Aula 11 – Mineração de Dados
Agrupamento de Dados: Validação

Profa. Elaine Faria

UFU

Validação de Agrupamento

- Em tarefas de classificação
  - A avaliação dos resultados do modelo de classificação é parte do processo de desenvolvimento
  - Há medidas e procedimentos de avaliação bem aceitos
    - Ex: acurácia, validação cruzada, etc.
- Em tarefas de agrupamento
  - A avaliação dos grupos (ou das partições) não está bem desenvolvida ou não é comumente usada como parte da análise de agrupamentos

Tendência e Validação

- Tendência de agrupamento
  - Observar antes de executar um algoritmo de agrupamento
    - Para garantir que existe nos dados uma estrutura significativa (não aleatória)
    - Utiliza testes estatísticos
- Validade de agrupamento
  - Verificar após executar um algoritmo de agrupamento
    - Estima o desempenho do algoritmo
    - Utiliza testes estatísticos e percepção do especialista (subjetivo)

Validação de Agrupamento

- Refere-se aos procedimentos que avaliam os resultados da análise de agrupamento de uma forma quantitativa e objetiva

Como avaliar os grupos gerados por um algoritmo de agrupamento?

Por que avaliar agrupamentos?

Validação de Agrupamento

- Questões importantes para o agrupamento
  - Determinar a tendência de agrupamento
    - Distinguir se há estrutura de grupos não-aleatória nos dados
  - Determinar o nro correto de grupos
  - Avaliar o quanto os resultados da análise de agrupamento se adequam aos dados, sem usar referência à informação externa
  - Comparar os resultados da análise de agrupamento a resultados conhecidos
  - Comparar duas partições para determinar qual é a melhor

(a) Original points.

(b) Three clusters found by DBSCAN.

(c) Three clusters found by K-means.

(d) Three clusters found by complete link.

Figure 8.26. Clustering of 100 uniformly distributed points.

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Medidas de avaliação

- Índice ou critérios internos
  - Medem a qualidade da partição sem usar informação externa
  - Ex: SSE
- Índice ou critérios externos
  - Medem o quanto os rótulos dos grupos casam com uma estrutura externa (ex: classes verdadeiras)
- Índice ou critérios relativos
  - Comparam duas partições ou dois grupos

Critério de Validação Externo

- Assumem que o *ground-truth* (rótulo verdadeiro) do agrupamento é conhecido
  - Informação externa é usada para avaliar a partição dos dados
- Como obter o rótulo
  - *Data sets* de classificação, que especificam o rótulo (classe) para cada exemplo podem ser usados
  - *Data sets* artificiais podem ser criados, com uma estrutura de grupos conhecida, com um rótulo para cada exemplo
  - Experiência do especialista de domínio

Critério de Validação Externo

- Formalizando o problema

  X_i, com i=1,2,..., N → *data set*

  $Y_i \in \{1,2,...,K\}$ → *ground-truth* de cada exemplo

  $T = \{T_1, T_2, ..., T_k\}$ → *ground-truth* do agrupamento, onde T_j corresponde a todos os pontos com o label *j*

  $C = \{C_1, ..., C_r\}$ → resultado do agrupamento usando algum algoritmo

- Nomenclatura
  - grupos da partição de referência (*ground truth*) → classes
  - grupos da partição sob avaliação → *clusters*

Critério de Validação Externo

- Orientado a classificação
  - Usam medidas de classificação como entropia, pureza, f-measure,etc. para avaliar o quanto um cluster contém objetos de uma única classe
- Orientado a similaridade
  - Está relacionado a medidas de similaridade para dados binários como, por exemplo, o coeficiente de Jaccard

Critério de Validação Externo

- Pureza
  - Mede o quanto cada *cluster* contém exemplos de uma única classe

$$p_i = \max_j p_{ij}$$

Pureza do i-ésimo *cluster*

$$pureza = \sum_{i=1}^{k} \frac{m_i}{m} pi$$

Pureza total

Critério de Validação Externo

Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Exemplo reitrado de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Critério de Validação Externo

- Critérios Orientados a similaridade
  - Comparam duas matrizes
    - A matriz de similaridade do agrupamento
      - 1 na célula ij, se dois objetos i e j estão no mesmo cluster
      - 0, caso contrário
    - A matriz de similaridade das classes
      - 1 na célula ij, se dois objetos i e j estão na mesma classe
      - 0, caso contrário

- Definindo
  - f_{11}: Nro de pares que pertencem à mesma classe e ao mesmo *cluster*
  - f_{10}: Nro de pares que pertencem à mesma classe e a *clusters* distintos
  - f_{01}: Nro de pares que pertencem a classes distintas e ao mesmo *cluster*
  - f_{00}: Nro de pares que pertencem a classes e *clusters* distintos

Critério de Validação Externo

- *Rand Index*
  - Medida orientada a similiaridade
    - Compara duas partições
  - Similar ao coeficiente de casamento simples

$$RandIndex = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Critério de Validação Externo

- Coeficiente de *Jaccard*
  - Medida orientada a similiaridade
    - Compara duas partições
  - Elimina o termo f_{00}

$$Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Critério de Validação Externo

- *Rand Index* e Coeficiente de *Jaccard*
  - Variam no intervalo [0,1]
  - Valores altos para esses índices indicam alto grau de similaridade entre a organização em grupos e a organização das partições

Critério de Validação Externo

- Exemplo
  - 5 objetos: p_1, p_2, p_3, p_4 e p_5
  - Dois *clusters*: $C_1=\{p_1,p_2,p_3\}$ e $C_2=\{p_4,p_5\}$
  - Duas classes: $L_1=\{p_1,p_2\}$ e $L_2=\{p_3,p_4,p_5\}$

Table 8.10. Ideal cluster similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Table 8.11. Ideal class similarity matrix.

Point	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

Exemplo reitrado de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Critério de Validação Externo

- Exemplo - continuação

  $f_{00} = 4$ $f_{01} = 2$ $f_{10} = 2$ $f_{11} = 2$

RandIndex= (2+4)/10 = 6/10= 0.6

Jaccard = 2/(2+2+2) = 2/6 = 0.33

Critério de Validação Externo de Hierarquias

- Método Direto
  - Aplicar um critério externo N-2 vezes
    - Um para cada nível intermediário das hierarquias em questão

  Compor esses resultados (soma, média, etc.)

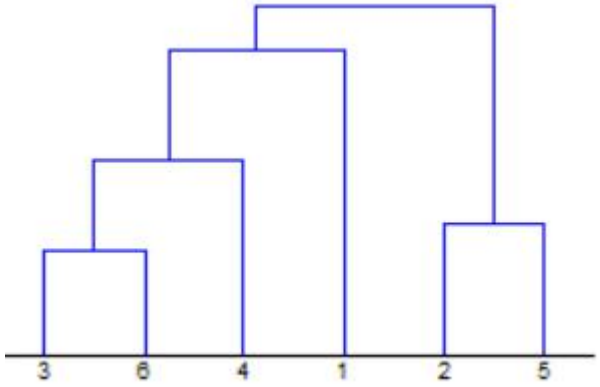
- Método Indireto
  - Avaliar a correlação entre duas matrizes que representam as hierarquias sendo comparadas

Critério de Validação Externo de Hierarquias

- Correlação estatística Γde Hubert

$$\Gamma = \frac{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}[X(i,j)-\mu_x][Y(i,j)-\mu_y]}{\sigma_x \sigma_y}$$

Matrizes: elemento (i,j) é igual a "c" se os objetos i e j aparecem unidos pela 1a vez no c-ésimo nível hierárquico



$$\mathbf{Y} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 6 & 5 & 5 & 6 & 5 \\ & 6 & 6 & 3 & 6 \\ & & 4 & 6 & 2 \\ & & & 6 & 4 \\ & & & & 6 \\ & & & & \end{bmatrix}$$

Critério de Validação Interno

- Usadas quando não se dispõe do *ground-truth*
- Medem o quanto os dados se ajustam a estrutura de grupos obtida
- Divididos em
  - Medidas de coesão dos grupos
    - Determina o quão relacionados estão dois objetos em um grupo
  - Medida de separação dos grupos
    - Determina o quão distintos ou bem separados estão um grupo de outros

Critério de Validação Interno

- Em geral, usam uma matriz de distâncias (ou matriz de proximidade)
- SSE (já vimos ao estudar o K-Means)
  - Pode ser usado por exemplo para identificar o nro de grupos

$$J = \sum_{c=1}^{k} \sum_{x_j \in C_c} d(x_j, \bar{x}_c)^2$$

Critério de Validação Interno de Hierarquias

- Usar um critério interno em sucessivas partições produzidas por um algoritmo hierárquico

- O resultado pode indicar um ponto de corte no dendograma → regra de parada

Critério de Validação Relativos

- Usado para indicar qual a melhor dentre duas ou mais partições
  - O termo pode ser usado para critério de validação interno → depende do contexto
- Usado para comparar diferentes execuções de um algoritmo com diferentes parâmetros
  - Ex: diferentes valores de k

Critério de Validação Relativos

- Largura da Silhueta (SWC)
  - Combina coesão com separação
  - Calculada para cada objeto que faz parte de um agrupamento
  - Baseada na proximidade entre os objetos de um *cluster* e na distância dos objetos de um *cluster* ao cluster mais próximo
  - Mostra quais objetos estão bem situados dentro dos seus *clusters* e quais estão fora de um *cluster* apropriado

Critério de Validação Relativos

- Largura da Silhueta (SWC)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$SWC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

$$SWC \in [-1, +1]$$

a(i): dissimilaridade média do i-ésimo objeto ao seu cluster

b(i): dissimilaridade média do i-ésimo objeto ao seu cluster vizinho mais próximo

Silhueta Original: a(i) e b(i) são calculados como a distância do i-ésimo objeto a todos os demais objetos do cluster em questão → Complexidade $O(N^2)$

Critério de Validação Relativos

- Largura da Silhueta (SWC) - Algoritmo

1. For the i^{th} object, calculate its average distance to all other objects in its cluster. Call this value a_i.
2. For the i^{th} object and any cluster not containing the object, calculate the object's average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters; call this value b_i.
3. For the i^{th} object, the silhouette coefficient is $s_i = (b_i - a_i)/\max(a_i, b_i)$.

Algoritmo reitrado de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Critério de Validação Relativos

- Largura da Silhueta Simplificada (SSWC)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i): dissimilaridade média do i-ésimo objeto ao seu cluster

b(i): dissimilaridade média do i-ésimo objeto ao seu cluster vizinho mais próximo

$$SWC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

Silhueta Simplificada: a(i) e b(i) são calculados como a dissimilaridade do i-ésimo objeto ao centróide do cluster em questão→ Complexidade O(N)

$$SWC \in [-1, +1]$$

Tendência de Agrupamento

- Teste de Hopkins é um dos testes usados para verificar se existe tendênciais de agrupamento nos dados
  - Ele mede a propabilidade de que um dado conjunto de dados é gerando por uma distribuição uniforme
    - Testa a aleatoriedade espacial dos dados

Tendência de Agrupamento

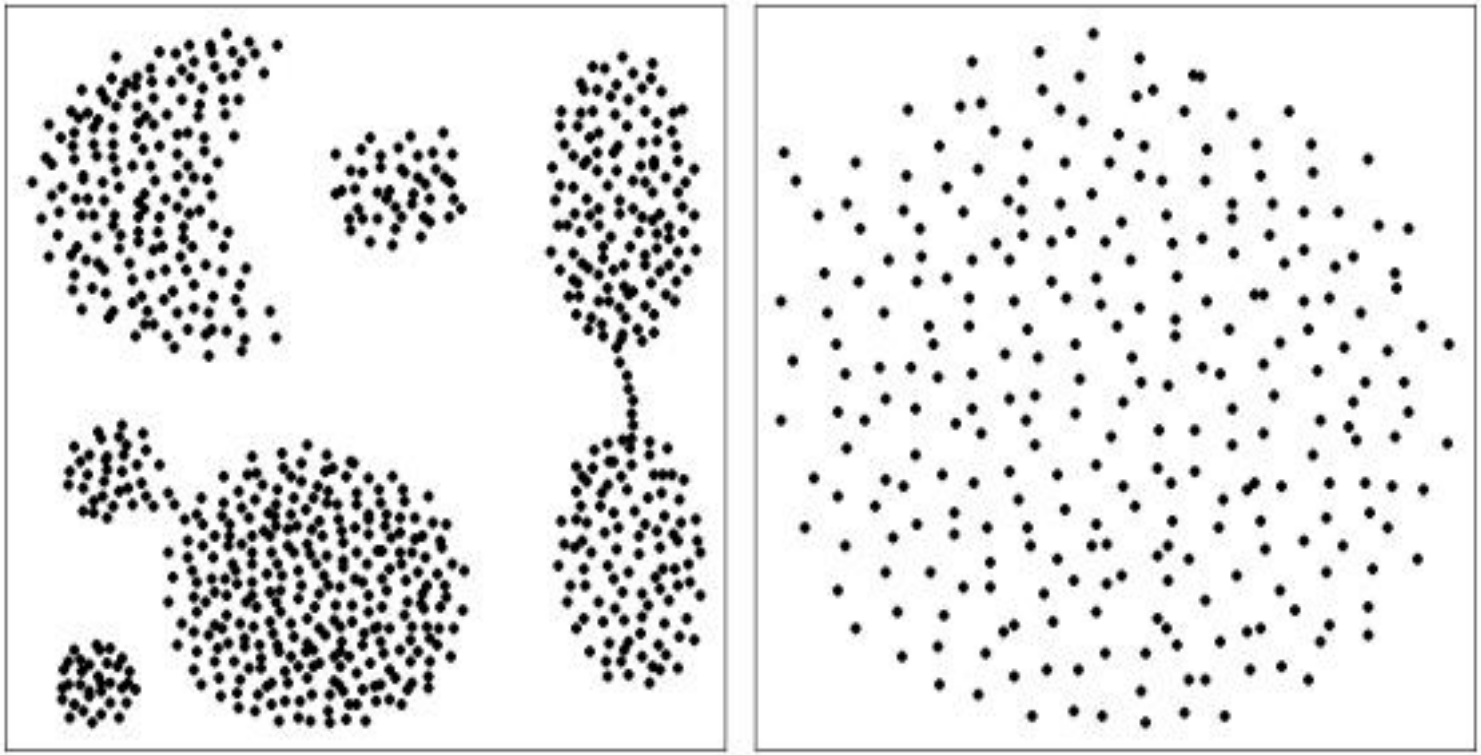


Figura 1 - Dados contendo agrupamentos naturais, com diferentes homogeneidades e separações (esquerda). Dados sem agrupamentos naturais (direita).

Imagem retirada da dissertação de mestrado: ANÁLISE DE AGRUPAMENTOS APLICADA À SETORIZAÇÃO GEOTÉCNICA DE MINAS A CÉU ABERTO - Ana Flávia Delbem Vidigal Nazareth - UFOP 2020

Hopkins - Algoritmo

- Amostre n pontos (p1,…, pn) a partir da base de dados D de forma aleatória.
- Para cada ponto pi, encontre seu vizinho mais próximo pj ∈D; compute a distância entre pi and pj e denote-a como xi=dist(pi,pj)
- Gere um conjunto de dados simulado (randomD) criado a partir de uma distribuição de dados uniforme como n points (q1,…, qn) com valores aleatórios no espaço de cada uma das p-dimensões do conjunto de dados D.
- Para cada ponto qi∈randomD, encontre o seu vizinho mais próximo in D; calcule a distância entre qi e qj e denote-a como yi=dist(qi,qj)
- Calcule a estátistica de Hopkins (H) de acordo com a fórmula

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Hopkins - Algoritmo

- Na Equação, busca-se a maximização de H, cujo valor pertence ao intervalo [0;1]
  - Em uma instância em que objetos estão em grupos bem definidos, coesos e bem separados, a distância média entre os objetos é pequena. Nesse caso o somatório de xi tende a ser próximo de 0 e, consequentemente, H é próximo de 1.
  - Em instâncias em que os objetos estão dispersos no espaço, os somatórios de xi e yi são próximos, ou seja, o valor de H é próximo a 0,5.

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
- Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.
- Zaki, M. J., Meira, W. Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.