MATB16

Laboratório de Inteligência Artificial

Validação Não-Supervisionado

Tatiane Nogueira Rios
Ricardo Araújo Rios

LabIA
Instituto de Computação - UFBA

Objetivo

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

(Jain and Dubes, Algorithms for Clustering Data, 1988)

Agenda

- Introdução
- Critérios Internos
- Critérios Externos
- Critérios Relativos

Objetivo

- Conhecer os diferentes índices de validação de agrupamento para comparar algoritmos ou verificar a qualidade dos grupos obtidos.

Introdução

- Análise dos resultados de agrupamentos não é uma tarefa fácil:
  - Dados não rotulados;
  - Diferentes tipos de agrupamento;
  - Diferentes algoritmos de agrupamento;
  - Diferentes parâmetros e valores para cada agrupamento

Introdução

- A validação é utilizada para:
  - Comparar algoritmos de agrupamento;
  - Validar grupos encontrados por algoritmos;

Introdução

- A validação é utilizada para:
  - Comparar algoritmos de agrupamento;
  - Validar grupos encontrados por algoritmos;

Nem sempre existe uma resposta esperada e não existe resposta única!

Introdução

- A avaliação de agrupamento deve ser objetiva;
- Normalmente utiliza índices estatísticos;
- Análise qualitativa das estruturas (grupos) encontrados;
- A forma de aplicação dos índices de agrupamento é definida por **critérios de validação.**

Introdução

- **Índices**
  - Estatística utilizada para testar a validade de um agrupamento
- **Critérios de validação**
  - Estratégia para validar um agrupamento
  - Define como os índices são utilizados

Introdução

- Tipo de critérios de validação
  - Relativo
  - Interno
  - Externo

Introdução

- Critérios relativos
  - Comparam diversos agrupamentos respeitando algum critério. Ex.: Estabilidade;
  - Utilizado para comparar algoritmos.
  - Utilizado para determinar os parâmetros mais apropriados para os algoritmos.

Introdução

- Critérios internos
  - Mede a qualidade do agrupamento com base nos dados originais;
  - Exemplo: Matriz de Similaridades

Introdução

- Critérios externos
  - Mede a qualidade do agrupamento com base em alguma estrutura previamente definida;
  - O agrupamento obtido é avaliado considerando uma informação conhecida ou esperada a priori;
    - Intuição do analista de dados

Introdução

- Embora seja uma técnica não-supervisionada, algum conhecimento prévio pode ser utilizado:
  - Visualização dos dados;
  - Especialista de domínio;
  - Bases sintéticas;
  - Definição de hipóteses;

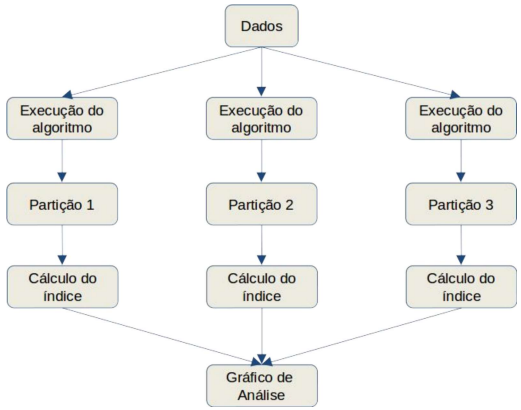
Introdução

- Um mesmo índice pode ser aplicado à diferentes critérios;
- O que diferencia a utilização de um índice entre critérios é a forma de sua aplicação;
- Critérios externos e internos são baseados em testes estatísticos;
  - Objetivo: Confirmar uma hipótese pré-determinada.

Introdução

- Índices devem:
  - Fazer sentido intuitivamente;
  - Ter uma base teórica;
  - Ser prontamente computável
  - Verificar uma estrutura não aleatória

Critério Relativo



Critério Relativo

- Exemplo de aplicação
  - Escolha do melhor número de **k** na aplicação de um algoritmo sobre um determinado conjunto de dados;
  - Neste caso, o algoritmo é aplicado para todos os possíveis números de **k**;
  - Índices são aplicados sobre cada resultado;
  - O melhor número de clusters pode ser definido pelo menor ou maior valor obtido ou ainda pela inflexão da curva observada;

Critério Relativo (índice)

- Silhueta
  - Baseia-se na similaridade entre objetos do mesmo grupo e na distância entre objetos de um cluster e objetos do cluster mais próximo.
  - Resultados no intervalo [-1, 1]
  - Melhor resultado quanto mais próximo de 1

Critério Relativo (índice)

- Silhueta

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

$$a(x_i, C_k) = \frac{1}{|C_k|} \sum_{x_j \in C_k} d(x_i, x_j)$$

$$b(x_i) = \min_{x_i \in C_i, C_i \neq C_j} a(x_i, Cj)$$

Critério Relativo (índice)

- Silhueta

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

$$a(x_i, C_k) = \frac{1}{|C_k|} \sum_{x_j \in C_k} d(x_i, x_j)$$

$$b(x_i) = \min_{x_i \in C_i, C_i \neq C_j} a(x_i, Cj)$$

Critério Relativo (índice)

- Silhueta
  - Quando a silhueta é calculada para cada objeto, seu valor será próximo de 1, se o objeto está bem situado dentro do seu cluster;
  - Valor próximo de -1 indica que o objeto deveria estar em outro cluster;

Critério Relativo (índice)

- Silhueta
  - Depende apenas da partição produzida e não do algoritmo utilizado;
    - Permite comparar resultados entre diferentes algoritmos
    - Permite melhorar o agrupamento escolhendo diferentes parâmetros

Critério Relativo (índice)

- Silhueta para cada cluster

$$sil(C_k) = \frac{1}{|C_k|} \sum_{x_i \in C_k} sil(x_i)$$

- Largura média das silhuetas

$$sil(\pi) = \frac{1}{n} \sum_{i=1}^{n} sil(x_i)$$

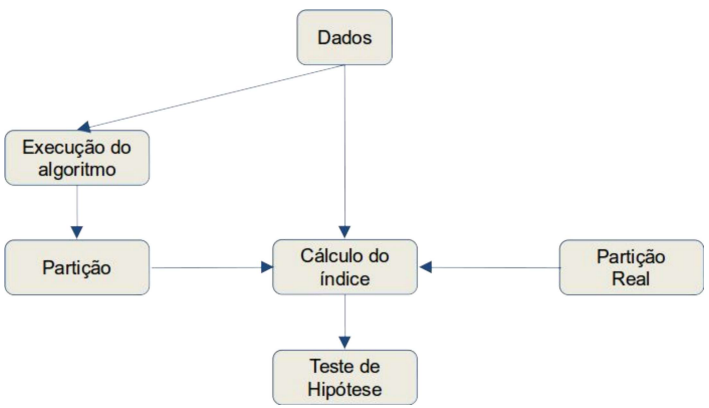
Critério Relativo (índice)

- Coeficiente de Silhueta
  - É o máximo valor de sil(π), tal que = 2, …, (n-1)
  - Quantifica a estrutura encontrada por um algoritmo
    - SC ≤ 0,25 → Nenhuma estrutura relevante encontrada
    - 0,26 ≤ SC ≤ 0,5 → Estrutura fraca/artificial
    - 0,51 ≤ SC ≤ 0,7 → Estrutura razoável
    - 0,71 ≤ SC ≤ 1 → Estrutura forte

Critério Relativo (índice)

- Silhueta
  - Indicada para clusters compactos e espaçados;
  - Funciona bem para clusters esféricos;
  - Favorece objetos associados a clusters com similaridade média/alta;
  - Largura de silhueta é tendenciosa para clusters sobrepostos, favorecendo agrupamento disjuntos.

Critério Externo



Critério Externo

- O agrupamento obtido confirma uma hipóstese pré-definida?
  - Utilização de testes de hipósteses
- Análise
  - π: partição obtida com algoritmo de agrupamento
  - β: partição real dos dados

Critério Externo

- Número de pares de objetos que:
  - a: pertencem ao mesmo grupo de π e β
  - b: pertencem ao mesmo grupo de π e a grupos diferentes de β
  - c: pertencem a grupos diferentes de π e ao mesmo grupo de β
  - d: pertencem a diferentes grupos de π e β

Critério Externo

- Número de pares de objetos que:
  - M = a + b + c + d
  - m_1 = a + b
  - m_2 = a + c

Critério Externo

- Índice Rand
  - Computa a probabilidade de dois objetos pertecerem ao mesmo grupo ou grupos diferentes nas duas partições.

$$R(\pi, \beta) = \frac{(a+d)}{M}$$

Critério Externo

- Índice Jaccard
  - Computa a probabilidade de dois objetos pertecerem ao mesmo grupo em ambas as partições.

$$J(\pi, \beta) = \frac{a}{a+b+c}$$

Critério Externo

- Índice Fowlkes e Mallows
  - Computa a semelhança entre duas partições;
  - Resultados variam no intervalo [0, 1]

$$FM(\pi, \beta) = \frac{a}{\sqrt{m_1 m_2}}$$

Exemplos de Datasets

- UCI (https://archive.ics.uci.edu/ml/datasets.html);
- UCI KDD Archive (http://kdd.ics.uci.edu/summary.data.application.html);
- Statlib (http://lib.stat.cmu.edu/);
- Delve (http://www.cs.utoronto.ca/~delve/);
- LETOR (http://research.microsoft.com/en-us/um/beijing/projects/letor/index.html);

Referências

- Faceli et al., Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina, LTC, 2015.
- Mitchell, T. M., Machine Learning, McGraw-Hill, 1997.
- Witten et al., Data Mining – Practical Machine Learning Tools and Techniques, 3d edition, Elsevier, 2011.
- Xu, R. and Wunsch, D.C., Clustering, 1a ed, Wiley, 2009
- J. Han; M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.