distância minima entre si inferior a 1. Assumiu-se como distância mínima entre dois grupos quaisquer C_k e $C'_{k'}, k \neq k'$, o menor valor da distância euclidiana entre todos os pares de pontos (p_a, p_b), $p_a \in C_k$ e $p_b \in C'_{k'}$. Apesar deste critério não ter sido aplicado aos cenários de (d.1) a (d.7), foi verificado que em todas suas execuções as distâncias eram superiores a 1.

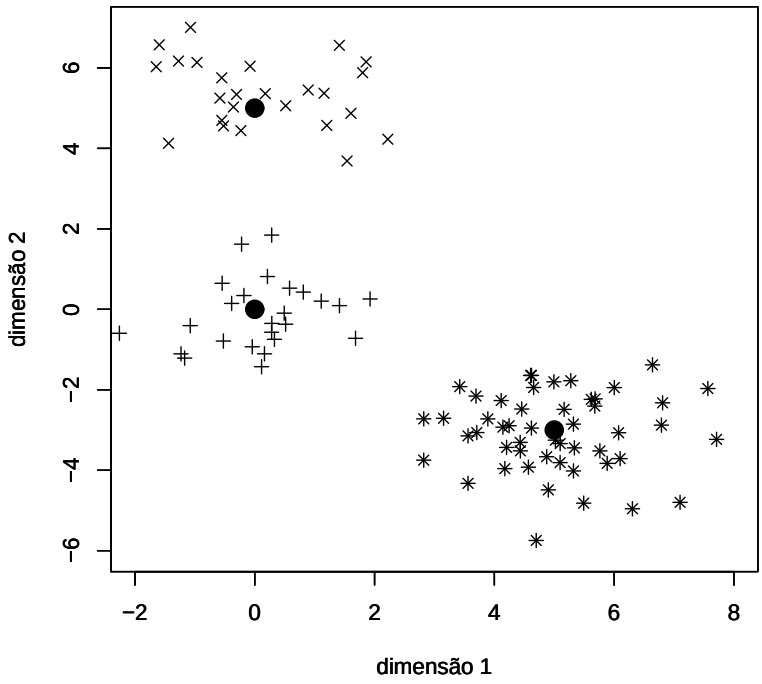

Figura 5.1: Exemplo de conjuntos de dados gerado no cenário (b): os indivíduos de cada grupo estão identificados por diferentes símbolos e seus centros são os pontos redondos.

Após gerar uma amostra A dentro das especificações do cenário sob estudo, com o respectivo conjunto de dados construiu-se agrupamentos contendo 1,2,...,10 grupos ($\boldsymbol{C_1}$, $\boldsymbol{C_2}$,

$\ldots$, $\boldsymbol{C_{10}}$). Para obtê-los, foram utilizados o método hierárquico com ligação média (*average linkage*) e o método de partição k-médias. Em cada cenário foram feitas 1000 execuções de ambos os métodos. Em (b) e (c), aproximadamente 50% dos bancos de dados gerados violaram a condição de distância mínima maior que 1 e portanto foram descartados. Todas as simulações foram feitas no pacote R (http://www.r-project.org/).

Para o método das k-médias utilizou-se a implementação da biblioteca *amap*. As sementes iniciais do processo iterativo foram determinadas aleatoriamente pelo próprio programa. Sabe-se que esse fato tem forte influência nos agrupamentos gerados, como explicado em Everitt et al. (2001), capítulo 5. Assim, para minimizar tal problema, utilizou-se a sugestão de Kapp & Tibshirani (2007): a cada execução da simulação, o algoritmo k-médias foi aplicado cinco vezes, sempre com um novo sorteio das sementes. Dentre os resultados obtidos, escolheu-se aquele com a menor soma das distâncias quadráticas intra-grupos[2]. Além disso, devido a limitações do próprio programa usado, o agrupamento gerado podia, ao final do processo, conter algum grupo vazio. Assim, implementou-se uma rotina que verificava isso e, em caso positivo, descartava o banco. Esses casos descartados não estão incluídos na taxa de 50% citada no parágrafo anterior.

Aplicou-se a cada conjunto de agrupamentos $\boldsymbol{C_1}, \ldots, \boldsymbol{C_{10}}$ os índices citados no início do capítulo e anotou-se o k ótimo indicado em cada um deles. Para o cálculo das estatísticas Gap gerou-se B=5 populações de referência em cada execução (vide seção 3.5).

No cenário (a), para o agrupamento hierárquico, as estatísticas Gap UNI e Gap CP tiveram desempenho muito bom, respectivamente com 85% e 90% de acertos. No método das k-médias seus desempenhos foram inferiores, acertando 78% e 80% das vezes. Os outros índices não são aplicáveis para k=1, mas destaca-se que a silhueta concentrou 94% das indicações para o seu mínimo k=2 no método hierárquico, mostrando-se sensível à estrutura compacta.

No cenário (b), para o agrupamento hierárquico, ambas estatísticas Gap tiveram desempenho excelente, com praticamente 100% de acertos, mas passaram a errar aproximadamente 90% das vezes quando se utilizou o k-médias. Nesse último caso, nota-se que elas concentraram mais de 50% das simulações nos valores de k inferiores ao real estabelecido. O *dHCE* teve um comportamento razoável somente no método hierárquico, com 78.3% de acertos (20 pontos percentuais acima do respectivo índice não corrigido).

[2] A soma das distâncias quadráticas intra-grupos para $\boldsymbol{C}$ é:

$$SSQ_{\boldsymbol{C}} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} d^2 \left(X_{ij}, \bar{X}_i \right),$$

onde X_{ij} é o vetor de dados da j-ésima observação do i-ésimo grupo, $\bar{X}_i$ o vetor de médias do i-ésimo grupo e d o critério de dissimilaridade considerado.

Tabela 5.1: Resultados (%) do estudo de simulação para cenário (a).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0*	93.8	0.0	0.2	0.4	0.1	0.8	0.6	1.0	3.1
Gap UNI	85.1*	13.4	1.3	0.2	0.00	0.0	0.0	0.0	0.0	0.0
Gap CP	90.2*	9.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0*	0.5	2.8	11.1	17.0	19.5	18.1	13.5	9.3	8.2
dHCE	0.0*	30.5	13.4	12.8	10.2	8.4	6.3	6.6	6.1	5.7
***k*-médias**										
Silhueta	0.0*	18.1	2.3	2.0	4.0	7.5	10.1	15.3	17.6	23.1
Gap UNI	78.1*	20.6	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	79.8*	19.5	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0*	0.1	1.0	4.5	7.20	11.7	13.5	19.7	21.2	21.1
dHCE	0.0*	38.0	18.1	13.9	8.6	6.4	3.9	5.1	3.9	2.1

* A coluna marcada indica o número correto de grupos.

Tabela 5.2: Resultados (%) do estudo de simulação para cenário (b).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	9.7	90.3*	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	0.0	0.0	99.9*	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	0.2	0.0	99.8*	0.0	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	6.2	56.3*	0.7	0.9	4.3	6.2	7.3	8.5	9.6
dHCE	0.0	6.2	78.3*	0.7	0.7	2.2	2.8	2.2	3.9	3.0
***k*-médias**										
Silhueta	0.0	19.4	80.5*	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	55.1	31.8	10.8*	2.2	0.1	0.0	0.0	0.0	0.0	0.0
Gap CP	54.5	32.4	12.1*	0.9	0.1	0.0	0.0	0.0	0.0	0.0
dHC	0.0	8.9	4.2*	4.5	9.3	10.7	15.1	13.3	18.5	15.5
dHCE	0.0	17.3	19.7*	8.8	10.6	11.2	11.6	6.8	7.6	6.4

* A coluna marcada indica o número correto de grupos.

Tabela 5.3: Resultados (%) do estudo de simulação para cenário (c).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	1.0	6.5	92.5*	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	3.0	0.2	0.0	96.8*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	8.6	0.5	0.0	90.9*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	0.1	2.4	96.3*	0.0	0.0	0.0	0.0	0.3	0.9
dHCE	0.0	0.1	2.4	97.0*	0.0	0.0	0.0	0.0	0.1	0.4
***k*-médias**										
Silhueta	0.0	3.2	11.1	84.1*	1.1	0.5	0.0	0.0	0.0	0.0
Gap UNI	42.4	25.9	9.5	18.8*	3.4	0.0	0.0	0.0	0.0	0.0
Gap CP	47.7	26.6	9.2	13.9*	2.2	0.3	0.1	0.0	0.0	0.0
dHC	0.0	5.8	16.1	24.2*	3.6	3.2	5.9	10.5	15.1	15.6
dHCE	0.0	18.8	25.5	37.6*	3.3	3.6	3.1	3.2	3.0	1.9

* A coluna marcada indica o número correto de grupos.

No cenário (c), todos os índices tiveram ótimo desempenho quando se aplicou o método hierárquico. Neste caso, *dHCE* foi o melhor (97%) e GAP CP o pior (90,9%). A silhueta apresentou taxas de acertos mais próximas entre os métodos hierárquico (92,5%) e k-médias (84,1%). Neste último método, os demais índices tiveram desempenho ruim e, como observado no cenário (b), as estatísticas Gap selecionaram novamente valores de k inferiores ao verdadeiro valor, em especial k=1 e k=2.

Nas Tabelas de 5.4 a 5.10 encontram-se os resultados para os cenários (d.1) a (d.7). Todos os índices diminuem seu desempenho à medida em que a variância intra-grupos cresce, ou seja, varia-se de (d.1) para (d.7). No k-médias, a silhueta teve a maior porcentagem de acertos em todos os cenários, exceto em (d.7). Suas maiores taxas de acerto foram 93,7%, 86,6% e 75,2%, em (d.1),(d.2) e (d.3), respectivamente. Os demais índices tiveram uma porcentagem de acertos inferior a 60% em todos os casos. Nos cenários de (d.1) a (d.3) no método hierárquico Gap UNI foi o melhor índice, seguido por *dHC* e *dHCE* , sempre empatados no segundo lugar. Entretanto, a diferença entre essas duas melhores posições observada em (d.1) é apenas de 1.5 ponto percentual, reduzindo-se mais à medida em que cresce a variância. Em (d.4), as diferenças de entropia já despontam como os critérios mais eficientes e mantém-se assim até o cenário (d.7). Porém, nos cenários (d.6) e (d.7) estes dois índices têm uma porcentagem de acertos inferior a 70%.

Tabela 5.4: Resultados (%) do estudo de simulação para cenário (d.1).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	0.1	1.5	98.4*	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	0.0	0.0	0.0	100.0*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	2.9	0.0	0.0	97.1*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	0.0	1.5	98.5*	0.0	0.0	0.0	0.0	0.0	0.0
dHCE	0.0	0.0	1.5	98.5*	0.0	0.0	0.0	0.0	0.0	0.0
***k*-médias**										
Silhueta	0.0	0.2	5.1	93.7*	0.80	0.1	0.0	0.1	0.0	0.0
Gap UNI	8.9	5.8	1.9	58.8*	20.5	3.2	0.9	0.0	0.0	0.0
Gap CP	33.8	15.9	3.5	41.7*	5.0	0.1	0.0	0.0	0.0	0.0
dHC	0.0	6.8	28.6	31.6*	1.8	1.7	3.5	7.5	8.6	9.9
dHCE	0.0	15.8	37.4	42.3*	1.6	1.1	0.5	0.8	0.2	0.3

* A coluna marcada indica o número correto de grupos.

A despeito do fato de que a silhueta não foi o melhor índice no método hierárquico em nenhum dos casos, sua taxa de acertos apresenta a menor diferença quando comparada com aquela observada no método das k-médias, em cada um dos cenários utilizados.

Em todos os cenários, para ambos os métodos de construção de agrupamentos, *dHCE* teve uma leve tendência a reduzir o k indicado quando comparado a *dHC*. Mas para efeito de seus desempenhos, a partir do cenário (c) isso não representou uma vantagem ao observar-se apenas as suas taxas de acertos. Conforme dito na seção 2.1, em uma situação com apenas duas variáveis como no cenário (b) deve-se considerar sempre a construção de um gráfico de dispersão para uma avaliação visual, mesmo com o uso de índices como os aqui utilizados.

As figuras 5.2 e 5.3 mostram, para cada um dos métodos de agrupamento considerados, a variação das taxas de acerto dos índices nos cenários (d.1) a (d.7). No caso específico do hierárquico, *dHC* e *dHCE* são claramente os melhores pois seus valores são sempre superiores ou praticamente iguais aos demais. Ainda é possível perceber que, em ambos os métodos de agrupamento utilizados, os valores observados para as diferenças de entropia tendem a diminuir menos à medida em que a variância aumenta.

Tabela 5.5: Resultados (%) do estudo de simulação para cenário (d.2).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	1.0	8.0	91.0*	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	0.0	0.0	0.1	99.9*	0.0	0.0	0.0	0.0	0.0	0.0
GapCP	2.1	0.1	0.1	97.7*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	0.0	1.3	98.7*	0.0	0.0	0.0	0.0	0.0	0.0
dHCE	0.0	0.0	1.3	98.7*	0.0	0.0	0.0	0.0	0.0	0.0
***k*-médias**										
Silhueta	0.0	1.7	10.8	86.6*	0.7	0.0	0.1	0.1	0.0	0.0
Gap UNI	13.7	12.0	9.5	44.3*	17.4	3.1	0.0	0.0	0.0	0.0
Gap CP	31.5	24.2	10.5	28.9*	4.5	0.4	0.0	0.0	0.0	0.0
dHC	0.0	7.3	23.5	33.9*	2.3	2.0	5.1	6.3	10.0	9.6
dHCE	0.0	17.4	34.5	42.4*	1.9	1.0	1.4	0.8	0.4	0.2

* A coluna marcada indica o número correto de grupos.

Tabela 5.6: Resultados (%) do estudo de simulação para cenário (d.3).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	2.6	11.6	85.8*	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	0.0	0.0	0.1	99.9*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	1.0	0.3	0.2	98.5*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	0.0	1.2	98.8*	0.0	0.0	0.0	0.0	0.0	0.0
dHCE	0.0	0.0	1.2	98.8*	0.0	0.0	0.0	0.0	0.0	0.0
***k*-médias**										
Silhueta	0.0	4.9	17.7	75.2*	2.0	0.2	0.0	0.0	0.0	0.0
Gap UNI	17.2	18.2	15.7	33.8*	12.9	2.2	0.0	0.0	0.0	0.0
Gap CP	34.9	27.7	14.0	20.2*	2.9	0.3	0.0	0.0	0.0	0.0
dHC	0.0	6.6	22.9	29.7*	5.0	4.9	4.9	6.6	8.4	11.0
dHCE	0.0	18.6	36.2	37.2*	3.4	2.2	0.8	0.9	0.2	0.5

* A coluna marcada indica o número correto de grupos.

Tabela 5.7: Resultados (%) do estudo de simulação para cenário (d.4).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	12.5	22.7	64.8*	0.0	0.0	0.0	0.0	0.0	0.0
Gap UNI	0.1	0.6	5.6	93.7*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	1.9	1.0	6.4	90.7*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	0.1	3.9	94.2*	1.1	0.3	0.0	0.0	0.2	0.2
dHCE	0.0	0.4	5.0	94.0*	0.3	0.0	0.0	0.0	0.1	0.2
***k*-médias**										
Silhueta	0.0	11.9	26.0	60.1*	1.6	0.2	0.2	0.0	0.0	0.0
Gap UNI	27.8	32.0	19.7	14.3*	5.6	0.5	0.1	0.0	0.0	0.0
Gap CP	40.5	37.3	14.5	6.1*	1.4	0.2	0.0	0.0	0.0	0.0
dHC	0.0	6.6	22.6	29.5*	6.9	5.1	5.6	6.3	6.8	10.6
dHCE	0.0	18.5	37.6	35.3*	3.9	1.8	1.4	1.1	0.2	0.2

* A coluna marcada indica o número correto de grupos.

Tabela 5.8: Resultados (%) do estudo de simulação para cenário (d.5).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	18.8	26.7	53.8*	0.6	0.1	0.0	0.0	0.0	0.0
Gap UNI	0.9	4.3	15.8	79.0*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	5.7	5.1	16.8	72.4*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	1.4	9.2	83.5*	3.3	1.6	0.5	0.4	0.1	0.0
dHCE	0.0	2.6	12.1	83.5*	1.1	0.4	0.2	0.1	0.0	0.0
***k*-médias**										
Silhueta	0.0	21.2	30.1	46.8*	1.7	0.2	0.0	0.0	0.0	0.0
Gap UNI	39.7	37.0	16.3	5.6*	1.4	0.0	0.0	0.0	0.0	0.0
Gap CP	48.4	38.2	10.4	2.6*	0.4	0.0	0.0	0.0	0.0	0.0
dHC	0.0	5.5	23.4	26.0*	10.2	6.1	5.5	6.0	7.6	9.7
dHCE	0.0	17.7	39.0	32.1*	6.0	3.2	0.7	0.9	0.2	0.2

* A coluna marcada indica o número correto de grupos.

Tabela 5.9: Resultados (%) do estudo de simulação para cenário (d.6).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	28.1	28.0	41.2*	2.6	0.1	0.0	0.0	0.0	0.0
Gap UNI	4.3	10.0	27.5	58.2*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	11.5	10.2	28.2	50.1*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	1.9	15.9	68.0*	7.6	2.9	1.9	0.4	0.5	0.9
dHCE	0.0	5.8	21.7	67.7*	3.1	0.7	0.4	0.1	0.2	0.3
***k*-médias**										
Silhueta	0.0	30.3	31.5	37.0*	1.1	0.1	0.0	0.0	0.0	0.0
Gap UNI	47.9	38.3	9.8	3.1*	0.7	0.2	0.0	0.0	0.0	0.0
Gap CP	57.0	35.2	6.6	1.0*	0.2	0.0	0.0	0.0	0.0	0.0
dHC	0.0	6.4	20.8	26.4*	11.0	4.8	5.7	8.2	8.6	8.1
dHCE	0.0	19.9	36.2	32.4*	7.0	1.7	1.2	0.9	0.3	0.4

* A coluna marcada indica o número correto de grupos.

Tabela 5.10: Resultados (%) do estudo de simulação para cenário (d.7).

Método	Número de grupos (k)									
	1	2	3	4	5	6	7	8	9	10
Hierárquico										
Silhueta	0.0	51.1	24.8	20.6*	2.8	0.7	0.0	0.0	0.0	0.0
Gap UNI	19.5	25.6	34.9	20.0*	0.0	0.0	0.0	0.0	0.0	0.0
Gap CP	28.4	27.8	27.0	16.8*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	4.8	22.0	39.5*	13.4	7.3	4.5	3.9	2.6	2.0
dHCE	0.0	18.5	32.2	37.0*	6.7	2.2	0.9	0.6	0.9	1.0
***k*-médias**										
Silhueta	0.0	40.3	30.6	28.6*	0.5	0.0	0.0	0.0	0.0	0.0
Gap UNI	58.6	33.9	6.6	0.8*	0.1	0.0	0.0	0.0	0.0	0.0
Gap CP	63.2	31.9	4.5	0.4*	0.0	0.0	0.0	0.0	0.0	0.0
dHC	0.0	6.2	19.5	27.2*	9.1	5.5	5.2	8.0	8.4	10.9
dHCE	0.0	18.4	35.2	36.1*	6.7	2.5	0.9	0.1	0.0	0.1

* A coluna marcada indica o número correto de grupos.

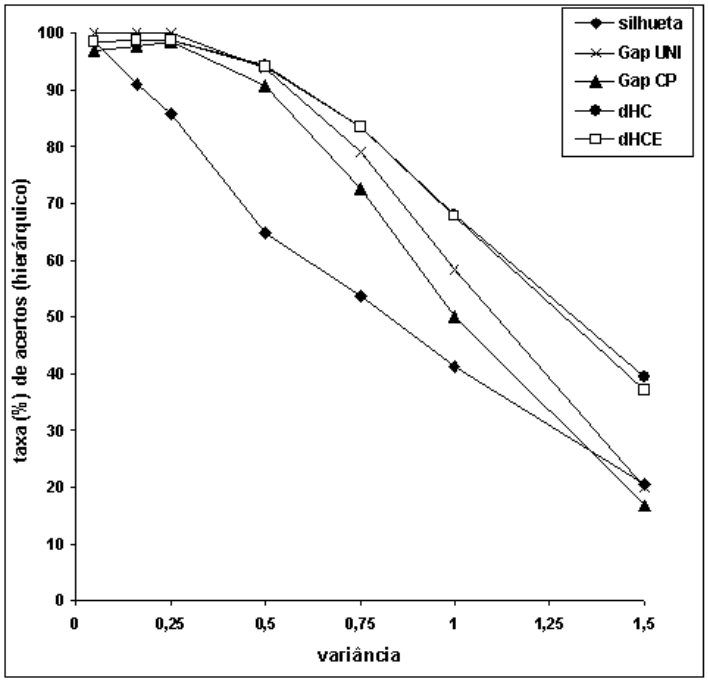

Figura 5.2: Taxas de acerto dos índices estudados à medida em que se aumenta a variância intra-grupos (cenários (d.1) a (d.7)) - método hierárquico de agrupamento.

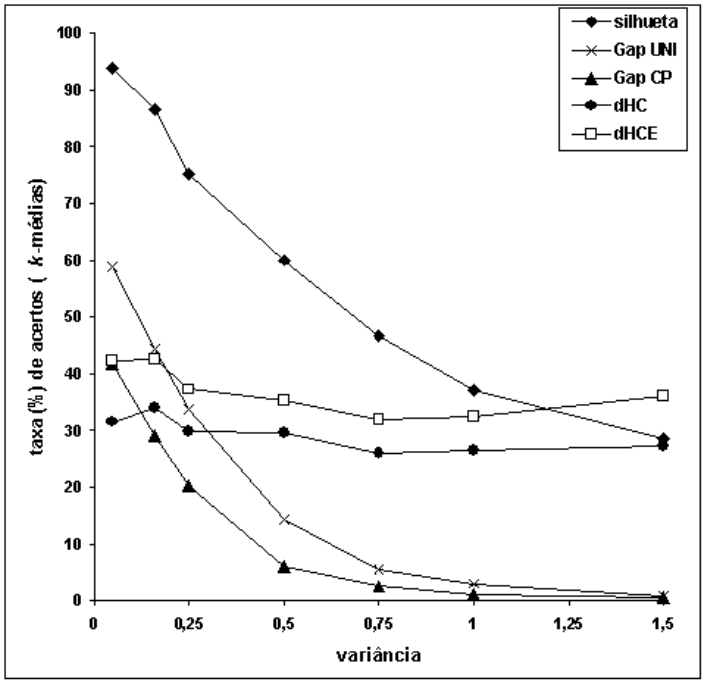

Figura 5.3: Taxas de acerto dos índices estudados à medida em que se aumenta a variância intra-grupos (cenários (d.1) a (d.7)) - método de agrupamento das k-médias.

Em relação aos resultados dos estudos de simulação de Tibshirani et al. (2001), que utilizaram apenas o método hierárquico com 50 realizações de cada cenário, algumas considerações são feitas a seguir.

No cenário (a), o artigo original apresenta taxas de acerto iguais a 98% e 100% para Gap UNI e Gap CP, superiores àquelas encontradas no presente trabalho, respectivamente 85,1% e 90,2%. Como os parâmetros da simulação foram idênticos, acredita-se que tal variação seja intrínseca aos índices pois em outro artigo, a saber Kapp & Tibshirani (2007), o mesmo cenário foi utilizado e, com 50 realizações, observou-se 84% e 92% de acertos.

Em (b), as estatísticas Gap tiveram desempenhos semelhantes. Porém, a silhueta acertou 100% das vezes em Tibshirani et al. (2001) enquanto que aqui essa taxa foi igual a 90,3%. Tal diferença pode ser devida à aplicação do critério de exclusão de grupos com distância mínima inferior a 1 ou à variação do próprio índice.

Já no cenário (c), apenas Gap UNI teve resultados semelhantes. Para Gap CP e a silhueta, Tibshirani et al. (2001) apresenta 84% em ambos índices, enquanto que o presente trabalho encontrou 90,9% e 92,5%, respectivamente. Duas possíveis justificativas para essas desigualdades são: o uso de diferentes variâncias intra-grupos e, novamente, características intrínsecas dos próprios índices.

Não serão feitas comparações com os cenários de (d.1) a (d.7) pois Tibshirani et al. (2001) utilizaram um critério de exclusão nas suas simulações. Além disso, aqui introduziu-se um "escalonamento" na variância, sendo esta uma modificação feita para avaliar a sensibilidade dos índices a incrementos na dispersão dos pontos dentro dos grupos gerados.

No que diz respeito aos tempos de execução computacional de cada um dos índices avaliados, realizou-se um procedimento simples para compará-los, apenas com intenção descritiva ou exploratória. É importante lembrar que em aplicações práticas, por exemplo na área de genética, é comum se encontrar situações com tamanho amostral e dimensões muito superiores às utilizadas aqui.

Considerando-se o cenário (d.4), realizou-se 10 simulações em cada método de obtenção de agrupamentos, variando-se o n total da amostra nas seguintes quantidades: 150, 300, 450, 600, 750, 900, 1050, 1200, 1350 e 1500. Apenas o agrupamento com quatro grupos foi construído a cada execução. As unidades amostrais foram distribuídas nos quatro grupos nas seguintes proporções: $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{3}$ e $\frac{1}{3}$. Por exemplo, no primeiro caso $n_1 = 25$, $n_2 = 25$, $n_3 = 50$ e $n_4 = 50$. O comando *date* do R foi utilizado para registrar os inícios e términos de cada processamento. Como sua precisão é limitada em segundos, deve-se assumir nos tempos mostrados na Tabela 5.11 a margem de erro $\pm 1s$.

No método hierárquico observa-se um crescimento muito rápido nos tempos das estatís-

Tabela 5.11: Tempos (segundos) para obter os índices variando-se o tamanho da amostra.

Método	Tamanho da amostra (n)									
	150	300	450	600	750	900	1050	1200	1350	1500
Hierárquico										
Silhueta	< 1	< 1	1	1	1	1	1	2	3	3
Gap UNI	4	14	41	91	173	288	444	650	915	1240
Gap CP	3	13	40	92	174	286	443	654	917	1236
dHC	< 1	< 1	< 1	< 1	1	< 1	< 1	< 1	1	1
dHCE	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
***k*-médias**										
Silhueta	< 1	< 1	1	1	< 1	1	1	2	3	3
Gap UNI	2	4	7	11	14	21	25	34	42	48
Gap CP	1	3	6	9	13	15	21	25	31	35
dHC	1	1	< 1	< 1	1	< 1	1	< 1	< 1	1
dHCE	<1	< 1	< 1	< 1	< 1	1	< 1	1	< 1	< 1

ticas Gap. Em ambas, na maior amostra, eles ultrapassaram 1200 segundos! Certamente, eles são muito influenciados pelo alto custo desta classe de métodos pois, a cada população de referência, sua respectiva árvore hierárquica deve ser obtida para cálculo da função W_k, e isso implica em um maior tempo de processamento, como já citado no capítulo 2. O tempo da silhueta aumenta muito pouco com n e as diferenças de entropia mantém-se estáveis.

Já no método das k-médias os tempos das estatísticas Gap também aumentam, mas de forma bem menos contundente. Seus máximos foram observados para n=1500, ambos abaixo de um minuto. Há uma leve tendência de Gap UNI demandar mais tempo do que Gap CP. Do ponto de vista de seus algoritmos, aparentemente não há razão para que isso aconteça. Mas, por conveniência de programação, o uso do k-médias para obtenção das partições da amostra original foi implementado dentro do cálculo de Gap UNI. Seus tempos, portanto, incluem também as construções do agrupamentos originais sem que isso, no entanto, invalide o aumento dos mesmos, apenas atenua. A silhueta, *dHC* e *dHCE* tiveram desempenhos praticamente iguais ao do método hierárquico, o que era esperado pois o método em si não modifica seus algoritmos. As diferenças de entropia utilizam apenas as alocações dos elementos de A nos k grupos e, então, dependem só de n e k. A silhueta, além disso, utiliza as medidas originais de todas as observações, e depende também de número de variáveis (r).

Para encerrar o capítulo, apresenta-se a seguir alguns comentários.

Independentemente do método de agrupamento usado, as estatísticas Gap são boas alter-

nativas para se verificar a possibilidade de k=1. Se essa hipótese é descartada e o método de agrupamento utilizado é o k-médias, deve-se aplicar somente a silhueta pois foi o único índice com boas taxas de acerto nessas circuntâncias. Porém, deve-se ter em mente que seu desempenho é sensível ao aumento da variância intra-grupos. No caso do método hierárquico, em nenhum momento ela superou os demais índices.

Não foram observadas grandes diferenças entre Gap UNI e Gap CP, mas isso já era esperado pois o uso das componentes principais foi proposto para o caso de grupos alongados. No presente estudo, somente grupos de forma aproximadamente esférica foram gerados.

Quando $r \geq 3$, as diferenças de entropia tiveram um desempenho semelhante ou superior às estatísticas Gap, com a vantagem de serem menos sensíveis ao aumento da variância intra-grupos e de menor custo computacional. Em particular, esta última característica é especialmente útil quando se está em processos de mineração de dados, onde a construção e avaliação de agrupamentos muitas vezes é aplicada de forma intensiva. Ainda com respeito ao número de variáveis, o aumento destas igualou os desempenhos de *dHCE* e *dHC* e, assim, este último pode ser preferido uma vez que sua implementação é mais simples.

A robustez dos índices quanto a perturbações nos dados devidas a variáveis de diferentes distribuições, *outliers* etc, deve ser objeto de estudos de simulação futuros. Para tanto, a consulta de Milligan (1980), Milligan (1985) e Qiu & Joe (2006) é recomendada pois eles discutem e apresentam procedimentos para introduzir "ruídos" nos bancos de dados de teste.

Capítulo 6

Aplicações

Neste capítulo serão apresentadas duas aplicações práticas, ambas da área de genética. A primeira aplica a variação da informação (VI) e a segunda aborda a questão da escolha do número ótimo de grupos com o uso de *dHC*.

6.1 Comparação de duas classificações de linfoma

Nesta seção será apresentado um exemplo de aplicação para a variação da informação. Os dados utilizados fazem parte de um estudo sobre o perfil do linfoma grande difuso da célula B (*diffuse large B-cell lymphoma*, DLBCL), que é a forma mais comum de linfomas malignos em adultos. Seus mecanismos de ação e desenvolvimento nos portadores ainda não são plenamente conhecidos, razão pela qual ela é objeto de diversas pesquisas.

O estudo original foi publicado por Monti, Savage, Kutok, Feuerhake, Kurtin, Mihm, Wu, Pasqualucci, Neuberg, Aguiar, Cin, Ladd, Pinkus, Salles, Harris, Dalla-Favera, Habermann, Aster, Golub & Shipp (2005). A amostra consistiu de material genético de tumores extraídos por biópsias de 176 pacientes diagnosticados e sem tratamento, dos quais um nível de expressão quantitativo foi obtido para aproximadamente 33.000 genes. Dentre estes, foram selecionados 2.118 para a construção do banco de dados.

Como existem estudos que indicam heterogeneidade clínica e genética do DLBCL, pode ser útil a identificação das chamadas "assinaturas" moleculares de tumores com características similares. A classificação destes em subconjuntos diferentes poderiam indicar os caminhos para as possíveis intervenções terapêuticas.

Assim, com o objetivo de avaliar a possível presença de diferentes classes de tumores entre os pacientes, foram contruídos agrupamentos contendo $k = 2, \ldots, 9$ grupos, utilizando-se três

métodos diferentes: hierárquico, mapas auto-organizáveis (*self-organizing maps*, SOM) e um método bayesiano, esse último proposto por Cheeseman & Stutz (1996). Com base nos resultados dessas classificações, Monti et al. (2005) concluíram que k=3 era a melhor opção. Em seguida, a partir dos valores observados das expressões gênicas em cada grupo, os autores caracterizaram seus perfis biológicos, sendo então denominados:

- OxPhos: ligado a processos de fosforilação oxidativa, entre outros;
- BCR/Prol: altos níveis de receptores de células B, entre outros; e
- Reação do paciente (*host response*, HR): ação de processos inflamatórios contra doença.

Foram classificados 50 pacientes no grupo OxPhos, 77 no BCR/Prol e 49 no HR.

Monti et al. (2005) decidiram comparar seus resultados com o agrupamento obtido por um outro método de classificação, afim de investigar possíveis relações entre eles. Então, estes autores reclassificaram sua amostra de tumores segundo a assinatura da célula de origem (*cell-of-origin (COO) signature*), proposta no trabalho de Wright, Tan, Rosenwald, Hurt, Wiestner & Staudt (2003). Esta metodologia está baseada num escore de predição para os seguinte tipos biológicos: célula B ativada (*activated B cell*, ABC), célula B de centro germinal (*germinal center B cell*, GCB) e outros (tipos não identificados).

A Tabela 6.1 apresenta o cruzamento da classificação dos 176 pacientes de sua amostra segundo essas duas propostas.

Tabela 6.1: Distribuição dos 176 pacientes de Monti et al. (2005) nas classificações por células de origem e perfis biológicos

células	**perfil biológico**			
de origem	OxPhos	BCR/Prol	HR	**total**
ABC	9	18	8	35
GCB	23	41	15	79
Outros	18	18	26	62
total	50	77	49	176

A distribuição dos pacientes não evidencia nenhuma correlação entre as duas formas de classificação. A casela central, com 41 pacientes, foi a que concentrou a maioria dos pacientes em relação às marginais. No caso das linhas ela possui 52% do total (41÷79) e em relação ao

total da sua coluna esse valor aumenta para 53% (41÷77). Todas as outras caselas ficaram iguais ou inferiores a esta.

Fazendo observações de natureza descritiva, como acima, e biológica, Monti et al. (2005) concluíram que ambas classificações revelam aspectos bem diferentes da doença. Portanto, não concorrem entre si e podem ser usadas simultaneamente para obter informações adicionais sobre a mesma. Entretanto, acredita-se que seria útil poder quantificar o quão diferentes esses procedimentos de classificação são.

Então, é proposto o cálculo da variação da informação para esta tabela. Obteve-se VI=0,398, valor este que já está padronizado pelo máximo $\log n$ para que VI fique no intervalo [0,1], como explicado na propriedade 3 do capítulo 4. Ele está razoavelmente distante de 0 e assim pode-se afirmar que este índice também indica duas classificações diferentes entre si. Naturalmente, a simples observação da variação da informação não permite comentar as relações biológicas existentes entre os dois agrupamentos, como fizeram Monti et al. (2005) em seu artigo, pois eles levaram em conta os valores individuais anotados em cada casela da Tabela 6.1. A vantagem de se utilizar a VI é poder fornecer um parâmetro de comparação entre os dois sistemas de classificação apresentados. Evidentemente, ela pode ser útil para comparar também classificações obtidas por outros sistemas.

Se outras propostas relacionadas à identificação de assinaturas moleculares do linfoma grande difuso da célula B estão disponíveis na literatura, e utilizam genes disponíveis no banco de dados do presente estudo, a aplicação descrita acima pode ser estendida com o intuito de se verificar quão diferentes são as classificações obtidas com os sistemas disponíveis.

6.2 Escolha do número ótimo de grupos num conjunto de pacientes leucêmicos

Os dados utilizados formam parte de um estudo sobre classificação de câncer baseado nas expressões gênicas de pacientes. O trabalho original foi dividido em duas partes: (1) levantamento de genes importantes na identificação do tipo de câncer por meio de uma amostra de treinamento e construção de uma classe preditiva, e (2) validação por meio de uma amostra independente. Utilizou-se os dados referentes à primeira parte, isto é, da amostra de treinamento, pois na mesma justifica-se o uso da análise de agrupamentos para separar os pacientes segundo as variáveis observadas, no caso as expressões gênicas. O conjunto de dados foi obtido em uma base pública (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi) e originalmente publicado por Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, Bloomfield & Lander (1999). A análise foi realizada no pacote R.

Amostras da medula de 38 pacientes foram extraídas no momento do diagnóstico da doença, dos quais 27 eram portadores de leucemia aguda linfoblástica (ALL) e 11 portadores de leucemia aguda mielóide (AML). A distinção entre os tipos ALL e AML é um aspecto fundamental para o tratamento. Obteve-se níveis de expressão quantitativa para 6817 genes de interesse no estudo. Com essas medidas e utilizando o método SOM, os autores concluíram que seria possível separar a amostra inicial de 38 pacientes em duas classes. Dentre os 6817 genes, Golub et al. (1999) selecionaram os 50 mais correlacionados com os dois tipos de leucemia e com eles procederam na segunda parte da pesquisa. No presente trabalho, procurou-se confirmar esta escolha com o uso da análise de agrupamentos e os índices avaliados no capítulo 5.

Com os dados referentes aos 38 pacientes e os 50 genes selecionados por Golub et al. (1999), contruiu-se grupos com a técnica hierárquica de agrupamento, usando ligação média (*average linkage*). Os índices aplicados para encontrar o melhor número de grupos foram a silhueta, Gap UNI, Gap CP e *dHC* , com k=1,...,6 grupos. Acreditou-se que $k \geq 7$ poderia gerar situações com grupos contendo poucos elementos, dificultando a criação das classes preditivas citadas acima.

Usualmente em uma análise de agrupamentos, apresenta-se no ínicio do resultados o respectivo dendrograma. Entretanto, no presente trabalho procurou-se dar um destaque maior para os índices utilizados. Assim, a seguir serão discutidas as conclusões obtidas somente com a aplicação dos mesmos e, posteriormente, será apresentado o dendrograma.

Todos os índices recomendaram a escolha de k=2 grupos. Assim, após obter este agrupamento, 29 pacientes foram classificados no grupo I, dos quais 27 tem diagnóstico de ALL.

O grupo II foi formado por nove pacientes portadores de AML (vide Tabela 6.2).

Tabela 6.2: Distribuição dos pacientes nos dois grupos obtidos, segundo o tipo de leucemia

	ALL	AML	total
Grupo I	27	2	29
Grupo II	0	9	9
total	27	11	38

A Tabela 6.3 mostra os valores de silhueta média e *dHC* para o intervalo de k considerado. Observa-se que, segundo o critério proposto em Kaufman & Rousseeuw (1990), o valor 0,48 encontrado para o primeiro índice indica uma estrutura fraca de agrupamento. Na figura 6.1 vê-se como ficou a disposição geral das silhuetas. O baixo coeficiente de silhueta do grupo 2 (0,24) indica a pouca homogeneidade de seus integrantes em relação às variáveis observadas.

Tabela 6.3: Valores obtidos de silhueta média e *dHC*.

	número de grupos (k)				
	2	3	4	5	6
Silhueta	0,48	0,43	0,42	0,38	0,27
dHC	-0,21	-0,05	-0,02	0,17	-0,16

O índice *dHC* atinge o mínimo de -0,21 para $k = 2$ e é crescente até $k = 5$, voltando a decrescer para -0,16 quando a amostra é dividida em seis grupos ($k = 6$).

No capítulo 5 foi observada uma variação na taxa de acertos de Gap UNI e Gap CP para indicar o k ótimo, quando se comparou os resultados dos estudos de simulação do presente trabalho com outros da literatura, realizados sob condições idênticas. Assim, durante a construção do programa da presente aplicação, os algoritmos das estatísticas Gap foram executados 15 vezes. Em duas delas, Gap CP indicou como melhor opção $k = 1$ e nas restantes obteve-se $k = 2$. Devido a esse fato, assumiu-se que Gap CP recomenda o uso de dois grupos. Gap UNI indicou $k = 2$ em todas elas. A título de ilustração na figura 6.2 apresenta-se os gráficos do número de grupos (k) versus a estatística Gap para as referências uniforme e componentes principais, calculadas em uma das 15 execuções do algoritmo onde ambas indicaram dois grupos como a melhor escolha.

A figura 6.3 exibe o dendrograma. Para visualizar-se os grupos acima mencionados, uma linha tracejada separa a amostra em $k = 2$ grupos. É direto ver que as expressões

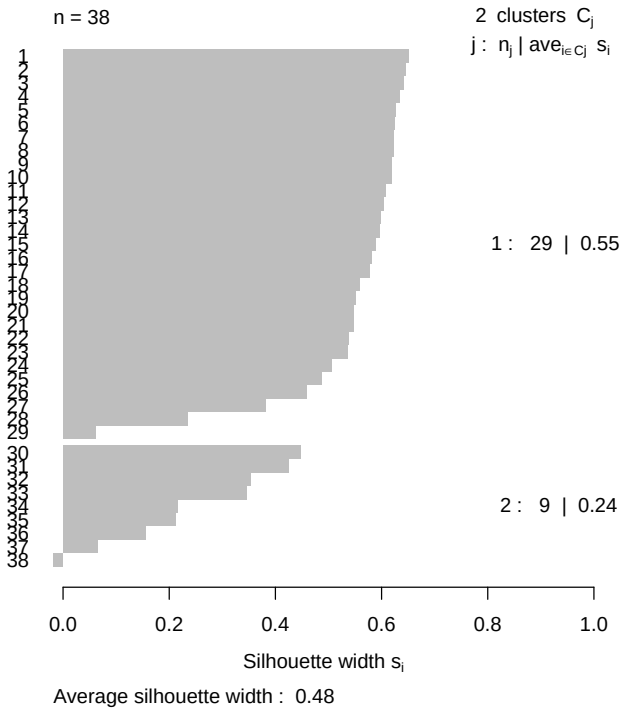

Figura 6.1: Silhueta dos dados de leucemia.

gênicas consideradas no estudo contém informação relevante para distinguir os dois tipos de câncer, pois os pacientes AML estão, em sua maioria, concentrados no lado direito, oposto aos portadores de ALL. Outra observação importante é que a maior parte dos pacientes do Grupo I conecta-se no dendrograma antes daqueles do Grupo II. Isso significa que as distâncias entre eles, que estão expressas no eixo y, são menores do que aquelas entre os pacientes do Grupo II, e isso implica que sua dispersão intra-grupo é menor. Como o Grupo I é formado majoritariamente por portadores de ALL (mais de 90%), pode-se concluir que esta variedade de leucemia apresenta medidas mais homogêneas do que AML, segundo as variáveis observadas. É importante lembrar que essa última observação tira proveito da opção pela ligação média (*average linkage*) na obtenção do agrupamento via dendrograma. Neste caso, as coordenadas dos centróides dos sucessivos subgrupos formados, utilizadas para se fazer as ligações entre eles, são as médias de seus respectivos elementos, para cada variável.

Apesar de *dHC* recomendar o uso de k=2, onde ela atingiu o mínimo igual a -0,21, nota-se que para k=6 seu valor está próximo, sendo igual a -0,16. Assim, decidiu-se investigar também o agrupamento quando o mesmo possui seis grupos (vide Tabela 6.4).

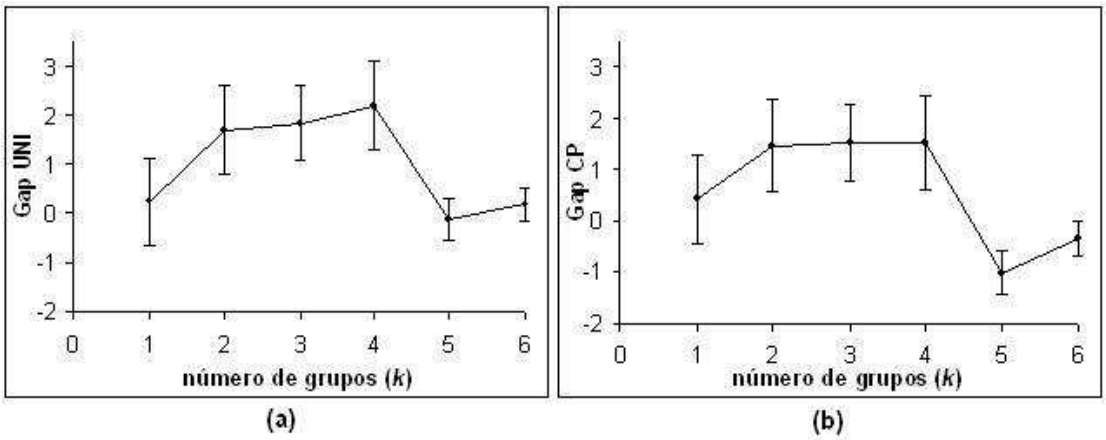

Figura 6.2: Gráficos da estatística Gap versus o número de grupos aplicada aos dados de leucemia: (a) Gap UNI e (b) Gap CP.

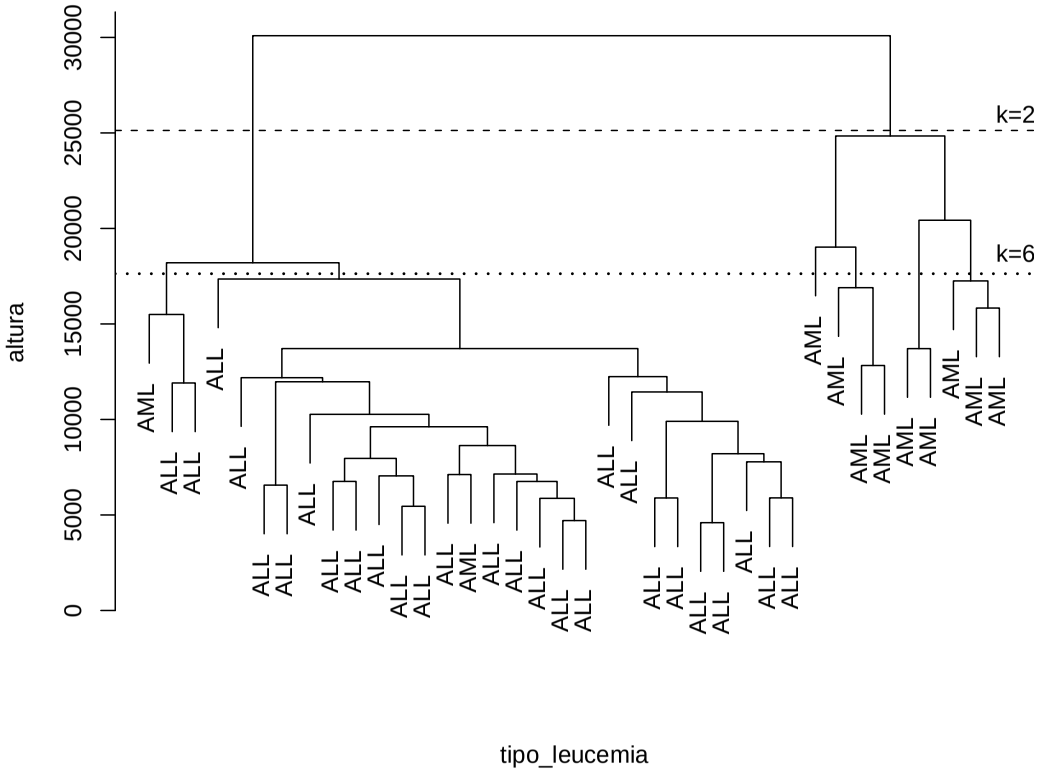

Figura 6.3: Dendrograma para a aplicação com dados de leucemia.

Tabela 6.4: Distribuição dos pacientes nos seis grupos obtidos, segundo o tipo de leucemia

	ALL	AML	total
Grupo I	25	1	26
Grupo II	2	1	3
Grupo III	0	3	3
Grupo IV	0	3	3
Grupo V	0	1	1
Grupo VI	0	2	2
total	27	11	38

É interessante notar que o Grupo I original, obtido quando k=2, mantém quase igual nesta nova configuração e apenas dois portadores de ALL e um de AML são deslocados para outros grupos. Já o Grupo II anterior é completamente subdividido em novas classes. Isso é devido ao fato das dispersões das variáveis utilizadas no estudo serem diferentes entre os dois tipos de leucemia, como citado acima. Pode-se observar essa divisão também no dendrograma da figura 6.3, onde uma linha pontilhada foi acrescentada para k=6.

Nos dois gráficos para as estatísticas Gap, mostrados na figura 6.2, nota-se um comportamento não monótono nas funções de Gap UNI e Gap CP, a partir de k=5. Segundo Tibshirani et al. (2001), isso pode ocorrer quando há pequenos subgrupos dentro de grupos maiores, estes últimos bem separados.

Portanto, além dos valores recomendados pelos índices, também é interessante observar o que ocorre em todo o intervalo considerado para os valores de k, pois outras configurações de agrupamento podem revelar características de interesse do problema sob investigação. No presente caso, a silhueta não forneceu nenhuma indicação neste sentido.

Por meio das análises aqui apresentadas se chegou às mesmas conclusões de Golub et al. (1999). Os índices detectaram k=2 como o melhor agrupamento a partir de um número menor de variáveis. Além disso, *dHC* , Gap UNI e Gap CP também sinalizaram outro corte interessante no dendrograma.

Capítulo 7

Considerações Finais

A análise de agrupamentos é uma ferramenta estatística para uso essencialmente descritivo e exploratório. Seus resultados podem auxiliar na elaboração de hipóteses sobre o assunto estudado, na compreensão da estrutura global da população da qual se retirou a amostra e em um primeiro passo na obtenção de modelos de previsão e classificação. O presente trabalho enfocou dois de seus aspectos: a escolha do número ótimo de grupos para segmentar uma amostra e a comparação de duas classificações para avaliar sua semelhança.

Foi proposto um novo índice (dHC) para a escolha do número ótimo de grupos que utiliza as diferenças de entropias, cujo desempenho nos estudos de simulação mostrou-se encorajador. Quando utilizado o método hierárquico com três ou mais variáveis, a nova proposta foi bastante eficiente, no sentido de detectar o número correto de grupos simulado, quando comparada com dois índices conhecidos: a estatística Gap e a silhueta.

É útil lembrar a recomendação, explicada no capítulo 2.2, do uso simultâneo de mais de um índice para a escolha do número de grupos uma vez que nenhuma técnica acerta 100% das vezes.

No que diz respeito à comparação de agrupamentos, apresentou-se a medida chamada "variação da informação", também baseada na entropia. Originalmente publicada em Meilă (2002), suas propriedades foram aqui demonstradas e uma aplicação prática realizada.

O leitor tem a oportunidade de aplicar as técnicas aqui exploradas pois os códigos em R utilizados estão disponíveis em http://br.geocities.com/hideste/tiriba.html. Assim, espera-se que o presente trabalho venha a se tornar uma ferramenta de uso imediato.

Como perspectiva de estudos futuros, seria interessante avaliar a nova medida em situações onde os grupos não possuem forma esférica, como as utilizadas nas simulações, ou na presença de "ruídos" como: *outliers*, variáveis sem informação etc. Além disso, poderia-se investigar modificações em sua fórmula com o intuito de melhorar o desempenho quando o

método de obtenção de agrupamentos é supervisionado, como o k-médias. Em bancos de dados de grandes dimensões, como os encontrados nos processos de mineração de dados, esta classe de métodos é preferida dada sua grande vantagem em termos de processamento computacional. Assim, as possibilidades de aplicação do índice aqui proposto aumentariam consideravelmente.

Quanto à variação da informação (VI), seria interessante realizar aplicações também com o uso simultâneo de outros índices de comparação de agrupamentos, como por exemplo aqueles apresentados na seção 3.2, pois assim poderia-se obter indicações das situações em que seu uso seria mais recomendado.

Referências Bibliográficas

Ben-Hur, A., Elisseeff, A. & Guyon, I. (2002), A stability based method for discovering structure in clustered data, *in* 'Pacific Symposium on Biocomputing', pp. 6–17.

Bussab, W., Andrade, D. F. & Miazaki, E. (1990), *Introdução à Análise de Agrupamentos*, 9^o SINAPE - Simpósio Nacional de Probabilidade e Estatística, São Paulo. Texto de minicurso.

Cheeseman, P. & Stutz, J. (1996), Bayesian classification (autoclass): Theory and results, *in* U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, eds, 'Advances in Knowledge Discovery and Data Mining', AAAI Press, Menlo Park. Disponível em: http://www.cs.ubc.ca/~murphyk/Teaching/Papers/autoclass_kdd95.ps.

Cormack, R. M. (1971), 'A review of classification', *Journal of The Royal Statistical Society A* **134**(3), 321–367.

Everitt, B. S. (1979), 'Unresolved problems in cluster analysis', *Biometrics* **35**, 169–181.

Everitt, B. S., Landau, S. & Leese, M. (2001), *Cluster Analysis*, 4^{th} edn, Arnold, UK.

Fowlkes, E. B. & Mallows, C. L. (1983), 'A method for comparing two hierarchical clusterings', *Journal of the American Statistical Association* **78**(383), 553–569.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Gordon, A. D. (1981), *Classification*, Chapman and Hall, London.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, NY.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of Classification* **2**, 193–218.

Kapp, A. & Tibshirani, R. (2007), Using the in-group proportion to estimate the number of clusters in a dataset. Artigo não publicado, disponível em http://www.stanford.edu/~akapp/work.html.

Kaufman, L. & Rousseeuw, P. J. (1990), *Finding Groups in Data - An Introduction to Cluster Analysis*, Wiley, NY.

Larsen, B. & Aone, C. (1999), Fast and effective text mining using linear-time document clustering, *in* 'Proceedings of the conference on Knowledge Discovery and Data Mining', pp. 16–22.

Loganantharaj, R., Cheepala, S. & Clifford, J. (2006), 'Metric for measuring the effectiveness of clustering of DNA microarray expression', *BMC Bioinformatics* **7**. (Suppl 2):S5.

Manning, C. D. & Schütze, H. (2003), *Foundations of Statistical Natural Language*, The MIT Press, Massachusetts.

Meilă, M. (2002), Comparing clusterings, `UW` statistics technical 418, University of Washington. Uma cópia do relatório técnico pode ser obtido no endereço: http://www.stat.washington.edu/www/research/reports/2002/tr418.ps.

Meilă, M. (2005), Comparing clusterings - an axiomatic view, *in* 'Procceedings of the 22^{nd} International Conference on Machine Learning', Bonn, Germany.

Meilă, M. & Heckerman, D. (2001), 'An experimental comparison of model-based clustering methods', *Machine Learning* **42**(1/2), 9–29.

Milligan, G. W. (1980), 'An examination of the effect of six types of error perturbation on fifteen clustering algorithms', *Psychometrika* **45**(3), 325–342.

Milligan, G. W. (1985), 'An algorithm for generating artificial test clusters', *Psychometrika* **50**, 123–127.

Milligan, G. W. (1996), Clustering validation: results and implications for applied analyses, *in* P. Arabie, L. J. Hubert & G. D. Soete, eds, 'Clustering and Classification', World Scientific, Singapore, pp. 341–375. Reimpressão de 1999.

Milligan, G. W. & Cooper, M. C. (1985), 'An examination of procedures for determining the number of clusters in a data set', *Psychometrika* **50**(2), 159–179.

Mirkin, B. (1996), *Mathematical Classification and Clustering*, Kluwer Academic Press.

Monti, S., Savage, K. J., Kutok, J. L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R. C. T., Cin, P. D., Ladd, C., Pinkus, G. S., Salles, G., Harris, N. L., Dalla-Favera, R., Habermann, T. M., Aster, J. C., Golub, T. R. & Shipp, M. A. (2005), 'Molecular profiling of diffuse large b-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response', *Blood* **105**(5), 1851–1861.

Qiu, W. & Joe, H. (2006), 'Generation of random clusters with specified degree of separation', *Journal of Classification* **23**, 315–334.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association* **66**(336), 846–850.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics* **20**, 53–65.

Ruspini, E. H. (1970), 'Numerical methods for fuzzy clustering', *Information Sciences* **2**, 319–350.

Shannon, C. E. & Weaver, W. (1949), *The Mathematical Theory of Communications*, University of Illinois Press,Urbana.

Sugar, C. A. & James, G. M. (2003), 'Finding the number of clusters in a dataset:an information-theoretic approach', *Journal of the American Statistical Association* **98**(463), 750–763.

Tan, P.-N., Steinbach, M. & Kumar, V. (2006), *Introduction to Data Mining*, Addison-Wesley.

Teixeira, T. S. (2003), Número Ótimo de aglomerados estocásticos, Master's thesis, Instituto de Matemática e Estatística/Universidade de São Paulo.

Thorndike, R. L. (1953), 'Who belongs in the family ?', *Psychometrika* **18**(4), 267–276.

Tibshirani, R., Walther, G. & Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of The Royal Statistical Society B* **63**, 411–423.

van Dongen, S. (2000), Performance criteria for graph clustering and markov cluster experiments, Technical report ins-r0012, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands.

van Rijsbergen, C. J. (1980), *Information Retrieval*, Butterworths, London.

Wallace, D. L. (1983), 'Comment', *Journal of the American Statistical Association* **78**(383), 569–576.

Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A. & Staudt, L. M. (2003), 'A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma', *Proceedings of the National Academy of Sciences of USA* **100**(17), 9991–9996.