UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

Validação de Algoritmos de Agrupamento

Katti Faceli
André C P L F de Carvalho
Marcílio Carlos Pereira de Souto

Nº 254

RELATÓRIOS TÉCNICOS

São Carlos – SP
Fev./2005

Validação de Algoritmos de Agrupamento

Katti Faceli
André C.P.L.F. de Carvalho

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
e-mail: {katti, andre}@icmc.usp.br

Marcílio Carlos Pereira de Souto

Universidade Federal do Rio Grande do Norte
Departamento de Informática e Matemática Aplicada - DIMAp
Campus Universitario
59072-970 - Natal, RN, Brazil
e-mail: marcilio@dimap.ufrn.br

Resumo

A validação do resultado de um agrupamento, em geral, é feita com base em índices estatísticos, que julgam, de uma maneira qualitativa, o mérito das estruturas encontradas. Um índice quantifica alguma informação a respeito da qualidade de um agrupamento. A maneira pela qual um índice é aplicado para validar um agrupamento é dada pelo critério de validação. Existem três tipos de critérios para investigar a validade de um agrupamento: critérios internos, externos e relativos. Também são três os tipos de estrutura que podem ser avaliados na validação de um agrupamento: hierarquias, partições e clusters individuais. Este relatório apresenta uma descrição dos três critérios de validação, bem como das metodologias e índices mais comumente empregados na validação de partições.

Palavras-Chave: Validação de agrupamento, critérios de validação.

Fevereiro 2005
Draft impresso em 23 de fevereiro de 2005

Sumário

Lista de Figuras

Lista de Tabelas

Notação

n:	Número de padrões
d:	Número de dimensões
$X_{n \times d}$:	Matriz de padrões
x_i:	Padrão (linha da matriz de padrões)
$X_{i,j}$:	Valor da característica j do padrão x_i
$S_{n \times n}$:	Matriz de similaridade
$S_{i,j}$:	Similaridade entre os padrões x_i e x_j
C_i:	Cluster i
v_i:	Centro do cluster i
x_{0_i}:	Centróide do cluster i
K:	Número de clusters
K_{min}:	Número mínimo de clusters
K_{max}:	Número máximo de clusters
n_K:	Número de padrões no cluster K
$\delta(C_i, C_j)$:	Distância entre os clusters C_i e C_j
$\Delta(C_i)$:	Distância intra-cluster (dentro do cluster)
U:	Partição *fuzzy*

$u_{i,j}$:	Grau de pertinência *fuzzy* do padrão x_i ao cluster C_j
$d(x_i, x_j)$:	Distância entre os padrões x_i e x_j
P_c:	Partição obtida com a aplicação de um algoritmo de agrupamento
P_r:	Partição real, ou construída com base em uma intuição sobre a estrutura real dos dados
K_P:	Número de clusters da partição P

1 Introdução

A avaliação do resultado de um agrupamento deve ser objetiva com o propósito de determinar se os clusters são significativos, ou seja, se a solução é representativa para o conjunto de dados analisado. Uma estrutura de agrupamento é válida se não ocorreu por acaso, ou se é "rara" em algum sentido, já que qualquer algoritmo de agrupamento encontrará clusters, independentemente se existe ou não similaridade nos dados. Entretanto, se essa similaridade existe, alguns algoritmos podem encontrar clusters mais adequados que outros. Algumas técnicas para a validação dos resultados de técnicas de agrupamento têm sido discutidas na literatura, tais como (He 1999): testes de significância nas variáveis utilizadas para criar os clusters, replicação, testes de significância nas variáveis externas e procedimentos de Monte Carlo. Um estudo sobre processos de validação de agrupamentos pode ser encontrado em (Jain & Dubes 1988; Gordon 1999; Halkidi et al. 2001).

A validação do resultado de um agrupamento, em geral, é feita com base em índices estatísticos, que julgam, de uma maneira qualitativa, o mérito das estruturas encontradas. Um índice quantifica alguma informação a respeito da qualidade de um agrupamento. A maneira pela qual um índice é aplicado para validar um agrupamento é dada pelo critério de validação. Assim, um critério de validação expressa a estratégia utilizada para validar uma estrutura de agrupamento, enquanto que um índice é uma estatística pela qual a validade é testada. Existem três tipos de critérios para investigar a validade de um agrupamento: critérios internos, externos e relativos.

Três tipos de estrutura podem ser avaliados na validação de um agrupamento: hierarquias, partições e clusters individuais. Os três critérios de validação podem ser empregados na validação de qualquer um dos tipos de estrutura possíveis. Todos os índices e abordagens descritos neste trabalho enfocam a validação de partições. Deste ponto em diante, os índices que são mais comumente empregados em critérios externos, internos e relativos serão denominados simplificadamente de índices externos, internos e relativos, respectivamente.

As abordagens de validação de agrupamentos tradicionais, descritas na literatura, obtêm melhores resultados quando os clusters são geralmente compactos. Para aplicações que envolvem clusters de formas arbitrárias, os critérios de validação tradicionais (variância, densidade e continuidade, separação) não são suficientes. É necessário o desenvolvimento de medidas de qualidade que avaliem a qualidade de um particionamento levando em consideração a qualidade dentro dos clusters, a separação inter-clusters e a geometria dos clusters, utilizando conjuntos de pontos representativos ou mesmo curvas multidimensionais ao invés de um único ponto representativo (Halkidi et al. 2002a).

Em seguida serão resumidos alguns trabalhos tradicionais e algumas abordagens recentes de validação. Os critérios de validação serão detalhados na Seção 2. Nas seções 3, 4 e 5 serão detalhados alguns dos índices empregados respectivamente com critérios relativos, internos e externos, bem como as metodologias mais comuns para a aplicação dos índices.

1.1 Trabalhos que analisam índices de validação

Milligan (1996) resume as abordagens de testes de hipótese e análise de replicação para a validação de agrupamentos. Segundo ele, os testes para verificar se existem estruturas significativas nas partições encontradas pelos algoritmos escapa a muitos pesquisadores aplicados, porque eles assumem que realmente existem clusters nos dados. Milligan (1996) ressalta ainda que não devem ser utilizadas as técnicas de teste de hipótese tradicionais como ANOVA, MANOVA e análise discriminante diretamente nas variáveis utilizadas para determinar o agrupamento, argumentando a invalidade de tal análise. Ele discute brevemente as abordagens válidas que envolvem os critérios externos e internos de validação.

Sarle & Kuo (1993) discutem os problemas de várias abordagens de teste de hipótese existentes para a determinação do número de clusters apropriado. Segundo Sarle & Kuo (1993), possivelmente a melhor hipótese nula para este tipo de análise é que os dados sejam amostrados de uma distribuição uniforme.

Existem vários trabalhos na literatura descrevendo e comparando diversas técnicas e índices de validação. Os trabalhos de Jain & Dubes (1988), Halkidi et al. (2001), Halkidi et al. (2002a) e Halkidi et al. (2002b) constituem uma revisão geral das diversas abordagens de validação de agrupamento. Jain & Dubes (1988) descrevem como os vários tipos de critérios podem ser empregados para validar cada tipo de estrutura (hierarquias, partições e clusters individuais). Halkidi et al. (2001), Halkidi et al. (2002a) e Halkidi et al. (2002b) enfocam as análises mais comuns feitas com cada tipo de critério, principalmente utilizando técnicas de Monte Carlo.

Diversos artigos realizam uma análise comparativa entre índices de validação, avaliando índices que são comumente empregados em critérios externos, internos e relativos.

Milligan et al. (1983) analisam o efeito da dimensionalidade, tamanho dos clusters e número de clusters em agrupamentos, utilizando os índices externos Rand (R), correção do Rand proposta por Morey & Agresti (1984), Jaccard (J) e Folkes e Mallows (FM). Hubert & Arabie (1985) discutem vários índices externos para comparar partições e propõem uma correção para o índice Rand (CR). Milligan & Cooper (1986) compararam cinco índices externos na análise de agrupamentos hierárquicos. Os índices avaliados foram Rand, Jaccard, Fowlkes e Mallows e as correções do índice Rand propostas por Morey & Agresti (1984) e Hubert & Arabie (1985). Este estudo indicou que o índice Rand corrigido (CR) de Hubert & Arabie (1985) é o mais apropriado para o tipo de análise realizada.

Milligan & Cooper (1985) comparam trinta índices internos de validação para estimar o número de clusters apropriado, utilizando quatro algoritmos de agrupamento hierárquico. Essa avaliação foi baseada em conjuntos de dados pequenos (cerca de 50 padrões) com clusters bem separados (Halkidi et al. 2001). Segundo Sarle & Kuo (1993), os critérios que obtiveram melhor desempenho no trabalho de Milligan & Cooper (1985) são apropriados apenas para clusters que são compactos ou levemente alongados, de preferência aproximadamente multivariados normais.

Pal & Bezdek (1995) avaliam o papel do parâmetro m do algoritmo FCM (*Fuzzy c-means*) na validação das partições geradas por esse algoritmo. Para isso, comparam os

índices relativos de validação *fuzzy* coeficiente de partição (PC), entropia da partição (PE), os índices de Xie-Beni, XB e XB estendido, e o índice de Fukuyama e Sugeno, FS, concluindo que o índice de Xie-Beni (XB) foi o mais estável dentre os investigados.

1.2 Abordagens recentes

Recentemente, têm sido descritas diversas abordagens para validação de agrupamento. Algumas abordagens de validação interna e relativa utilizam análise de Monte Carlo, outras *bootstrapping* e outras ainda se baseiam em análise de replicação. Existem ainda outras propostas que não fazem uso dessas metodologias para a validação, especialmente no caso dos critérios relativos.

1.2.1 Abordagem que utiliza Monte Carlo

Tibshirani et al. (2001b) propõem um índice interno, a estatística *Gap*, para estimar o número de clusters e o compara com os índices KL, proposto por Krzanowski & Lai (1985), H, proposto por Hartigan (1975) e silhueta, proposto por Rousseeuw (1987). Este método emprega análise de Monte Carlo simulando múltiplos conjuntos de dados de uma distribuição referência e desenvolvendo uma regra de rejeição para o teste estatístico.

1.2.2 Abordagem baseada em análise de replicação

Dudoit & Fridlyand (2002) propõem o método de reamostragem baseado em predição, *Clest*, para estimar o número de clusters de um agrupamento. Esse método é uma generalização e extensão da análise de replicação. O número de clusters é estimado comparando o valor observado da estatística empregada (R, J ou FM) para cada K com o seu valor esperado sob um distribuição nula apropriada com $K = 1$. O número estimado de clusters é definido como sendo o valor de K correspondente à maior evidência significativa contra a hipótese nula de $K = 1$.

1.2.3 Abordagens que empregam bootstrapping

Além da sua utilização para a construção de um modelo nulo citada anteriormente, existem outras formas de aplicação de *bootstrapping* para validação de um agrupamento.

Law & Jain (2003) propõem uma forma de validar uma partição interpretando um algoritmo de agrupamento como um estimador estatístico e avaliando a variabilidade desse estimador por meio de *bootstrapping*. Se uma partição é válida, sua variabilidade deve ser baixa. O procedimento proposto pode ser utilizado para comparar a validade de partições obtidas por um algoritmo de agrupamento utilizando diferentes valores para seus parâmetros, diferentes algoritmos de agrupamento ou diferentes medidas de similaridade usadas para o agrupamento. Se um conjunto de dados não tem estrutura, o método proposto não pode identificar tal fato explicitamente, embora uma variabilidade alta para diferentes valores de K possa sugerir a falta de estrutura. A medida proposta é independente da qualidade do ajuste do algoritmo aos dados.

O trabalho de Kerr & Churchill (2001) utiliza *bootstrapping* para avaliar a estabilidade da associação de um padrão a um cluster. Esses autores aplicam *bootstrapping* para avaliar

a estabilidade de um agrupamento, aplicando o método em agrupamentos de genes com base em seus valores de expressão. Para isso, utilizam um modelo de análise de variância (ANOVA) para descrever dados de *microarray*. Segundo os autores, a forma exata do modelo ANOVA que deve ser utilizado depende dos dados que estão sendo empregados, ou seja, cada conjunto de dados deve ser avaliado individualmente para se determinar as fontes de variação presentes e construir o modelo apropriado. O método de *boostrapping* é utilizado criando um grande número de conjuntos de dados simulados, de acordo com o modelo ANOVA dos dados, e aplicando o algoritmo de agrupamento a cada um desses conjuntos (os agrupamentos obtidos dessa forma podem ser vistos como uma amostra dos agrupamentos que estão próximos do agrupamento original no espaço de todos os agrupamentos possíveis). O casamento de um padrão com uma partição é dito 95% estável se ocorre na análise dos dados reais e em pelo menos 95% dos agrupamentos gerados com o método *bootstrapping*.

1.2.4 Outras abordagens

Dom (2002) propõem uma medida de validação externa baseada em teoria da informação e a compara com os índices R, J, FM e Γ.

As abordagens de Yeung et al. (2001) e Tibshirani et al. (2001a) se baseiam na idéia de que, se um agrupamento reflete uma estrutura verdadeira, um preditor construído com base nos clusters desse agrupamento deve estimar precisamente os rótulos dos clusters de novas amostras de teste (Jiang et al. 2004).

Yeung et al. (2001) propõem uma metodologia, utilizando uma figura de mérito (*Figure of merit*), FOM, para avaliar os resultados de algoritmos de agrupamentos baseada no poder preditivo dos clusters resultantes. A metodologia é proposta no contexto de agrupamento de genes com base no seu valor de expressão em várias condições experimentais. O índice Rand corrigido (CR) é utilizado para validar externamente o índice proposto. A metodologia proposta não é apropriada para comparações entre agrupamentos com diferentes números de clusters. A FOM é calculada com base no resultado da aplicação do algoritmo de agrupamento a amostras dos dados que excluem um dos atributos.

Tibshirani et al. (2001a) propõem a medida força de predição, PS, para avaliar quantos grupos podem ser preditos a partir dos dados e quão bem eles podem ser preditos. Os autores aplicam a medida em dados de expressão gênica. O índice PS é comparado com os índices Gap, CH e KL. O objetivo dessa abordagem é utilizar um conjunto de treinamento e um conjunto de teste para medir o quão bem os centros dos clusters obtidos a partir do conjunto de treinamento predizem a co-pertinência dos padrões do conjunto de teste.

Pakhira et al. (2004) propõem o índice PBM e sua versão *fuzzy* e os comparam com outros índices. Pakhira et al. (2004) utilizam o índice XB para partições *crisp*, comparando os resultados dos índices D, DB, XB e PBM nos resultados obtidos com os algoritmos k-means e EM. Além disso, comparam o índice XB com a versão *fuzzy* do índice proposto, $PBMF$, nos resultados obtidos com o algoritmo FCM (*Fuzzy c-mean*). Em ambos os casos, o índice proposto apresentou os melhores resultados.

2 Critérios de validação

Como já foi dito, um critério de validação expressa a estratégia utilizada para validar um agrupamento, enquanto que um índice é uma estatística pela qual a validade é testada. Ou, de outra forma, o critério de validação indica a maneira pela qual um índice é aplicado para validar um agrupamento. Existem três tipos de critérios para investigar a validade de um agrupamento:

Critérios relativos: Comparam diversos agrupamentos para decidir qual deles é o melhor em algum aspecto (qual é mais o estável ou qual é o mais adequado aos dados, por exemplo). Podem ser utilizados para comparar diversos algoritmos de agrupamento ou para determinar o valor mais apropriado de algum parâmetro do algoritmo aplicado, como o número de clusters. Por exemplo, pode-se medir quantitativamente qual dentre dois algoritmos melhor se ajusta aos dados ou determinar o número de clusters mais apropriado para um agrupamento feito com um determinado algoritmo.

Critérios internos: Medem a qualidade de um agrupamento com base apenas nos dados originais (matriz de padrões ou matriz de similaridade). Por exemplo, um critério interno pode medir o grau em que uma partição obtida por um algoritmo de agrupamento é justificado pela matriz de similaridade.

Critérios externos: Avaliam um agrupamento de acordo com uma estrutura pré-especificada, imposta ao conjunto de dados e que reflete a intuição do pesquisador sobre a estrutura presente nos dados. Essa estrutura pré-especificada pode ser uma partição que se sabe previamente existir nos dados, ou um agrupamento construído por um especialista da área com base em conhecimento prévio. Por exemplo, um critério externo pode medir o grau de correspondência entre o número clusters obtidos com o agrupamento e os rótulos dos dados conhecidos previamente.

Três tipos de estrutura podem ser avaliados: hierarquias, partições e clusters individuais. Os três critérios de validação podem ser empregados na validação de qualquer um dos tipos de estrutura possíveis. Neste relatório, as abordagens descritas se restringem a avaliação de partições.

Os critérios relativos de validação têm como objetivo encontrar o melhor agrupamento que um algoritmo pode obter sob certas suposições e valores para seus parâmetros ou o algoritmo mais apropriado para os dados/estruturas analisados.

Existem vários índices que podem ser empregados com critérios relativos. Alguns dos índices mais comuns e também alguns propostos recentemente são apresentados na Seção 3.1. Esses índices, em geral, podem também ser empregados em critérios internos (Jain & Dubes 1988). O que distingue a utilização de um índice em um ou outro critério é a maneira como o índice é aplicado. A forma mais comum de aplicação de um índice com um critério relativo, consiste do cálculo do seu valor para vários agrupamentos que estão sendo comparados, obtendo-se uma seqüência de valores. O melhor agrupamento é determinado pelo valor que se destaca nessa seqüência, como o valor máximo, mínimo ou inflexão na curva do gráfico construído com a seqüência.

A Figura 1 resume a metodologia mais comum de aplicação do critério relativo de validação. Como indica a Figura 1, nesse critério, vários algoritmos, ou um mesmo algoritmo com diferentes valores para seus parâmetros, são aplicados ao mesmo conjunto de dados. O índice é calculado para cada uma das partições obtidas. Esses valores do índice são comparados, em geral com o auxílio de um gráfico, para se determinar o melhor algoritmo ou o melhor valor para um ou mais parâmetros de um algoritmo.

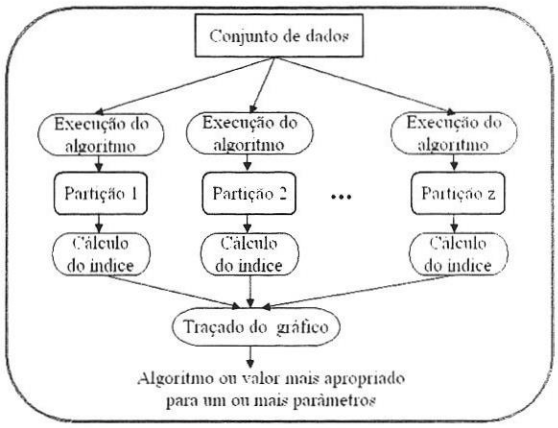

Figura 1: Critério relativo de validação.

Além da forma mais comum de validação relativa previamente descrita, existem outras abordagens que podem ser consultadas na literatura. Algumas dessas abordagens serão descritas na Seção 3.2. Serão detalhadas a análise de replicação, proposta por McIntyre & Blashfield (1980) e Morey et al. (1983), a abordagem de Law & Jain (2003), que calcula a variabilidade do algoritmo utilizando *bootstrapping* e as abordagens de Yeung et al. (2001) e Tibshirani et al. (2001a), que se baseiam no poder preditivo do algoritmo.

Os critérios externos e internos são baseados em testes estatísticos e têm um alto custo computacional (Halkidi et al. 2001). Seu objetivo é medir o quanto o resultado obtido confirma uma hipótese pré-especificada. Neste caso, são utilizados testes de hipótese para determinar se uma estrutura obtida é apropriada para os dados. Isto é feito testando se o valor do índice utilizado é extraordinariamente grande ou pequeno. Isto requer o estabelecimento de uma população base ou de referência. O mesmo índice pode ser utilizado em um critério externo e interno, embora as distribuições de referência do índice sejam diferentes (Jain & Dubes 1988).

A proposição de um índice para validação é fácil, porém é muito difícil estabelecer *thresholds* com base nos quais se possa afirmar que o valor do índice é grande ou pequeno o suficiente para se considerar o agrupamento "raro", ou válido.

Os índices de validação, ou estatísticas, são funções dos dados que contêm informações úteis, como o erro quadrático de um agrupamento ou a compactação de seus clusters. Um índice é uma variável aleatória. Sua distribuição descreve a freqüência relativa com a qual seus valores são gerados sob alguma hipótese. Uma hipótese é uma afirmação sobre a freqüência relativa de eventos no espaço amostral que expressa um conceito. Esse conceito pode ser a ausência de estrutura nos dados.

No caso de validação de agrupamentos, a hipótese nula, H_0, é uma afirmação de não estrutura ou aleatoriedade dos dados. Jain & Dubes (1988) e Gordon (1999) descrevem algumas hipóteses nulas comumente utilizadas para a validação de agrupamentos. Além disso, Jain & Dubes (1988) discutem aplicações de cada uma dessas hipóteses.

A seleção da hipótese nula depende do tipo dos dados e do aspecto dos dados que estão sendo analisados. Jain & Dubes (1988) resumem também os principais aspectos relacionados à utilização de um índice de validação:

Definição do índice: o índice deve fazer sentido intuitivamente, deve ter uma base teórica e deve ser prontamente computável.

Distribuição de probabilidade base: uma distribuição base é uma distribuição derivada de uma população que não possui estrutura. Uma população referência é definida ou implicada pela distribuição base.

Teste para verificar estrutura não aleatória: o valor de um índice de validação é comparado com um *threshold* que estabelece um dado nível de significância. O *threshold* é definido a partir da distribuição base, que raramente é conhecida na teoria.

Teste para verificar um tipo de estrutura: a habilidade do índice de validação em recuperar uma estrutura conhecida indica seu poder estatístico. A escolha da estrutura depende da aplicação particular.

As figuras 2 e 3 resumem, respectivamente, os critérios externos e internos de validação. Conforme dito anteriormente, nesses critérios é utilizado um teste de hipótese que depende da distribuição do índice sob uma hipótese nula, H_0. A diferença entre esses critérios está nas informações utilizadas. Nos critérios externos pode-se observar a utilização de uma partição dos dados conhecida previamente, chamada aqui "partição real", no cálculo do índice. Já nos critérios internos, o cálculo do índice depende somente dos próprios dados, seja na forma de matriz de padrões (conjunto de dados original) ou de matriz de similaridade.

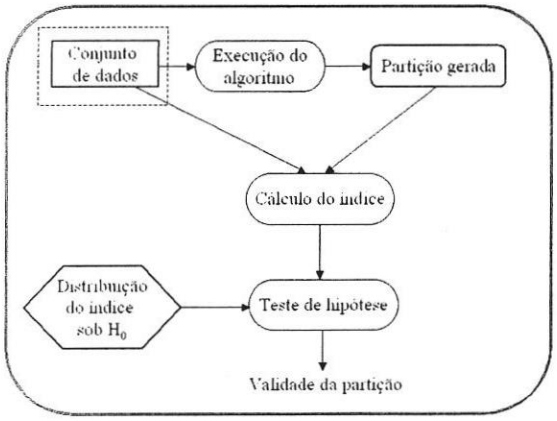

Figura 2: Critério interno de validação.

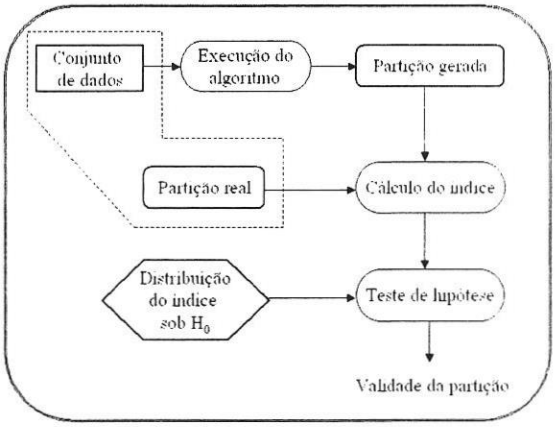

Figura 3: Critério externo de validação.

Um grande problema em validação externa ou interna de agrupamentos é o estabelecimento da distribuição dos índices (estatísticas) sob a hipótese nula e conseqüentemente a determinação dos *thresholds* que dizem se uma partição é adequada de acordo o índice. Os testes reais de validação são geralmente definidos utilizando ferramentas estatísticas, como análise de Monte Carlo e *bootstrapping*.

Análise de Monte Carlo: É um método para estimar parâmetros e taxas de probabilidade por meio de amostragem por computador. É utilizada quando tais valores são difíceis ou impossíveis de serem calculados diretamente.

Em validação de agrupamentos, uma das formas mais comuns de utilização de análise de Monte Carlo é no estabelecimento da distribuição referência de um índice sob a hipótese nula. Inicialmente, é gerada uma grande quantidade de conjuntos de dados sintéticos de acordo com a distribuição considerada na hipótese nula H_0. Cada um desses conjuntos é agrupado e o valor do índice é calculado em cada caso. Com esses valores do índice é traçado um gráfico de dispersão, que é uma aproximação da função de densidade de probabilidade do índice. Dado o valor do índice para o agrupamento que está sendo validado e a distribuição estimada, determina-se se a hipótese H_0 deve ser aceita ou rejeitada. Com a utilização da análise de Monte Carlo, o *threshold* para determinar se o valor do índice é grande ou pequeno o suficiente (ou seja, rejeitar H_0) pode ser visto como um intervalo de valores, ao invés do valor único obtido quando a distribuição sob H_0 é conhecida (Jain & Dubes 1988). A Figura 4 resume esta forma de aplicação de análise de Monte Carlo para validação de agrupamento.

A definição do modelo nulo não é uma tarefa trivial. Existe uma grande quantidade de tipos de modelos nulos, cada um com suas vantagens e desvantagens (Gordon 1996). Segundo Filho (2003), (Gordon 1996) sugere a utilização de mais de um modelo nulo, o que torna o processo de validação ainda mais complexo e demorado. Segundo Jain & Dubes (1988), a parte mais difícil na análise de Monte Carlo é a obtenção de uma amostra de distribuição arbitrária. Geralmente existem geradores de números aleatórios que geram amostras pseudo aleatórias de distribuição $U(1,0)$ e existem técnicas para transformar essas amostras uniformes em amostras de outras

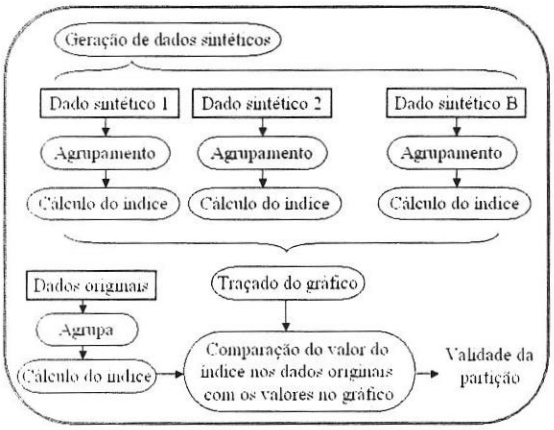

Figura 4: Análise de Monte Carlo para validação de um agrupamento.

distribuições.

Outro problema da análise de Monte Carlo é a necessidade de grande quantidade de recursos computacionais, uma vez que é necessário um grande número de replicações para construir a distribuição de referência. Nos últimos anos, esse problema tem sido minimizado devido à crescente disponibilidade de tais recursos.

***Bootstrapping*:** As técnicas de *bootstrapping* utilizam reamostragem dos dados, com substituição, para criar um conjunto de dados "falso", que é utilizado como se fosse uma replicação de um experimento de Monte Carlo (Jain & Dubes 1988). Neste caso, as amostras *bootstrap* são utilizadas para construir o modelo nulo. As amostras podem ser obtidas reamostrando-se os padrões ou os atributos do conjunto de dados. Com *bootstrapping* evita-se os problemas da análise de Monte Carlo relacionados com a escolha do modelo nulo. Existem outras formas de aplicação de *bootstrapping* em validação, algumas das quais são empregadas em índices descritos posteriormente.

3 Índices e metodologias empregados em critérios relativos

Como já foi dito, o objetivo da validação relativa é encontrar um agrupamento que melhor se ajuste aos dados, a partir de vários agrupamentos obtidos com um algoritmo sob certas suposições e valores para seus parâmetros, ou encontrar o algoritmo mais apropriado para agrupar os dados/estruturas analisados.

A forma mais comum de utilização dos critérios relativos é na determinação do número mais adequado de clusters. Neste caso, o algoritmo de agrupamento é executado para todos os possíveis números de clusters, K, entre K_{min} e K_{max} fornecidos. Em seguida, os valores do índice obtidos a partir dessas execuções são apresentados como função de K. O melhor número de clusters é dado pelo mínimo, máximo ou inflexão na curva observada. Halkidi et al. (2002b) descrevem resumidamente a utilização de índices em critérios relativos para determinar outros parâmetros dos algoritmos de agrupamento.

A Seção 3.1 apresenta uma grande variedade de índices, tanto tradicionais quanto propostos recentemente, que podem ser empregados com critérios relativos.

Além dessa metodologia para validação relativa, existem algumas outras abordagens descritas na literatura, algumas das quais serão apresentadas na Seção 3.2.

3.1 Índices empregados em critérios relativos

Existe um grande número de índices de validação utilizados em critérios relativos para avaliar um agrupamento. Alguns dos índices tradicionais mais comuns são:

- Estatística Hubert Γ modificada (Jain & Dubes 1988),
- Família de índices Dunn (Halkidi et al. 2002b),
- Índice Davies-Bouldin (DB) (Jain & Dubes 1988),
- Silhuetas (Rousseeuw 1987),
- Índice de Calinski-Harabasz (CH) (Calinski & Harabasz 1974),
- Índice de Krzanowski e Lai (KL) (Krzanowski & Lai 1985) e
- Índice de Hartigan (H) (Hartigan 1975; Tibshirani et al. 2001b).

Outras abordagens mais recentes incluem:

- Índice SD (Halkidi et al. 2002b),
- Índice S_Dbw (Halkidi & Vazirgiannis 2001),
- Índice PBM (Pakhira et al. 2004).

Os índices D, DB e SD, resumidos por Halkidi et al. (2001), medem a qualidade dos clusters gerados a partir dos conceitos de homogeneidade e separação. Esses índices medem em que grau os padrões são similares dentro de um cluster e diferentes em clusters separados.

Além desses índices, existem numerosos índices que são apropriados para a avaliação de agrupamentos *fuzzy*. Nesse tipo de agrupamento, cada padrão apresenta um grau de pertinência em relação a cada cluster, ao invés de ser associado unicamente a um cluster. Alguns dos índices mais comuns empregados para validação de agrupamentos *fuzzy* podem ser citados:

- Coeficiente de partição (PC) (Pal & Bezdek 1995),
- Entropia da partição (PE) (Pal & Bezdek 1995),
- Índice de Xie-Beni (XB) (Xie & Beni 1991; Pal & Bezdek 1995),
- Índice de Xie-Beni estendido (Pal & Bezdek 1995),
- Índice de Fukuyama-Sugeno (FS) (Pal & Bezdek 1995),
- Separação da partição (S) (Yang & Wu 2001) e
- Índice $PBMF$ (Pakhira et al. 2004).

Dentre os índices citados para validação de partições *fuzzy*, os índices PC e PE utilizam apenas o valor de pertinência dos padrões aos clusters para calcular seu valor. Já os demais índices utilizam, além do valor de pertinência, o próprio conjunto de dados.

3.1.1 Estatística Hubert Γ modificada

A estatística Hubert Γ modificada é dada pela Equação 1, em que $M = n(n-1)/2$, S é a matriz de proximidade do conjunto de dados e Q é uma matriz $n \times n$, na qual cada célula (i, j) é a distância entre os pontos representativos (v_{C_i}, v_{C_j}) dos clusters dos padrões x_i e x_j. A estatística Γ normalizada é dada pela Equação 2, podendo também ser escrita na forma da Equação 53.

$$\Gamma = (1/M) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S(i,j) \cdot Q(i,j) \tag{1}$$

$$\hat{\Gamma} = \frac{[(1/M) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (S(i,j) - \mu_S)(Q(i,j) - \mu_Q)]}{\sigma_S \sigma_Q} \tag{2}$$

Esse índice decresce monotonicamente com o aumento do número de clusters. O número de clusters mais apropriado é aquele no qual pode-se observar uma inflexão significativa no gráfico de Γ normalizada contra K, que corresponde a um aumento significativo

da Γ normalizada. O valor desse índice é 1 para o agrupamento trivial, em que cada padrão pertence a um cluster individual, e não está definido para $K = 1$. O gráfico tende a ser plano quando os dados contêm clusters, mas tende a ter uma inclinação positiva para dados aleatórios.

Um valor alto de Γ ($\hat{\Gamma}$) indica a existência de clusters compactos. O índice Γ pode ser aplicável apenas quando os dados se agrupam em clusters hiperesféricos (Jain & Dubes 1988), não sendo apropriado para clusters de formas arbitrárias.

Γ e $\hat{\Gamma}$ podem ser usadas para validação de agrupamentos testando se a associação entre S e Q é extraordinariamente grande sob a hipótese de rótulos aleatórios (todas as permutações de rótulos de linhas (e colunas) de Q são igualmente prováveis) (Bezdek & Pal 1998).

3.1.2 *Família de índices Dunn*

A família de índices Dunn é representada pela Equação 3, em que $\delta(C_i, C_j)$ é uma função de dissimilaridade entre os clusters C_i e C_j e $\Delta(C_i)$ é a distância intra-cluster do cluster C_i, uma medida de dispersão do cluster. No índice Dunn original, $\delta(C_i, C_j)$ é dada pela Equação 4 e $\Delta(C_i)$ é dada pela Equação 5.

O índice Dunn é a razão da separação dentro dos clusters e entre os clusters (Pakhira et al. 2004).

$$D(K) = \min_{i=1,\ldots,K} \left\{ \min_{j=i+1,\ldots,K} \left\{ \frac{\delta(C_i, C_j)}{\max_{k=1,\ldots,K} diam(C_k)} \right\} \right\} \tag{3}$$

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{4}$$

$$\Delta(C_i) = \max_{x,y \in C_i} d(x, y) \tag{5}$$

O índice Dunn é apropriado para a identificação de clusters compactos e bem separados (Halkidi et al. 2002b). Valores altos do índice indicam a presença desse tipo de cluster. O índice $D(K)$ não apresenta nenhuma tendência em relação ao número de clusters. O ponto de máximo no gráfico de $D(K)$ contra K pode ser uma indicação do número de clusters que mais se ajusta aos dados.

Os problemas do índice Dunn original são a sua complexidade e sensibilidade a ruído. Outros índices da família Dunn são mais robustos à presença de ruídos. Bezdek & Pal (1998) propõem várias generalizações desse índice e as compara com os índices Γ modificado, DB e Dunn original. Azuaje (2002) apresenta uma ferramenta que emprega 18 índices baseados no índice Dunn para determinar o melhor número de clusters. Esses índices são baseados em diferentes combinações de diferentes distâncias inter-clusters e intra-cluster. Esse índice não é apropriado para clusters de formas arbitrárias.

3.1.3 Índice Davies-Bouldin (DB)

Seja $R_{i,j}$ uma medida de similaridade entre dois clusters C_i e C_j, dada pela Equação 6, em que e_{C_j} é o erro médio para o cluster C_j e $d(C_j, C_k)$ é a distância Euclideana entre os centros dos clusters C_j e C_k. Seja o índice para o k-ésimo cluster dado pela Equação 7. O índice de Davies-Bouldin, DB, é dado pela Equação 8.

Este índice é uma função da razão da soma da dispersão dentro dos clusters pela dispersão entre os clusters (Pakhira et al. 2004).

$$R_{j,k} = \frac{e_{C_j} + e_{C_k}}{d(C_j, C_k)} \tag{6}$$

$$R_k = \max_{j \neq k} \{R_{j,k}\} \tag{7}$$

$$DB(K) = \frac{1}{K} \sum_{k=1}^{K} R_k \tag{8}$$

O melhor agrupamento é dado pelo pelo menor valor do índice DB. Para se determinar o número de clusters mais apropriado, deve-se observar o valor de K que minimiza o índice no gráfico de $DB(K)$ contra K. O valor desse índice é 0 para o agrupamento trivial, em que cada padrão pertence a um cluster individual e deve ser computado apenas quando cada cluster contém um número razoável de padrões.

O índice DB pode ser aplicável apenas quando os dados se agrupam em clusters hiperesféricos (Jain & Dubes 1988), não sendo apropriado para clusters de formas arbitrárias.

3.1.4 Silhuetas

A medida de silhueta (Rousseeuw 1987) é calculada para cada padrão que faz parte de um agrupamento. As silhuetas medem a qualidade dos clusters com base na proximidade entre os padrões de um cluster e na distância dos padrões de um cluster ao cluster mais próximo. As silhuetas (Rousseeuw 1987) mostram quais padrões estão bem situados dentro dos seus clusters e quais estão fora de um cluster apropriado. Elas podem ser calculadas para agrupamentos realizados utilizando tanto medidas de similaridade quanto medidas de dissimilaridade (distâncias). Dados $a(x_i)$, a dissimilaridade média do padrão x_i em relação a todos os outros padrões do cluster C_i, $d(x_i, C_j)$, a dissimilaridade média do padrão x_i em relação aos padrões do cluster C_j e $b(x_i)$, a menor dissimilaridade média de x_i em relação a todos os demais clusters, dada pela Equação 9, a silhueta de um padrão, $s(x_i)$, empregando dissimilaridade, é dada pela Equação 10:

$$b(x_i) = \min_{C_i \neq C_j} d(x_i, C_j) \tag{9}$$

$$s(x_i) = \begin{cases} 1 - a(x_i)/b(x_i), & a(x_i) < b(x_i) \\ 0, & a(x_i) = b(x_i) \\ b(x_i)/a(x_i) - 1, & a(x_i) > b(x_i) \end{cases} \quad (10)$$

Para o cálculo da silhueta utilizando similaridade, deve-se empregar $a'(x_i)$ e $d'(x_i, C_j)$, as respectivas similaridades médias, $b'(x_i)$, dada pela Equação 11, e $s(x_i)$, dada pela Equação 12:

$$b'(x_i) = \min_{C_i \neq C_j} d'(x_i, C_j) \quad (11)$$

$$s(x_i) = \begin{cases} 1 - b'(x_i)/a'(x_i), & a'(x_i) > b'(x_i) \\ 0, & a'(x_i) = b'(x_i) \\ a'(x_i)/b'(x_i) - 1, & a'(x_i) < b'(x_i) \end{cases} \quad (12)$$

O valor de uma silhueta está no intervalo $[-1, 1]$. Se um padrão está bem situado dentro do seu cluster, sua silhueta apresenta um valor próximo de 1. Um valor próximo de -1 indica que o padrão deveria ser associado a outro cluster (Yeung 2001).

As silhuetas dependem apenas da partição dos dados, não dependendo do algoritmo de agrupamento empregado. Elas podem ser usadas para melhorar os resultados de uma análise de cluster ou para comparar os resultados de diferentes algoritmos aplicados aos mesmos dados.

Além da silhueta de cada objeto, pode ser calculada a silhueta de cada cluster e a largura média da silhueta, $\overline{s}(K)$, que é o valor médio sobre todos os padrões do conjunto de dados. Um modo de escolher o melhor valor de K é selecionar aquele que resulta no maior valor de $\overline{s}(K)$.

O coeficiente silhueta, SC, é o máximo $\overline{s}(K)$ para $K = 2, 3, ...(n-1)$. SC é uma medida da quantidade de estrutura descoberta por um algoritmo de agrupamento. Uma interpretação para SC é: $SC \leq 0.25$ significa que não foi encontrada uma estrutura substancial, $0.26 \leq SC \leq 0.5$ indica que a estrutura encontrada é fraca e pode ser artificial, $0.5 \leq SC \leq 0.7$ significa que uma estrutura razoável foi encontrada e $0.71 \leq SC \leq 1$ indica que foi encontrada uma estrutura forte.

As silhuetas são apropriada nos casos em que a proximidade está em uma escala de proporção, como a distância Euclideana, e para a identificação de clusters compactos e bem separados. Essa medida funciona melhor com clusters aproximadamente esféricos (Rousseeuw 1987). Uma vez que a definição da silhueta favorece padrões que estão associados a clusters com a similaridade média mais alta, as larguras das silhueta são tendenciosas contra clusters verdadeiros potencialmente sobrepostos, favorecendo agrupamentos disjuntos.

3.1.5 *Índice de Calinski-Harabasz (CH)*

Segundo (Tibshirani et al. 2001b), o índice CH, proposto por (Calinski & Harabasz 1974), é o índice com melhor desempenho no estudo de (Milligan & Cooper 1985) envol-

vendo 30 procedimentos de validação para determinar o melhor número de clusters. O índice CH é ilustrado pela Equação 13, em que $B(K)$ e $W(K)$ são as somas dos quadrados das distâncias ao centróide, respectivamente entre os clusters e dentro dos clusters, com K clusters (Tibshirani et al. 2001b).

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)} \tag{13}$$

O índice não está definido para $K = 1$. O melhor número de clusters é aquele que maximiza o índice CH.

3.1.6 Índice de Krzanowski e Lai (KL)

O índice KL, proposto por (Krzanowski & Lai 1985), é dado pela Equação 14:

$$DIFF(K) = (K-1)^{2/p} W_{k-1} - K^{2/p} W_K \tag{14}$$

$$KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right| \tag{15}$$

Este índice não está definido para $K > 1$. Segundo este índice, o melhor número de clusters é aquele que maximiza $KL(K)$ dado pela Equação 15.

3.1.7 Índice de Hartigan (H)

O índice de Hartigan, H (Hartigan 1975), é dado pela Equação 16:

$$H(K) = \left\{ \frac{W(K)}{W(K+1)} - 1 \right\} \Big/ (n-K-1) \tag{16}$$

Segundo (Tibshirani et al. 2001b), Hartigan sugere que um cluster seja adicinado se $H(K) > 10$. Portanto, o número estimado de clusters é o menor $K \geq 1$ tal que $H(K) \leq 10$. Esse índice é definido para $K = 1$ e pode potencialmente discriminar entre um ou mais clusters.

3.1.8 Índice SD

O índice SD é baseado nos conceitos de dispersão média por clusters, $Scatt(K)$, dada pela Equação 17, e separação total entre os clusters, $Dis(K)$, dada pela Equação 18. O índice é dado pela Equação 19, em que a é um fator de ponderação igual a $Dis(C_{max})$, com C_{max} sendo o número máximo de clusters de entrada.

$$Scatt(K) = \frac{1}{K} \sum_{i=1}^{K} \|\sigma(v_i)\| \,/\, \|\sigma(X)\| \tag{17}$$

$$Dis(K) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{K} \left(\sum_{z=1}^{K} \|v_k - v_z\| \right)^{-1} \tag{18}$$

$$SD(K) = a \cdot Scatt(K) + Diss(K) \tag{19}$$

O menor valor do índice $SD(K)$ indica o melhor número de clusters. Esse índice não é apropriado para clusters de formas arbitrárias.

3.1.9 Índice *S_Dbw*

A definição do índice *S_Dbw* (Halkidi & Vazirgiannis 2001) leva em consideração a compactação dos clusters (variância dos clusters), $Scatt(K)$, da Equação 17, e a densidade entre os clusters. A densidade entre clusters avalia a densidade média na região entre os clusters em relação à densidade dos clusters. Essa densidade é dada pela Equação 20, em que v_i e v_j são os centros dos clusters C_i e C_j, respectivamente, e z_{ij} é o ponto médio do segmento de reta definido por v_i e v_j. A densidade do ponto z, $density(z)$, representa o número de padrões na vizinhança de z e é dada pela Equação 21, em que n_{ij} é o número de padrões que pertencem aos clusters C_i e C_j, $x_l \in C_i, C_j$ e $f(x, z)$ é dada pela Equação 22, sendo *stdev* o desvio padrão médio dos clusters, dado pela Equação 23. O índice *S_Dbw* é dado pela Equação 24.

$$Dens\_bw = \frac{1}{K \cdot (K-1)} \sum_{i=1}^{K} \left(\sum_{\substack{j=1 \\ i \neq j}}^{K} \frac{density(z_{ij})}{\max\{density(v_i), density(v_j)\}} \right) \tag{20}$$

$$density(z) = \sum_{l=1}^{K} f(x_l, z) \tag{21}$$

$$f(x, z) = \begin{cases} 0, d(x, z) > stdev \\ 1, otherwise \end{cases} \tag{22}$$

$$stdev = \frac{1}{K} \sqrt{\sum_{i=1}^{K} \|\sigma(v_i)\|} \tag{23}$$

$$S\_Dbw(K) = Scatt(K) + Dens\_bw(K) \tag{24}$$

Esse índice avalia uma combinação da compactação, da separação e da variação de densidade entre os clusters. O número de clusters mais apropriado é aquele que minimiza o índice *S_Dbw*. Esse índice não é apropriado para clusters de formas arbitrárias.

3.1.10 Índice PBM

O índice PBM (Pakhira et al. 2004) é dado pela Equação 25, em que u_{ij} é a pertinência do padrão x_j ao cluster C_i (no caso *crisp*, $u_{ij} = 1$, se $x_j \in C_k$ e $u_{ij} = 0$, caso contrário), E_K é dado pela Equação 26, com E_k dado pela Equação 27 e D_K dado pela Equação 28.

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^2 \tag{25}$$

$$E_K = \sum_{k=1}^{K} E_k \tag{26}$$

$$E_k = \sum_{j=1}^{n} u_{kj} \left\| x_j - v_k \right\| \tag{27}$$

$$D_K = \max_{i,j=1}^{K} \left\| v_i - v_j \right\| \tag{28}$$

O objetivo do índice é manter o menor número possível de clusters aumentando sua compactação e separação.

O valor máximo do índice PBM indica o particionamento apropriado. Traçando-se o gráfico de PBM contra K, o valor máximo no gráfico indica o número correto de clusters.

3.1.11 Coeficiente de partição (PC)

Segundo Pal & Bezdek (1995), o índice coeficiente de partição (*Partition Coefficient*), PC, é dado pela Equação 29, em que u_{ij} é o grau de pertinência do padrão x_i ao cluster C_j.

$$PC = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} u_{ij}^2 \tag{29}$$

O valor do índice PC está no intervalo $[1/K, 1]$. Quanto mais próximo de 1, mais *crisp* é a partição. Esse índice tem seu máximo a cada partição *hard*. Por outro lado, quanto mais próximo de $1/K$ é o valor do índice, mais *fuzzy* é a partição. Além disso, um valor próximo de $1/K$ indica que não há tendência de agrupamento no conjunto de dados, ou o algoritmo aplicado não conseguiu revelar a estrutura. Não se pode afirmar que o agrupamento encontrado é bom se o valor de PC é próximo de 1.

3.1.12 Entropia da partição (PE)

Segundo Pal & Bezdek (1995), o índice coeficiente de entropia da partição (*Partition Entropy*), PE, é dado pela Equação 30, em que u_{ij} é o grau de pertinência do padrão x_i ao cluster C_j e a é a base do logaritmo.

$$PE = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} u_{ij} \cdot \log_a(u_{ij}) \tag{30}$$

O índice é computado para $K > 1$ e tem valores no intervalo $[0, \log_a K]$. Quanto mais próximo de 0, mais *crisp* é a partição. O índice tem seu mínimo a cada partição *hard*. Um valor próximo de $\log_a K$ indica que não há estrutura no conjunto de dados, ou que o algoritmo aplicado não conseguiu revelar essa estrutura. Não se pode afirmar que o agrupamento encontrado é bom se o valor de PE é próximo de 0.

3.1.13 Índice de Xie-Beni (XB)

O índice de Xie-Beni (Xie & Beni 1991), XB, mede a compactação e separação médias de uma partição *fuzzy* e é dado pela Equação 31, em que u_{ij} é o grau de pertinência do padrão x_i ao cluster C_j e v_i é o centro do cluster C_i

$$XB = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n} u_{ij}^2 \left\| v_i - x_j \right\|^2}{n(\min_{i,j} \left\| v_i - v_j \right\|^2)} \tag{31}$$

O melhor agrupamento é aquele que tem o menor valor do índice XB. O índice decresce monotonicamente quando K está próximo de n. Esse índice é apropriado apenas para clusters compactos e bem separados.

3.1.14 Índice Xie-Beni estendido

Esta extensão do índice Xie-Beni (Xie & Beni 1991) é recomendada quando o critério que está sendo otimizado é diferente de J_2 do algoritmo FCM.

$$XB_{estendido} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n} u_{ij}^m \left\| v_i - x_j \right\|^2}{n(\min_{i,j} \left\| v_i - v_j \right\|^2)} \tag{32}$$

3.1.15 Índice de Fukuyama-Sugeno (FS)

O índice proposto por Fukuyama & Sugeno (1989), FS, é dado pela Equação 33, em que $1 < m < \infty$.

$$FS = \sum_{k=1}^{n} \sum_{i=1}^{K} (u_{ik})^m \cdot (\left\| x_k - v_i \right\|_A^2 - \left\| v_i - \bar{v} \right\|_A^2) \tag{33}$$

O valor mínimo do índice FS indica um bom agrupamento.

3.1.16 Separação da partição (S)

O índice separação da partição (*Partition Separation*), S, proposto por Yang & Wu (2001), é dado pela Equação 34, em que S_k é dado pela Equação 35, com u_M dado pela Equação 36, β_T dado pela Equação 37.

$$S(K) = \sum_{k=1}^{K} S_k \tag{34}$$

$$S_k = \sum_{i=1}^{n} u_{ik}^2 / u_M - \exp(-\min_{j \neq k}\{\|v_k - v_j\|^2\}/\beta_T) \tag{35}$$

$$u_M = \max_{1 \leq k \leq K} \sum_{i=1}^{n} u_{ik}^2 \tag{36}$$

$$\beta_T = \frac{\sum_{k=1}^{K} \|v_k - \overline{v}\|^2}{K} \tag{37}$$

O melhor valor de K é aquele que maximiza $S(K)$, para $2 \leq K \leq n$.

3.1.17 $PBMF$

A versão *fuzzy* do índice PBM, $PBMF$ (Pakhira et al. 2004) é obtida incorporando distâncias *fuzzy* e é dada pela Equação 38, em que J_m é dado pela Equação 39 e os demais componentes são os mesmos definidos para o índice PBM.

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{J_m} \times D_K\right)^2 \tag{38}$$

$$J_m = \sum_{j=1}^{n} \sum_{k=1}^{K} (u_{kj})^m \|x_j - v_k\| \tag{39}$$

O valor máximo do índice $PBMF$ indica o particionamento apropriado. Traçando-se o gráfico de $PBMF$ contra K, o valor máximo no gráfico indica o número correto de clusters.

3.2 Outras metodologias de validação relativa

Nesta seção são descritas algumas outras abordagens para validação relativa. As metodologias detalhadas são a análise de replicação, proposta por McIntyre & Blashfield (1980) e Morey et al. (1983), a abordagem de Law & Jain (2003), que calcula a variabilidade do algoritmo utilizando *bootstrapping*, e as abordagens de Yeung et al. (2001) e Tibshirani et al. (2001a), que se baseiam no poder preditivo do algoritmo.

A análise de replicação (McIntyre & Blashfield 1980; Morey et al. 1983) se baseia num argumento semelhante ao procedimento de *cross-validation* em análise de regressão. A idéia é medir a estabilidade de um algoritmo de agrupamento. Essa estabilidade é medida comparando a partição obtida pelo algoritmo, com uma partição obtida em um subconjunto independente dos dados.

A forma de validação de um agrupamento proposta por Law & Jain (2003) se baseia na interpretação de um algoritmo de agrupamento como um estimador estatístico e avaliação da variabilidade desse estimador por meio de *bootstrapping*. Se uma partição é válida, sua variabilidade deve ser baixa.

As abordagens de Yeung et al. (2001) e Tibshirani et al. (2001a) se baseiam na idéia de que, se um agrupamento reflete uma estrutura verdadeira, um preditor construído com base nos clusters desse agrupamento deve estimar precisamente os rótulos dos clusters de novas amostras de teste (Jiang et al. 2004).

Cada uma dessas abordagens é detalhada a seguir.

3.2.1 Análise de replicação

A análise de replicação, proposta por McIntyre & Blashfield (1980), Morey et al. (1983) e descrita por Milligan (1996) é uma abordagem similar ao processo de *cross-validation* para validar um agrupamento. Seu objetivo é medir a estabilidade (ou replicabilidade) de um método de agrupamento.

A análise de replicação é ilustrada na Figura 5 e pode ser resumida da seguinte maneira:

- Obter duas amostras dos dados, o que pode ser feito dividindo o conjunto de dados original em dois subconjuntos.
- Realizar todos os passos de agrupamento na primeira amostra e determinar os centróides dos clusters obtidos.
- Determinar as distâncias entre os padrões da segunda amostra e os centróides encontrados.
- Associar os padrões da segunda amostra ao centróide mais próximo. Isso resulta em um agrupamento da segunda amostra com base nas características da primeira.
- Aplicar os mesmos passos de agrupamento na segunda amostra.
- Calcular uma medida de concordância entre os dois agrupamentos da segunda amostra. Essa medida pode ser o índice Rand corrigido (Hubert & Arabie 1985). O nível de concordância entre as duas partições reflete a estabilidade do algoritmo de agrupamento.

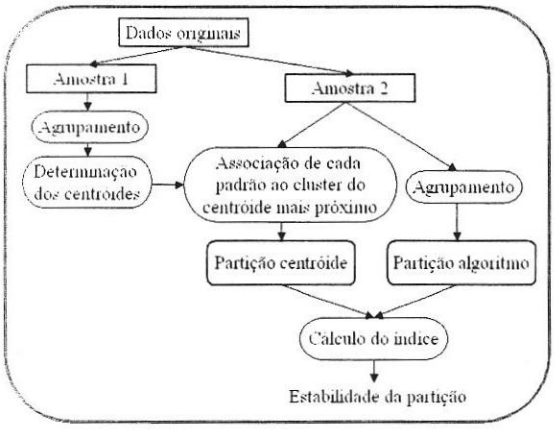

Figura 5: Análise de replicação para validação de um agrupamento.

3.2.2 *FOM*

Yeung et al. (2001) propõem uma metodologia para avaliação da qualidade do resultado de um agrupamento com base no poder preditivo do algoritmo. O poder preditivo é dado pela medida figura de mérito (*Figure of merit*), FOM, definida pelos autores. A FOM foi proposta para aplicação em agrupamentos de genes com base no seu nível de expressão medido em diversos experimentos. A intuição por trás dessa medida é a tendência dos genes pertencentes a um mesmo cluster terem um nível de expressão similar em experimentos adicionais, não empregados na geração dos clusters, se o agrupamento for significativo do ponto de vista biológico.

Um algoritmo de agrupamento é aplicado a todos os atributos, exceto um atributo, denominado e. Esse atributo e é utilizado para estimar o poder preditivo do algoritmo, por meio da variância dentro dos clusters dada pela FOM. Clusters significativos devem ter menos variação nos experimentos restantes do que clusters gerados ao acaso. Assim, quanto maior a similaridade intra-cluster calculada a partir da exclusão do atributo j, mais forte é o poder preditivo e melhor o esquema de agrupamento. Seja X_{ij} o valor do atributo j para o padrão x_i (ou, no contexto de expressão gênica, o nível de expressão do gene x_i sob a condição j) na matriz de dados bruta. Seja $\mu_{C_i}(j)$ a média dos valores do atributo j nos padrões pertencentes ao cluster C_i (ou o nível de expressão médio na condição j dos genes no cluster C_i). $FOM(j, K)$ é dada pela Equação 40, para K clusters usando o atributo j.

$$FOM(j, K) = \sqrt{\frac{1}{n} \sum_{i=1}^{K} \sum_{x \in C_i} (X_{ij} - \mu_{C_i}(j))^2} \tag{40}$$

Fazendo isso para cada um dos d atributos obtém-se a figura de mérito agregada, dada pela Equação 41, que é uma estimativa do poder preditivo total de um algoritmo sobre todas as amostras para K clusters.

$$FOM(K) = \sum_{j=1}^{d} FOM(j, K) \tag{41}$$

O valor da FOM tem a tendência de diminuir com o aumento do número de clusters. Para corrigir esse efeito, Yeung et al. (2001) utilizam a FOM ajustada, dada pela Equação 42.

$$FOM(K) = \frac{\sum_{j=1}^{d} FOM(j, K)}{\sqrt{(n-K)/n}} \tag{42}$$

Um valor baixo para a FOM indica um algoritmo com um poder preditivo maior.

A metodologia proposta por Yeung et al. (2001) tem uma abordagem preditiva, ou seja, o modelo proposto assume que o atributo excluído contém informação dos experimentos que são usados para produzir os clusters. Essa abordagem não é aplicável a todas as situações. Se todos os atributos contêm informações independentes, ela não se aplica. Além disso, não é uma abordagem segura para se comparar agrupamentos com diferentes números de clusters ou obtidos com medidas de similaridade diferentes.

3.2.3 Força de predição (PS)

A força de predição (Tibshirani et al. 2001a), PS é uma quantidade para avaliar o número de clusters em um conjunto de dados. Essa medida avalia quantos clusters podem ser preditos dos dados e quão bem eles podem ser preditos.

Para o cálculo da força de predição, Tibshirani et al. (2001a) dividem as amostras em um conjunto de treinamento X_{tr} e um conjunto de teste X_{te}. A idéia principal é usar o resultado do agrupamento gerado com a aplicação de um algoritmo aos dados de treinamento para predizer a co-pertinência (se duas amostras pertencem ao mesmo cluster) no conjunto de teste.

Inicialmente, agrupa-se os dados de teste em K clusters C_1, C_2, ... C_K. A operação de agrupamento é denotada por $C(X_{te}, K)$ e a co-pertinência por $D[C(X_{te}, K), X_{te}]_{ij} = 1$ se os padrões x_i e x_j estão no mesmo cluster e zero, caso contrário. Em seguida, agrupa-se os dados de treinamento em K clusters e mede-se o quão bem os centros dos clusters do conjunto de treinamento predizem as co-pertinências no conjunto de teste.

A força de predição (*prediction strength*), PS, de um agrupamento $C(., K)$ é dada pela Equação 43, em que $D[C(X_{tr}, K), X_{te}]_{ij}$ é a co-pertinência dos padrões de teste x_i e x_j agrupados de acordo com os centros do agrupamento dos padrões de treinamento, n_k é o número de padrões do cluster C_k e $I(c) = 1$, se a condição c for verdadeira e zero, caso contrário. Para cada cluster de teste, é computada a proporção de pares observados nesse cluster que também foram associados ao mesmo cluster pelos centros dos clusters derivados do conjunto de treinamento. A força de predição é o mínimo dessa quantidade sobre os K clusters de teste.

$$PS(K) = \min_{1 \leq k \leq K} \frac{1}{n_k(n_k - 1)} \sum_{i \neq j \in C_k} I(D[C(X_{tr}, K), X_{te}]_{ij} = 1) \quad (43)$$

Como em geral não se tem um conjunto de teste, Tibshirani et al. (2001a) utilizam *2-fold cross-validation* para dividir o conjunto de dados X em X_{tr} e X_{te}. Neste caso, a força de predição é a média sobre todas as divisões de X.

Quanto maior o valor de PS, melhor a qualidade do agrupamento. O número ótimo de clusters é o maior K tal que $PS(K)$ esteja acima de algum *threshold*. Tibshirani et al. (2001a) sugerem que um *threshold* entre 0,8 e 0,9 funciona bem para clusters bem separados.

3.2.4 Variabilidade

Law & Jain (2003) propõem uma forma de validação para comparar os resultados de diferentes agrupamentos, interpretando um algoritmo de agrupamento como um estimador estatístico e utilizando *bootstrapping* para estimar sua variabilidade. Se uma partição é válida, sua variabilidade deve ser baixa.

O algoritmo de agrupamento é considerado um estimador de ponto θ baseado em uma amostra X de tamanho n, independente e identicamente distribuída. A medida de variabilidade proposta é calculada da seguinte maneira:

- Gerar B amostras *bootstrap*, X_b^*, $b = 1, ..., B$, de tamanho n, reamostrando X com reposição.
- Executar o algoritmo de agrupamento para agrupar cada um dos conjuntos de dados X_b^*, gerando as partições P_{eb}^*.
- Calcular a variabilidade V de acordo com a Equação 44, em que $d(.,.)$ é uma medida de distância entre partições. Law & Jain (2003) utilizam cinco medidas diferentes: os índices Rand, Jaccard, Folkes e Malows e Hubert, descritos anteriormente, e a medida de dissimilaridade proposta em Lange et al. (2003). As quatro primeiras medidas são subtraídas de 1 para obter distâncias, uma vez que são medidas de similaridade entre duas partições.

$$V = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{j=i+1}^{B} d(P_{ei}^*, P_{ej}^*) \quad (44)$$

- Calcular a variabilidade ajustada $V' = V/V_{ran}$, em que V_{ran} é a variabilidade estimada utilizando B partições aleatórias de tamanho n.

Todo algoritmo de agrupamento apresenta um *bias*. Em geral, existe uma explicação para o *bias* de qualquer algoritmo de agrupamento razoável, e uma partição com baixa variabilidade, mesmo se diferente dos clusters reais, provavelmente revela informações úteis sobre os dados (Law & Jain 2003).

Este método não identifica explicitamente se um conjunto de dados não tem estrutura, mas uma alta variabilidade para diferentes valores de K pode sugerir ausência de estrutura.

4 Índices e metodologias empregados em critérios internos

Os índices empregados em critérios internos, ou simplesmente índices internos, medem o grau em que uma partição obtida por um algoritmo de agrupamento é justificada, baseado apenas na matriz de padrões ou na matriz de similaridade. Em outras palavras, medem o ajuste entre a partição gerada e os dados empregados. Esse tipo de validação está relacionado à determinação do número "verdadeiro" de clusters.

A aplicação de índices internos para a validação de uma partição apresenta diversas dificuldades (Jain & Dubes 1988; Halkidi et al. 2001). Uma delas decorre do fato de que os índices mais utilizados, como erro quadrático, provavelmente apresentam valores menores para dados que realmente possuem clusters do que com dados aleatórios, sem estrutura. Isso pode levar a conclusão errônea de que uma solução encontrada é válida, mesmo que ela não contenha o número correto de clusters. Uma solução para essa questão é a aplicação dos índices internos como índices relativos. Uma outra dificuldade é a dependência desses índices dos valores utilizados para os parâmetros do problema, tais como número de padrões, número de dimensões, número de clusters e o espalhamento dos dados (Jain & Dubes 1988).

Além dos índices relativos descritos anteriormente, que também podem ser empregados como internos, duas abordagens são tratadas a seguir: o índice *Gap* (Tibshirani et al. 2001b) e o procedimento *Clest* (Dudoit & Fridlyand 2002).

4.1 Estatística *Gap*

O índice *Gap* (Tibshirani et al. 2001b) é dado pela Equação 45. O valor de W_K é obtido pela Equação 46 e E_n^* é a esperança sob uma amostra de tamanho n de uma distribuição de referência. D_r da Equação 46 é, por sua vez, definido pela Equação 47.

A idéia dessa abordagem é padronizar o gráfico de $log(W_K)$ comparando-o com sua esperança sob uma distribuição de referência dos dados apropriada. A estimativa do número ótimo de clusters é o valor de K para o qual $log(W_K)$ se situa o mais abaixo possível dessa curva de referência.

$$Gap_n(K) = E_n^*\{\log(W_K)\} - \log(W_K) \tag{45}$$

$$W_K = \sum_{r=1}^{K} \frac{1}{2n_r} D_r \tag{46}$$

$$D_r = \sum_{i,j \in C_r} d(x_i, x_j) \tag{47}$$

Esse índice foi proposto para estimar o número de clusters presentes em um conjunto de dados. O valor estimado de K é o valor que maximiza $Gap_n(K)$ após levar em consideração a distribuição da amostra.

Essa estimativa é muito geral, aplicável a qualquer algoritmo de agrupamento e medida de distância. Para tornar a estatística *Gap* um procedimento operacional, é preciso encontrar uma distribuição referência apropriada e avaliar a distribuição amostral da estatística *Gap*.

A população de referência pode ser determinada de duas formas:

a. gerar cada atributo de referência uniformemente no intervalo dos valores observados para aquele atributo;

b. gerar os atributo de referência a partir de uma distribuição uniforme sobre uma caixa alinhada com os componentes principais dos dados.

O método `a` é mais simples, mas o método `b` leva em consideração a forma da distribuição dos dados e torna o procedimento invariante à rotação, se o método de agrupamento for invariante.

Para calcular a estatística *Gap*, estima-se $E_n^*\{\log(W_K)\}$ pela média de B cópias $log(W_k^*)$, cada uma computada de uma amostra de Monte Carlo extraída da distribuição referência. Em seguida, é preciso avaliar a distribuição amostral da estatística *Gap*.

O cálculo da estatística *Gap* deve ser feito da seguinte maneira. Agrupa-se os dados observados para cada número de clusters $K = 1, 2, ..., K_{max}$, calculando a medida de dispersão dentro dos clusters, W_K. Gera-se B conjuntos de referência, utilizando `a` ou `b` descritos anteriormente. Agrupa-se cada conjunto referência, calculando as medidas de dispersão $W_{Kb}^*, b = 1, 2, ..., B, K = 1, 2, ..., K_{max}$. Computa-se a estatística *Gap* estimada, dada pela Equação 48. Computa o desvio padrão sd_K, dado pela Equação 49, em que $\bar{l} = (1/B)\sum_b \log(W_{Kb}^*)$. Define-se $s_K = sd_K\sqrt{1 + 1/B}$. Escolhe-se o melhor número de clusters como sendo o menor K tal que $Gap(K) \geq Gap(K + 1) - s_{K+1}$.

$$Gap_n(K) = (1/B)\sum_b \log(W_{Kb}^*) - \log(W_K) \tag{48}$$

$$sd_K = [(1/B)\sum_b \{\log(W_{Kb}^*) - \bar{l}\}^2]^{1/2} \tag{49}$$

A derivação para a estatística *Gap* assume que há clusters uniformes bem separados. Quando existem subclusters menores dentro de clusters maiores bem separados, a estatística *Gap* pode exibir um comportamento não monotônico. Portanto, é importante examinar a curva *Gap* inteira, ao invés de simplesmente buscar o seu máximo.

4.2 *Clest*

Dudoit & Fridlyand (2002) propõem o método de reamostragem baseado em predição, *Clest*, para estimar o número de clusters de um agrupamento. Esse método é uma generalização e extensão da análise de replicação.

O número de clusters é estimado comparando o valor observado da estatística empregada (R, J ou FM) para cada K com o seu valor esperado sob um distribuição nula apropriada com $K = 1$. O número estimado de clusters é definido como sendo o valor de K correspondente à maior evidência significativa contra a hipótese nula de $K = 1$. Os conjuntos de dados referência são gerados sob a hipótese de uniformidade. O procedimento utilizado é o seguinte:

- Para todos os números possíveis de cluster k, $2 \leq k \leq K_{max}$:
  - Dividir aleatoriamente B vezes o conjunto original em conjuntos de treinamento $X_{b_{tr}}$ e teste $X_{b_{te}}$, sem sobrebreposição.
  - Executar o algoritmo de agrupamento para agrupar os conjuntos de treinamento $X_{b_{tr}}$, obtendo as partições $P_{b_{tr}}$.
  - Construir classificadores usando os conjuntos de treinamento $X_{b_{tr}}$ e os rótulos dos clusters das partições $P_{b_{tr}}$.
  - Aplicar os classificadores aos conjuntos de teste $X_{b_{te}}$, obtendo as classificações $C_{b_{te}}$.
  - Executar o algoritmo de agrupamento para agrupar os conjuntos de teste $X_{b_{te}}$, obtendo as partições $P_{b_{te}}$.
  - Calcular o valor do índice, ind_b (R, J ou FM), entre a partição $P_{b_{te}}$ e a classificação $C_{b_{te}}$, obtida com a aplicação do classificador, para todas as B subdivisões do conjunto de dados.

- Calcular a estatística de similaridade observada para a partição de k clusters dos dados, $t_k = mediana(ind_1, ..., ind_B)$.

- Gerar B_0 conjuntos de dados de referência sob uma hipótese nula apropriada. Para cada conjunto de referência repete os passos anteriores para obter B_0 estatísticas de similaridade, $t^*_{k1}, ..., t^*_{kB_0}$.

- Seja $t^*_k = [1/B_0] \sum_{b=1}^{B_0} t^*_{kb}$, e p_k a proporção de $t^*_{kb}, 1 \leq b \leq B_0$, que é pelo menos tão grande quanto a estatística observada t_k. Seja $d_k = t_k - t^*_k$ a diferença entre a estatística de similaridade observada e seu valor esperado estimado sob a hipótese nula de $K = 1$.

  Definir o conjunto $S = \{2 \leq k \leq K_{max} : p_k \leq p_{max}, d_k \geq d_{min}\}$, em que p_{max} e d_{min} são *thresholds* pré-estabelecidos. Se esse conjunto é vazio, então o número de clusters estimado $\hat{K}$ é 1. Caso contrário, $\hat{K} = argmax_{k \in S} d_k$.

5 Índices e metodologias empregados em critérios externos

O objetivo da validação externa é medir o quanto o agrupamento obtido confirma uma hipótese pré-especificada. Para isso, são utilizados testes de hipótese para determinar se uma estrutura obtida é apropriada para os dados. Isto é feito testando se o valor do índice utilizado é extraordinariamente grande ou pequeno.

Conforme dito anteriormente, para utilizar os índices em critérios externos, aplicando testes estatísticos, é preciso conhecer suas funções de densidade de probabilidade sob a hipótese nula H_0, de estrutura aleatória do conjunto de dados. O cálculo da função de densidade de probabilidade desses índices é difícil. Uma metodologia comumente empregada para isso é a aplicação de análise de Monte Carlo. O procedimento para a utilização de análise de Monte Carlo para determinar a distribuição referência desses índices sob a hipótese nula H_0 e responder se um agrupamento é válido de acordo com eles é descrito por Halkidi et al. (2002a) e Jain & Dubes (1988):

- Gerar B conjuntos de dados X_b^*, $b = 1, ..., B$, de tamanho n, com a mesma dimensão do conjunto de dados original X, seguindo uma distribuição uniforme (Para um nível de significância α de 0,05, Jain & Dubes (1988) indicam a utilização de $B = 100$).
- Para os B conjuntos de dados X_b^*, associar cada padrão do conjunto a um dos clusters existentes na partição real P_r. Os padrões são associados aleatoriamente aos clusters, de forma que cada cluster tenha o mesmo tamanho dos clusters na partição P_r. Isso gera, para cada conjunto de dados X_b^*, uma partição aleatória, P_{rb}^*.
- Executar o mesmo algoritmo de agrupamento usado para gerar P_e, para agrupar cada um dos conjuntos de dados X_b^*, gerando as partições P_{eb}^*.
- Calcular o índice escolhido, ind, entre a partição P_{eb}^* e P_{rb}^*, obtendo os valores $ind(P_{eb}^*)$ para cada conjunto X_b^*.
- Calcular o valor do índice entre a partição P_e e P_r, $ind(P_e)$.
- Criar um gráfico de dispersão para os B valores do índice, $ind(P_{eb}^*)$ (esse gráfico é uma aproximação da função de densidade de probabilidade do índice).
- Comparar o índice calculado para a partição P_e, $ind(P_e)$, com os valores do gráfico, considerando um dado nível de significância. Como, segundo Halkidi et al. (2002a), os índices R, J, FM e Γ têm uma função de densidade de probabilidade unicaudal à direita sob a hipótese nula, H_0, de estrutura aleatória do conjunto de dados, se s valores de $ind(P_{eb}^*)$ são maiores que o valor de $ind(P_e)$, com $s = (1 - alpha)B$, a hipótese nula é aceita, indicando que os dados são distribuídos aleatoriamente.

Em seguida, são descritos alguns índices mais tradicionalmente empregados em validação externa, bem como o índice proposto recentemente por Dom (2002). Além das formas tradicionais desses índices, existem equações corrigidas que são freqüentemente empregadas. Uma dessas correções, bastante utilizada, é a correção da estatística *Rand*,

proposta por Hubert & Arabie (1985), que também é apresentada neste trabalho. Os índices mais tradicionais são comparados na Seção 5.7.

Os índices externos mais tradicionais, freqüentemente utilizados na validação de partições, comparam uma partição resultante da aplicação de um algoritmo, P_e, com uma partição independente dos dados, construída com base na intuição ou conhecimento a priori sobre a estrutura real dos dados, P_r. São os casos da estatística *Rand* (R), descrita na Seção 5.1, do coeficiente de Jaccard (J), descrito na Seção 5.2, do índice de Folkes e Mallows (FM), descrito na Seção 5.3 e da estatística Huberts Γ normalizada, descrita na Seção 5.4. Esses índices se baseiam nas seguintes informações a respeito da relação entre os pontos de P_e e P_r. Um par de pontos, ou padrões, (x_v, x_u) é dito:

- SS: se pertencem ao mesmo cluster de P_e e ao mesmo cluster de P_r.
- SD: se pertencem ao mesmo cluster de P_e e a clusters diferentes de P_r.
- DS: se pertencem a clusters diferentes de P_e e ao mesmo cluster de P_r.
- DD: se pertencem a clusters diferentes de P_e e a clusters diferentes de P_r.

Sejam $a1$, $a2$, $a3$ e $a4$ os números de pares SS, SD, DS e DD, respectivamente. Define-se $M = a1 + a2 + a3 + a4$ como o número máximo de todos os pares no conjunto de dados ($M = n(n-1)/2$). Define-se também $m_1 = a1 + a2$ e $m_2 = a1 + a3$.

5.1 Índice Rand

O índice Rand computa a probabilidade de que dois pontos pertençam ao mesmo cluster ou pertençam a clusters diferentes nas duas partições P_e e P_r (Boutin & Hascoët 2004). A estatística *Rand* (R) é dada pela Equação 50.

$$R = \frac{(a1 + a4)}{M} \tag{50}$$

5.2 Índice Jaccard

Esse índice computa a probabilidade de que dois pontos pertencentes ao mesmo cluster em uma das partições também pertençam ao mesmo cluster na outra partição (Boutin & Hascoët 2004). O coeficiente de Jaccard (J) é dado pela Equação 51.

$$J = \frac{a1}{(a1 + a2 + a3)} \tag{51}$$

5.3 Índice Folkes e Mallows

O índice de Folkes e Mallows (FM) é dado pela Equação 52.

$$FM = \frac{a1}{\sqrt{(m_1)(m_2)}} \tag{52}$$

5.4 Índice Hubert

A estatística Huberts Γ normalizada é dada pela Equação 53.

$$\Gamma = \frac{Ma1 - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \tag{53}$$

5.5 Índice Rand corrigido

Além das formas tradicionais dos índices descritos anteriormente, existem equações corrigidas que são freqüentemente empregadas. Uma correção bastante comum é a normalização do índice para que ele apresente o valor 0 quando as partições são selecionadas ao acaso e 1 quando a partição casa perfeitamente com a partição real (Jain & Dubes 1988; Gordon 1999). A correção da estatística *Rand*, proposta por Hubert & Arabie (1985), resulta na estatística *Rand* corrigida (CR), dada pela Equação 54, em que n_{ij} é o número de objetos comuns aos clusters C_i de P_e e C_j de P_r, $n_{i.}$ é o número de objetos no cluster C_i de P_e, $n_{.j}$ é o número de objetos no cluster C_j de P_r, K_{P_e} e K_{P_r} são os números de clusters nas partições P_e e P_r respectivamente.

$$CR = \frac{\sum_{i=1}^{K_{P_e}} \sum_{j=1}^{K_{P_r}} \binom{n_{ij}}{2} - \left[\sum_{i=1}^{K_{P_e}} \binom{n_{i.}}{2} \sum_{j=1}^{K_{P_r}} \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\left[\sum_{i=1}^{K_{P_e}} \binom{n_{i.}}{2} + \sum_{j=1}^{K_{P_r}} \binom{n_{.j}}{2}\right] /2 - \left[\sum_{i=1}^{K_{P_e}} \binom{n_{i.}}{2} \sum_{j=1}^{K_{P_r}} \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{54}$$

O índice *Rand* corrigido é limitado superiormente pelo valor 1 e assume o valor 0 quando o índice atinge seu valor esperado (Hubert & Arabie 1985). Também pode ser apropriado ter um limite inferior bem definido, mas a normalização necessária não oferece vantagens práticas, uma vez que valores negativos do índice não tem uso real (Hubert & Arabie 1985).

Uma outra correção bastante utilizada é a correção sugerida por Morey & Agresti (1984) e utilizada por Milligan & Cooper (1986). Porém, segundo Hubert & Arabie (1985) essa correção assume inadequadamente que a esperança de uma variável aleatória ao quadrado é o quadrado da esperança.

5.6 Índice de Dom

Um outro índice externo proposto recentemente por Dom (2002) mede qualitativamente o quão úteis os rótulos de clusters dos padrões são como preditores de suas classes.

A medida proposta é baseada na entropia condicional da teoria da informação, podendo também ser definida em termos da informação mútua.

Dados n_{ij}, o número de objetos comuns aos clusters C_i de P_e e C_j de P_r, $n_{i.} = \sum_{j=1}^{K_{P_r}} n_{ij}$, o número de objetos no cluster C_i de P_e, $n_{.j} = \sum_{i=1}^{K_{P_e}} n_{ij}$, o número de objetos no cluster C_j de P_r e K_{P_e} e K_{P_r}. os números de clusters nas partições P_e e P_r, respectivamente. O índice proposto é dado pela Equação 55, em que $\tilde{H}(P_r|P_e)$, a entropia condicional empírica, é dada pela Equação 56.

$$Q_0(P_r|P_e) = \tilde{H}(P_r|P_e) + \frac{1}{n}\sum_{i=1}^{K_{P_e}} \log \begin{pmatrix} n_{i.} + K_{P_r} - 1 \\ K_{P_r} - 1 \end{pmatrix} \tag{55}$$

$$\tilde{H}(P_r|P_e) = \sum_{j=1}^{K_{P_r}} \sum_{i=1}^{K_{P_e}} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n_{i.}} \tag{56}$$

Os menores valores do índice indicam agrupamentos melhores. Para que o índice tenha valores no intervalo $(0, 1]$, com valor 0 representando o pior agrupamento e o valor 1 representando o melhor, como os índices R, J, FM e Γ, Dom (2002) transforma Q_0 para a forma dada pela Equação 57.

$$Q_2(P_r|P_e) = \frac{\frac{1}{n}\sum_{j=1}^{K_{P_r}} \log \begin{pmatrix} n_{.j} + K_{P_r} - 1 \\ K_{P_r} - 1 \end{pmatrix}}{Q_0(P_r|P_e)} \tag{57}$$

Dom (2002) faz uma análise comparando a medida proposta com outros índices de validação externos, mostrando que sua medida tem todas as propriedades desejáveis para uma medida de validação, argumentando, em bases filosóficas, que sua medida é superior por ser baseada em teoria da informação. Ele mostra que as outras medidas analisadas por ele dão resultados contra-intuitivos (se não incorretos) em certos casos, enquanto que a medida proposta tem sempre o comportamento desejado. Mostra ainda que a medida proposta apresenta resultados diferentes daqueles produzidos pelas outras medidas comumente utilizadas.

5.7 Comparação dos índices tradicionais para validação externa

Os índices R, J, FM, Γ e CR são os mais comuns para utilização em validação externa. A maioria deles pode ser sensível ao número de clusters de uma partição ou à distribuição dos padrões nos clusters (Filho 2003). Os índices Γ e R têm a tendência de apresentar valores mais elevados para partições com maior número de clusters. O índice J apresenta valores mais elevados para partições com menor número de clusters. O índice CR não apresenta essas características indesejáveis.

Valores altos de R, FM e Γ indicam forte concordância entre as duas partições P_e e P_r. A estatística Γ é uma correlação e seu valor está compreendido no intervalo $[-1, 1]$. R e J possuem valores no intervalo $[0, 1]$, sendo que o valor máximo 1 não é atingível quando as duas partições têm números de clusters diferentes.

A estatística Rand se aproxima de seu limite superior de 1 a medida em que o número de clusters cresce sem limite (Milligan et al. 1983). Os índices Rand, Jaccard e Folkes e Mallows têm como limite inferior o valor 0, mas na prática nunca vão produzir esse valor em um agrupamento de dados real. Milligan & Cooper (1986) confirmaram que o índice CR apresenta valores próximos de 0 quando os clusters são gerados a partir de dados aleatórios.

Nas investigações de Milligan et al. (1983) sobre alguns desses índices, foi encontrado um alto grau de consistência. Esses autores encontraram alta similaridade entre os índices J e FM, talvez devido ao fato de ambos utilizarem o mesmo numerador (número de pares de pontos que foram apropriadamente agrupados pelo algoritmo). Da mesma forma, os índices R e CR se comportaram de forma semelhante, o que é razoável, uma vez que a estatística Rand corrigida (CR) é apenas um ajuste da R original. Milligan et al. (1983) recomendam a utilização de apenas um índice de cada par de índices similares.

A estatística Rand corrigida tem uma variabilidade aumentada, o que parece ser desejável (Milligan et al. 1983). O fato desse índice poder assumir valores entre 0 e 1 para conjuntos de dados reais dá ao índice um forte apelo intuitivo para pesquisadores de áreas aplicadas. Da mesma maneira, o maior intervalo de variância do índice J em relação ao FM, o torna a melhor escolha do grupo. Estudos mais detalhados que envolvam outras características dos índices podem mudar essa recomendação.

6 Conclusão

A avaliação do resultado de um agrupamento deve ser objetiva, com o propósito de determinar se os clusters são significativos, ou seja, se a solução é representativa para o conjunto de dados analisado.

A validação do resultado de um agrupamento, em geral, é feita com base em índices estatísticos, que julgam, de uma maneira qualitativa, o mérito das estruturas encontradas. Um índice quantifica alguma informação a respeito da qualidade de um agrupamento. A maneira pela qual um índice é aplicado para validar um agrupamento é dada pelo critério de validação.

Este relatório apresentou uma descrição de vários critérios, métodos e índices comumente empregados para a validação de agrupamento, enfocando a validação de uma partição encontrada por um algoritmo de agrupamento para um dado conjunto de dados.

Referências

Azuaje, F. (2002). A cluster validity framework for genome expression data. *Bioinformatics 18*(2), 319–320.

Bezdek, J. C., & N. R. Pal (1998, Jun). Some new indexes of cluster validity. *IEEE Trans. Syst., Man, Cybernetics - Part B: Cybernetics 28*(3), 301–315.

Boutin, F., & M. Hascoët (2004). Cluster validity indices for graph partitioning. In *Eighth International Conference on Information Visualisation (IV'04)*, London, England, pp. 376–381.

Calinski, R., & J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics 3*, 1–27.

Dom, B. E. (2002). An information-theoretic external cluster-validity measure. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, San Francisco, CA, pp. 137–145. Morgan Kaufmann Publishers.

Dudoit, S., & J. Fridlyand (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology 3*(7), research0036.1–0036.21.

Filho, I. G. C. (2003, Apr). Comparative analysis of clustering methods for gene expression data. Master's thesis, Centro de Informática, Universidade Federal de Pernambuco, Recife.

Fukuyama, M., & M. Sugeno (1989). A new method of choosing the number of clusters for the fuzzy c-means method. In *Proceedings of 5th Fuzzy Systems Symposium*, pp. 247–250.

Gordon, A. (1999). *Classification*. Chapman & Hall/CRC.

Gordon, A. D. (1996). *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization*, Chapter Null models in cluster validation, pp. 32–44. Springer-Verlag.

Halkidi, M., Y. Batistakis, & M. Vazirgiannis (2001). On clustering validation techniques. *Intelligent Information Systems Journal 17*(2-3), 107–145.

Halkidi, M., Y. Batistakis, & M. Vazirgiannis (2002a, Jun). Cluster validity methods: Part I. *SIGMOD Record 31*(2), 40–45.

Halkidi, M., Y. Batistakis, & M. Vazirgiannis (2002b, Sep). Cluster validity methods: Part II. *SIGMOD Record 31*(3), 19–27.

Halkidi, M., & M. Vazirgiannis (2001). A data set oriented approach for clustering algorithm selection. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 165–179.

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

He, Q. (1999). A review of clustering algorithms as applied in IR. Technical Report UIUCLIS–1999/6+IRG, Information Retrieval Group, University of Illinois.

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

Jain, A., & R. Dubes (1988). *Algorithms for Clustering Data.* Prentice Hall.

Jiang, D., C. Tang, & A. Zhang (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering 16*(11), 1370–1386.

Kerr, M. K., & G. A. Churchill (2001, Jul). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. In *Proc. Natl. Acad. Sci. USA*, Volume 98, pp. 8961–8965.

Krzanowski, W., & Y. Lai (1985). A criterion for determining the number of groups in a dataset using sum of squares clustering. *Biometrics 44*, 23–34.

Lange, T., M. Braun, V. Roth, & J. Buhmann (2003). Stability-based model selection. In *Advances in Neural Information Processing Systems*, Volume 15, pp. 617–624.

Law, M. H., & A. K. Jain (2003). Cluster validity by bootstrapping partitions. Technical Report MSU-CSE-03-5, Department of Computer Science and Engineering, Michigan State University.

McIntyre, R. M., & R. K. Blashfield (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research 15*, 225–238.

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

Milligan, G., & M. Cooper (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research 21*, 441–458.

Milligan, G., T. Soon, & L. Sokol (1983, Jan). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence 5*(1), 40–47.

Milligan, G. W. (1996). *Clustering and Classification*, Chapter Clustering validation: results and implications for applied analyses. World Scientific Publ.

Morey, L., & A. Agresti (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement 44*, 33–37.

Morey, L. C., R. K. Blashfield, & H. A. Skinner (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research 18*, 309–329.

Pakhira, M. K., S. Bandyopadhyay, & U. Maulik (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition 37*(3), 487–501.

Pal, N. R., & J. C. Bezdek (1995, Aug). On cluster validity for fuzzy c-means model. *IEEE Transactions on Fuzzy Systems 3*(3), 370–379.

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.

Sarle, W. S., & A.-H. Kuo (1993). The MODECLUS procedure. Technical Report P-256, Cary, NC: SAS Institute Inc.

Tibshirani, R., G. Walther, D. Botstein, & P. Brown (2001). Cluster validation by prediction strength. Technical report, Department of Statistics, Stanford University.

Tibshirani, R., G. Walther, & T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

Xie, X. L., & G. Beni (1991, Aug). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*(8), 841–847.

Yang, M.-S., & K.-L. Wu (2001). A new validity index for fuzzy clustering. In *IEEE International Fuzzy Systems Conference*, pp. 89–92.

Yeung, K. (2001). *Cluster analysis of gene expression data.* Ph. D. thesis, University of Washington.

Yeung, K. Y., D. R. Haynor, & W. L. Ruzzo (2001). Validating clustering for gene expression data. *Bioinformatics 17*(4), 309–318.