0.1 Análise de agrupamento

A técnica de análise de agrupamento (AA), também conhecida como análise de conglomerados, classificação ou *cluster analysis* corresponde a um método que busca uma partição dos elementos de uma amostra em grupos de tal forma que (a) as observações de um mesmo grupo sejam similares entre si em relação às variáveis medidas e que (b) as observações de grupos diferentes sejam heterogêneas em relação a essas mesmas variáveis. Portanto, dada uma amostra de tamanho n, com cada objeto medido segundo p variáveis, a análise de agrupamento classifica os objetos em grupos com elevado grau de homogeneidade interna e heterogeneidade externa.

De acordo com a classificação de dados em grupos pode ser realizada com o objetivo de simplificá-los e realizar predições. A partir do método, é possível detectar o relacionamento e estrutura do conjunto de dados. Em muitas aplicações, os pesquisadores podem estar interessados na descrição de um conjunto de dados maior e a atribuição de novos objetos, bem como fazer previsão e descobrir hipóteses para explicar a estrutura dos dados.

"Análise de agrupamento" é uma expressão genérica que compreende vários métodos numéricos que pretendem descobrir grupos de observações homogêneas. O que esses métodos tentam é formalizar o que os seres humanos conseguem fazer bem em duas ou três dimensões, por exemplo, por meio de diagramas de dispersão. Os grupos são identificados pela avaliação das distâncias entre os pontos

Como forma de ilustração, considere um conjunto de dados fictícios em que há $n = 23$ observações (pessoas) e registros sobre suas idades e rendas mensais, ou seja, há $p = 2$ variáveis, ambas medidas na escala contínua. O objetivo é agrupar as pessoas em relação às duas variáveis. O diagrama de dispersão obtido se encontra no gráfico 5 e é possível identificar três agrupamentos, apenas pela análise visual. Se houvesse mais uma variável e, consequentemente mais uma dimensão, ainda seria possível imaginar um gráfico com os pontos. Porém, com mais de três dimensões, torna-se mais difícil a visualização dos dados [**?**].

Figura 1: Diagrama de dispersão de dados fictícios de idade e renda de várias pessoas.

Figura 2: *
Fonte: modificado a partir de (BARTHOLOMEW et al., 2008, p.18).

Há dois objetivos possíveis de um agrupamento: agrupar as n observações em um número de grupos desconhecidos ou classificar as observações em um conjunto predefinido de grupos. A análise de agrupamento deve ser utilizada no primeiro caso e análise discriminante no segundo. Na análise de agrupamento, geralmente, o número de grupos não é conhecido a princípio e encontrar o melhor agrupamento não é uma tarefa simples

De acordo com qualquer processo de agrupamento tem como base duas etapas:

1. Obter as distâncias de todos os pares de objetos para construção da matriz de proximidades;

2. Desenvolver um algoritmo para formação de grupos com base nessas distâncias.

As distâncias da etapa 1 são determinadas com base em medidas de similaridade ou dissimilaridade, que indicam a proximidade dos objetos. As medidas de dissimilaridade correspondem às distâncias, ao passo que as de similaridades complementam as distâncias, assim, quanto maior a medida de similaridade entre dois objetos menor será a de dissimilaridade e mais próximos eles serão A seguir são apresentadas algumas orientações sobre medidas de distâncias que podem ser utilizadas.

0.1.1 Distâncias

Para realizar o procedimento de agrupamento escolhido é necessário que a medida de similaridade ou dissimilaridade seja definida *a priori*. Alguns tipos comuns de distâncias que podem ser calculadas

entre os pares de observações são a distância euclidiana, distância euclidiana padronizada e distância de Mahalanobis, entre outras. Essas medidas são de dissimilaridade, ou seja, quanto menor seus valores, mais próximos ou similares são os objetos comparados. A escolha da métrica interfere diretamente no resultado final do agrupamento

A distância entre as observações i e j aparece na i-ésima linha e j-ésima coluna da matriz de distâncias. Por exemplo, se há $n = 4$ elementos na amostra, a matriz de distâncias terá dimensão 4×4 e será da forma

$$\begin{bmatrix} - & d_{12} & d_{13} & d_{14} \\ d_{21} & - & d_{23} & d_{24} \\ d_{31} & d_{32} & - & d_{34} \\ d_{41} & d_{42} & d_{43} & - \end{bmatrix},$$

em que d_{ij} é a distância entre os elementos i e j. Geralmente, essa matriz é simétrica, ou seja, $d_{12} = d_{21}$, $d_{13} = d_{31}$, e assim por diante

Dentre as distâncias citadas, um tipo muito simples e comum é a distância euclidiana, calculada por

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2},$$

em que d_{ij} é a distância euclidiana entre os elementos i, com os valores X_{i1}, X_{i2}, ..., X_{ip}, e j, com os valores X_{j1}, X_{j2}, ..., X_{jp}. Na aplicação da distância euclidiana há dois caminhos diferentes. Como as variáveis geralmente são medidas em unidades distintas, é necessário que os dados sejam padronizados. Dessa forma, a cada variável padronizada é atribuído o mesmo peso. No entanto, caso seja aplicada a técnica de componentes principais para a redução da dimensionalidade dos dados, o peso difere de acordo com o componente. Nessa situação, é atribuído ao primeiro componente um peso maior na determinação da similaridade entre os objetos De acordo com o uso dessa métrica faz com que variáveis com maior variabilidade dominem a classificação e ordenação dos objetos, portanto é mais indicada para grupos de variáveis com escalas similares.

As distâncias de Mahalanobis e euclidiana padronizada são uma generalização da distância euclidiana. Dessa forma, seja a distância generalizada entre dois elementos $X_{i.}$ e $X_{j.}$ definida por:

$$d_{ij} = (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}})^T \boldsymbol{A} (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}}).$$

A seleção dessa matriz $\boldsymbol{A}$ define a distância utilizada. Quando $\boldsymbol{A} = \boldsymbol{I}$ temos a distância euclidiana e quando $\boldsymbol{A} = \boldsymbol{D^{-1}}$ temos a distância euclidiana padronizada. Se $\boldsymbol{A} = \boldsymbol{S^{-1}}$, isto é, a matriz de covariâncias da matriz de dados, obtém-se a distância de Mahalanobis. Nesse caso, são consideradas as diferenças de variâncias e relações lineares entre as variáveis, a partir das covariâncias A definição dessa métrica propõe a ideia de que objetos situados na mesma direção das correlações entre as variáveis são mais similares entre si do que aqueles situados na direção oposta Além disso, a métrica produz agrupamentos compactos e convexos e elimina o efeito o efeito de domínio na classificação das variáveis de maior variabilidade

Há outras medidas de similaridade e dissimilaridade propostas na literatura, tais como as distâncias: euclidiana média, quarteirão (*city-block*) ou Manhattan, de Chebychev, angular, Canberra, entre outras. Embora a distância euclidiana seja uma das mais utilizadas, a de Mahalanobis é a mais indicada na maioria das situações aplicadas por levar em conta a colinearidade existente entre as variáveis usadas para realizar o agrupamento Além disso, outras distâncias conhecidas em situações práticas, como as de Kolmogorov, de Hellinger, de Rao, entre outras, são funções da distância de Mahalanobis sob pressuposições de normalidade e homocedasticidade, e sob outras condições

Após a definição da medida de distância utilizada, é necessário escolher um método de agrupamento. Segundo Ferreira (2011), esses métodos são divididos em hierárquicos (aglomerativos ou divisivos) e não hierárquicos. No primeiro tipo, há n grupos no início, cada um com uma observação, e no fim há um único grupo com todas as observações. A cada passo, cada observação ou grupo é unido a outra observação ou grupo (método hierárquico aglomerativo). A união ocorre com base no critério de similaridade, os objetos mais próximos entre si são alocados para um mesmo grupo, até que todos estejam em um único grupo. Portanto, a cada passo se perde um grupo, que é unido ao outro mais similar a ele. Nos métodos hierárquicos divisivos, há um único grupo com as n observações no início e, ao final, há n grupos. Nos métodos que não são hierárquicos é preciso definir o número k de grupos inicialmente para, em seguida, atribuir as n observações aos k grupos da melhor maneira possível. Sempre é preciso usar uma alocação arbitrária no início do processo e, iterativamente, buscar a alocação ótima.

0.1.2 Técnicas hierárquicas aglomerativas

De acordo com os agrupamentos resultantes a partir de métodos hierárquicos, aglomerativos ou divisivos, podem ser representados por um diagrama bidimensional muito utilizado, o dendrograma, também denominado de diagrama de árvore. Esse gráfico ilustra as fusões ou divisões realizadas a cada passo do processo. A Figura 3 mostra algumas das terminologias utilizadas para descrever os dendrogramas.

Ainda segundo os autores, o arranjo de nós e caules representam a topologia da árvore. O diagrama descreve o processo pelo qual foi obtida a hierarquia, assim há várias sub-árvores oriundas da raiz da árvore. O nó interno representa partições particulares, ou seja, os agrupamentos formados a partir dos nós terminais, que representam os objetos. A altura do nó interno corresponde ao ponto em que os objetos ou grupos foram unidos, ou seja, a proximidade entre eles. Dessa forma, a ordem de união dos grupos segue o princípio de ordem crescente da altura do nó.

Figura 3: Terminologia utilizada na descrição de dendrogramas

Figura 4: *
Fonte: elaboração própria a partir de [**?**].

Portanto, no caso representado pela Figura 3, os objetos denominados de A e C foram os primeiros a serem unidos em um único grupo, com nível de fusão de aproximadamente 1,7 (altura do nó). Esse valor corresponde à distância entre os elementos A e C nas variáveis medidas. Após essa fusão, a amostra formada por 3 elementos foi dividida em 2 grupos, o primeiro de tamanho 2 contendo os elementos A e C e, o segundo de tamanho 1, formado pelo elemento B. No próximo passo o elemento 3 é reunido ao primeiro grupo formado, com nível de fusão de aproximadamente 3,2, obtendo um único cluster de tamanho 3. Nesse exemplo foram considerados apenas 3 elementos para ilustrar a terminologia utilizada na descrição de dendrogramas, no entanto em uma análise real de muitos objetos, diversos grupos são obtidos.

Ao se utilizar um procedimento hierárquico aglomerativo, depois que uma fusão é realizada, ela não será mais desfeita. Assim, quando o método coloca dois elementos em um mesmo grupo, eles não mais aparecerão em grupos diferentes. Para o pesquisador encontrar a melhor solução com o número de agrupamentos ótimo, ele deverá adotar algum critério de divisão

O pesquisador tem a difícil tarefa de decidir em qual altura o corte no dendrograma deve ser realizado para escolha do número final de grupos. Isso ocorre porque o objetivo dos processos de agrupamentos hierárquicos é agrupar os n grupos de tamanho 1 em um único grupo com todas as observações. Contudo, o interesse do pesquisador é agrupar as observações em vários grupos e, para isso, é necessário decidir uma regra de parada do processo, para obtenção de k grupos. Esse tópico será discutido na seção 0.1.4.

Os principais métodos hierárquicos aglomerativos são: ligação simples (vizinho mais próximo), ligação completa (vizinho mais distante), ligação média, centroide e método de Ward

Com exceção de Ward, as demais técnicas seguem um processo iterativo geral, denominado método de grupo de pares que será descrito a seguir:

Em um método hierárquico aglomerativo, inicialmente, as n observações são alocadas em n grupos de tamanho 1 e, após todos os passos de fusão dos grupos, é formado um único grupo contendo os n elementos. O algoritmo básico para todos os métodos é semelhante

(Início) Grupos $G_1, G_2, \ldots, G_n$ - cada um contendo uma única observação.
(1) Unir os grupos G_i e G_j mais próximos entre si e diminuir o número de grupos de 1.
(2) Se o número de grupos é igual a 1, parar; senão, retornar ao passo (1).

Porém, antes do processo iniciar, a matriz de distâncias entre os objetos precisa ser obtida (conforme apresentado na seção 0.1.1) e deverá ser recalculada a cada novo passo, de forma a contabilizar todas as distâncias entre grupos. Com base nessa matriz, o algoritmo de agrupamento irá definir quais os grupos mais próximos entre si, ou seja, aqueles que apresentam menores distâncias.

Após selecionar a medida de dissimilaridade é preciso decidir qual método de agrupamento aplicar. Há vários métodos hierárquicos aglomerativos e eles diferem na forma com que definem a distância entre um grupo recém-formado a uma observação ou a outros grupos já existentes Os procedimentos aglomerativos incluem:

- Vizinho mais próximo (ligação simples): a distância entre dois grupos é a menor distância entre dois elementos dentro dos dois grupos.
- Vizinho mais distante (ligação completa): oposto ao vizinho mais próximo, define a distância entre dois grupos como sendo a maior distância entre quaisquer dois elementos dos dois grupos.
- Ligação média (distância média): a distância entre dois grupos é definida como a distância média entre todos os pares de elementos dos dois grupos.
- Centroide: o centro geométrico (centroide) de cada grupo é calculado primeiro. A distância entre os dois grupos é definida como a distância entre os dois centroides.
- Ward: método distinto dos demais, pois não combina os dois objetos mais semelhantes sucessivamente. Em vez disso, os objetos cuja fusão resulte na menor variância dentro do grupo são combinados.

O método de Ward, também denominado de método de variância mínima, foi proposto por Diferentemente dos outros métodos hierárquicos aglomerativos, não segue o algoritmo básico de agrupamento apresentado. A razão dessa diferença é que ele não busca a menor distância entre dois grupos, mas sim a menor soma de quadrados dentro do grupo, ou seja, a menor variância interna. Num primeiro momento, ele permite a redução dos n conjuntos iniciais a $n-1$ conjuntos mutuamente exclusivos considerando a fusão dos dois grupos, dentre todos os $n(n-1)/2$ pares possíveis, que resulte na menor soma de quadrados. Esse processo é repetido até haver um único grupo.

Segundo o processo iterativo dessa técnica segue os seguintes passos:

(1) Inicialmente, as n observações são alocadas em n grupos de tamanho 1, representados por G_1, G_2, $\ldots$, G_n.
(2) A cada passo do processo de agrupamento, a soma dos quadrados dentro de cada grupo é calculada como a soma do quadrado da distância euclidiana de cada elemento do grupo em relação ao vetor de médias do grupo. Assim, a soma de quadrados SQ_i de um grupo G_i é definida por:

$$SQ_i = \sum_{j=1}^{n_i} (\boldsymbol{X_{ij}} - \boldsymbol{\bar{X}_i})^T (\boldsymbol{X_{ij}} - \boldsymbol{\bar{X}_i}),$$

em que, n_i é o número de elementos no grupo G_i quando se está no passo m do processo, $\boldsymbol{X_{ij}}$ é o vetor de observações do j-ésimo elemento amostral que pertence ao i-ésimo grupo e $\boldsymbol{\bar{X}_i}$ é o vetor de médias do grupo.

No passo m, a soma de quadrados total dentro dos grupos é dada por:

$$SQT = \sum_{i=1}^{g_m} SQ_i,$$

em que, g_m é o número de grupos no passo m.

A distância entre dois grupos G_r e G_s é definida como a soma de quadrados entre eles, dada por:

$$d(G_r, G_s) = \left[\frac{n_r n_s}{n_r + n_s} \right] (\boldsymbol{\bar{X}_r} - \boldsymbol{\bar{X}_s})^T (\boldsymbol{\bar{X}_r} - \boldsymbol{\bar{X}_s}), \tag{1}$$

que é a soma de quadrados entre os grupos G_r e G_s. A cada passo de aplicação do processo iterativo, os dois grupos que minimizarem o valor de (1) são unidos.

Ainda de acordo com os métodos de Ward e da máxima verossimilhança são relacionados quando a distribuição dos dados é normal multivariada. Porém, para a aplicação do método de Ward não é necessário que os dados sejam normais multivariados, basta que as variáveis sejam quantitativas, já que ele se baseia no cálculo de médias.

Assim como no método do centroide, a distância entre dois grupos é definida considerando os vetores de médias amostrais, no entanto, o método de Ward considera a diferença entre o número de elementos em cada um dos grupos que estão sendo comparados. Dessa forma, o fator $\left[\frac{n_r n_s}{n_r + n_s} \right]$ pondera a distância de dois grupos r e s de tamanhos diferentes. Quanto maiores os valores de n_r e n_s, maior será o fator de ponderação e, portanto, maior a distância entre os vetores de médias comparados. Para a aplicação do método de Ward, é necessário que as p variáveis sejam quantitativas para que seja possível o cálculo dos vetores de médias

Em situações práticas, a tarefa de escolher um dentre os métodos hierárquicos aglomerativos pode se tornar difícil, já que os agrupamentos obtidos para o mesmo conjunto de dados podem ser bastante diferentes dependendo do método utilizado Para auxiliar nessa tarefa, dividem estudos empíricos sobre os métodos de agrupamento em dois tipos: estudos de simulação (em que se sabe a estrutura dos dados e uma avaliação dos métodos é feita em relação à recuperação dessa estrutura) e estudos reais (em que o critério é a interpretabilidade dos grupos obtidos). O que deve ficar claro é que não há um método que deve ser recomendado e que é melhor do que todos os outros, mas algumas observações gerais podem ser feitas e que são descritas a seguir.

O método do vizinho mais próximo tende a ser menos custoso computacionalmente, mas menos satisfatório do que outros métodos por causa do efeito de encadeamento ou *chaining* (novos grupos formados tendem a se unir a uma nova observação simples e não a um grupo já existente) Já o vizinho mais distante, por ser baseado em distâncias máximas, é muito afetado por *outliers* e os grupos obtidos tendem a ser compactos Os métodos da ligação média e centroide tendem a formar grupos com melhores partições do que os da ligação simples e completa. Além disso, os grupos resultantes possuem aproximadamente a mesma variância interna

enumeram alguns trabalhos em que o método de Ward retornou melhores resultados do que os outros, mas ressaltam que ele tende a parecer um bom método, mas tende a impor uma estrutura

esférica aos dados que pode não existir. No mesmo sentido, de acordo com muitos trabalhos apontam os métodos de Ward e da ligação média como tendo os melhores desempenhos de forma geral, mas seu desempenho depende dos dados. comparou os métodos para conjuntos de dados simulados e concluiu que o método de Ward foi o mais preciso e é o mais indicado para dados quantitativos contínuos. Geralmente, os grupos resultantes pelo método de Ward possuem o mesmo número de objetos, são convexos e compactos Assim, é uma boa escolha aplicar esse método se é esperado que haja grupos de iguais tamanhos ou se esses grupos obtidos favoreçam a interpretabilidade

Além das técnicas hierárquicas, há outros métodos de agrupamento que particionam as observações em um número específico de grupos utilizando como critério a minimização ou maximização de algum critério. Um desses métodos de otimização mais populares é o das k-médias, a ser descrito em seguida

0.1.3 Técnicas não hierárquicas: k-médias

Como o próprio nome diz, os métodos hierárquicos não seguem a propriedade da hierarquia, isso significa que mesmo se dois objetos forem unidos em algum passo do processo, pode ser que eles não permaneçam no mesmo grupo na partição final. E, portanto, isso implica que não é possível construir dendrogramas para a representação dos agrupamentos formados passo a passo Há métodos não hierárquicos baseados em estimação de densidades, misturas de distribuição e partição. Os procedimentos de partição são os mais utilizados e um deles, o das k-médias, é o mais popular

O procedimento de agrupamento das k-médias (*k-means*) tem como principais características: aplicação do processo à matriz de dados $\boldsymbol{X}$ e número de grupos k definido *a priori* A técnica procura uma partição das n observações em k agrupamentos $(G_1, G_2, \ldots, G_k)$, em que G_i denota o conjunto de observações que está no i-ésimo grupo e k é dado por algum critério numérico de minimização. A implementação mais usada do método tenta encontrar a partição dos n elementos em k grupos que minimizem a soma de quadrados dentro dos grupos (SQDG) em relação a todas as variáveis. Esse critério pode ser escrito como

$$SQDG = \sum_{j=1}^{p} \sum_{l=1}^{k} \sum_{i \in G_l} (X_{ij} - \bar{X}_j^{(l)})^2,$$

em que $\bar{X}_j^{(l)} = \dfrac{1}{n_i} \displaystyle\sum_{i \in G_l} X_{ij}$ é a média dos indivíduos no grupo G_l em relação à variável j

Ainda de acordo com os autores, apesar de o problema parecer relativamente simples, ele não é tão direto. A tarefa de selecionar a partição com a menor soma de quadrados dentro dos grupos se torna complexa porque o número de partições possíveis se torna muito grande mesmo com um tamanho amostral não tão grande. Por exemplo, para $n = 100$ e $k = 5$, o número de partições é da ordem de 10^{68}. Esse problema levou ao desenvolvimento de algoritmos que não garantem encontrar a solução ótima, mas que levam a soluções satisfatórias.

Segundo o processo iterativo do método pode ser descrito pelos seguintes passos:

(1) Inicialmente são escolhidos k centroides, denominados de "sementes", calculados com base no número de grupos escolhido *a priori*;

(2) Uma medida de distância é aplicada para comparar cada objeto a cada centroide inicial, então o objeto é unido ao grupo de menor distância;

(3) Os valores dos centroides são recalculados considerando cada grupo formado, então o passo 2 é repetido com os novos vetores de médias calculados para os novos grupos;

(4) Os passos 2 e 3 são repetidos até que não haja mais realocação dos objetos entre os grupos.

O agrupamento final obtido através do método das k-médias depende diretamente da escolha das sementes (passo 1) Diversas sugestões para a definição das sementes são apresentadas na literatura,

apresenta algumas propostas, sendo elas: aplicação de técnicas hierárquicas aglomerativas, escolha aleatória ou via observação dos valores discrepantes do conjunto de observações.

Segundo a autora, as sementes iniciais podem ser escolhidas com base no número de grupos obtidos após a aplicação de uma técnica hierárquica aglomerativa. Nesse caso, o vetor de médias de cada grupo é calculado e utilizado como semente para o uso do método das k-médias. O método de Ward é frequentemente utilizado para selecionar os centroides iniciais porque o critério de fusão de grupos com base na menor soma de quadrados dentro do grupo, utilizado no método de Ward, é próximo ao critério do quadrado da soma de erros de partição do método k-médias A segunda sugestão se baseia na escolha aleatória a partir de um procedimento de amostragem aleatória simples repetido m vezes, produzindo para cada grupo o centroide das m sementes selecionadas. Outra regra de decisão se baseia na seleção de k elementos discrepantes, em relação às p-variáveis no conjunto de dados, como sementes de um agrupamento inicial

Não há um consenso sobre o melhor método para escolha do número de grupos inicial ou de seus centroides, contudo, alguns autores aconselham que o processo seja realizado com diferentes escolhas para buscar a melhor solução de agrupamento apresentam um estudo comparativo de diferentes métodos de inicialização para as k-médias e a escolha aleatória aparece como um dos melhores dentre os comparados, por tornar o procedimento das k-médias mais efetivo e independente do agrupamento inicial.

De acordo com o método não hierárquico das k-médias é superior aos métodos hierárquicos por ser menos afetado por *outliers* e por variáveis não relevantes para o agrupamento. E, por ser um método mais eficiente computacionalmente, pode ser aplicado a grandes conjuntos de dados. Uma desvantagem apontada pelo autor é a necessidade de definir previamente o número de grupos e esse ponto será discutido na seção 0.1.4.

0.1.4 Número de grupos

Nas aplicações de métodos de agrupamento, o pesquisador precisa, em algum momento, decidir o número apropriado de grupos, independente do método utilizado. Essa é a etapa final nos métodos de agrupamento hierárquicos aglomerativos e a inicial nos agrupamentos não hierárquicos. Isso acontece porque as técnicas hierárquicas aglomerativas iniciam o procedimento com k observações separadas em k grupos. A cada passo, o algoritmo reúne duas observações ou grupos e ao final, um único grupo com as k observações é obtido. Portanto, é preciso que uma regra de corte seja estabelecida, para que o número ideal de grupos seja escolhido. No uso de métodos não hierárquicos, como o das k-médias, a escolha do número de grupos acontece antes da aplicação do método porque, por definição, essas técnicas exigem que o número de grupos seja escolhido *a priori*

De acordo com a escolha do número apropriado de grupos está sujeita a dois tipos de erros diferentes. O primeiro acontece quando a regra de parada seleciona um número k de grupos maior do que o adequado. O segundo tipo ocorre quando a regra de decisão conduz a escolha de um número de grupos menor do que o apropriado. Apesar dos dois tipos de erros serem indesejáveis, o segundo produz consequências consideradas mais sérias, pois informação é perdida. De forma geral, essa não é considerada uma tarefa simples. A seguir, serão apresentadas diferentes abordagens propostas na literatura

Um método gráfico utilizado para a escolha do número adequado de agrupamentos é o corte no dendrograma. A questão, entretanto, é decidir onde o corte deve ser feito. Segundo uma maneira informal de tomar essa decisão é avaliar as mudanças na altura do gráfico nos diferentes passos e escolher a maior observada. A ideia é que a maior mudança representa uma maior diferença no nível de fusão, o que sugere que o grupo pode se tornar menos homogêneo internamente com essa união.

A Figura 5 ilustra um exemplo onde 23 observações foram agrupadas pelo método hierárquico distância média. O ponto de maior mudança é facilmente identificado, o que produz uma divisão final com 3 agrupamentos. Mas nem sempre é fácil visualizar onde o corte deve ser realizado. Na Figura 7, que ilustra o dendrograma de 20 outras observações agrupadas utilizando o método hierárquico

ligação simples, apesar do número de observações não ser muito grande, não é simples decidir onde o corte deve ser feito.

Figura 5: Ilustração de corte no dendograma com 23 observações

Figura 6: *

Fonte: elaboração própria

Figura 7: Ilustração de dendograma com 20 observações

Figura 8: *

Fonte: elaboração própria

Outros critérios informais utilizam um gráfico da medida *versus* o número de grupos. Nesses métodos, o objetivo é identificar grandes mudanças no gráfico para um determinado número k. Segundo dentre os métodos heurísticos está o *scree plot*. A proposta do método é que se faça um gráfico do número de grupos k da solução de agrupamento *versus* uma medida de erro W_k correspondente a cada passo. A medida W_k decresce monotonicamente conforme k aumenta, contudo, em um certo ponto, W_k cai abruptamente formando uma quebra do tipo "cotovelo", que indica o número de grupos que deve ser escolhido.

Como os critérios gráficos são subjetivos, várias técnicas mais formais têm sido propostas e alguns trabalhos avaliaram suas propriedades Nesse sentido, segundo a estatística *gap* foi proposta com o objetivo de formalizar a ideia de procurar pelo "cotovelo" no gráfico do número de grupos *versus* algum critério de otimização. Os autores fizeram simulação de cinco cenários para a avaliação de seis critérios de escolha do número de grupos. As estatísticas comparadas foram as propostas por e duas variações da estatística *gap*. O trabalho de além de propor um método, também o compara com outros, inclusive a estatística *gap*, que se saiu melhor quando os dados seguiam a normal multivariada e com o uso da distância euclidiana. O mesmo ocorre com outras distribuições simétricas, contudo, ela falha quando os dados seguem distribuições assimétricas.

Os autores apresentaram um estudo de simulação de Monte Carlo para comparar 30 critérios de determinação do número de grupos. Os autores não incluíram nenhum método gráfico na análise, pois o objetivo foi testar as técnicas que buscam eliminar a subjetividade presente nesses métodos. Além disso, usaram dois critérios externos: índice de Jaccard e estatística Rand ajustada. Basicamente esses critérios usam informações externas ao processo de agrupamento para validação dos grupos obtidos. Nesse trabalho, a informação externa era a real estrutura dos grupos. Os autores utilizaram conjuntos de dados artificiais que continham 2, 3, 4 ou 5 grupos não sobrepostos, contendo 50 observações cada e bem separados. Com o intuito de obter diferentes partições finais, esses conjuntos de dados fictícios foram analisados por quatro métodos de agrupamento hierárquicos diferentes, sendo eles vizinho mais próximo, vizinho mais distante, distância média e método de Ward.

O trabalho citado identificou como as cinco técnicas com melhores desempenhos: em primeiro lugar a estatística pseudo F em seguida o critério Je(2)/Je(1) C-*Index* Gamma e Beale Esses também foram os 5 métodos identificados por como os melhores. A estatística pseudo F é a que aparece mais frequentemente na literatura como tendo o melhor dexempenho na maioria das situações ressaltam que essa medida apresenta um bom custo-benefício, por levar em conta simplicidade e adequacidade.

Ainda de acordo com a estatística Traço de W apresentou desempenho ruim, apesar de ser uma das mais populares. A estatística $|T|/|W|$, proposta por não acertou em nenhuma das 432 tentativas. No entanto, os autores ressaltaram que os resultados estão sujeitos a serem dependentes da estrutura de dados, ou seja, pode ser que a ordenação dos melhores testes seja modificada caso sejam testados com uma estrutura de dados diferentes. Os dados foram gerados com o uso da normal multivariada

e isso pode ter contribuído para que alguns métodos não tenham um bom desempenho. Ou seja, em outras situações, o resultado poderia ser diferente.

Diante dessa diversidade de abordagens e por não haver um critério melhor em todas as situações, é melhor levar em conta considerações práticas. Em alguns casos, pode haver alguma ideia *a priori* ou uma teoria que sugira uma estrutura nos dados. Entretanto, o mais importante é que os resultados sejam interpretáveis e tenham significado prático e útil