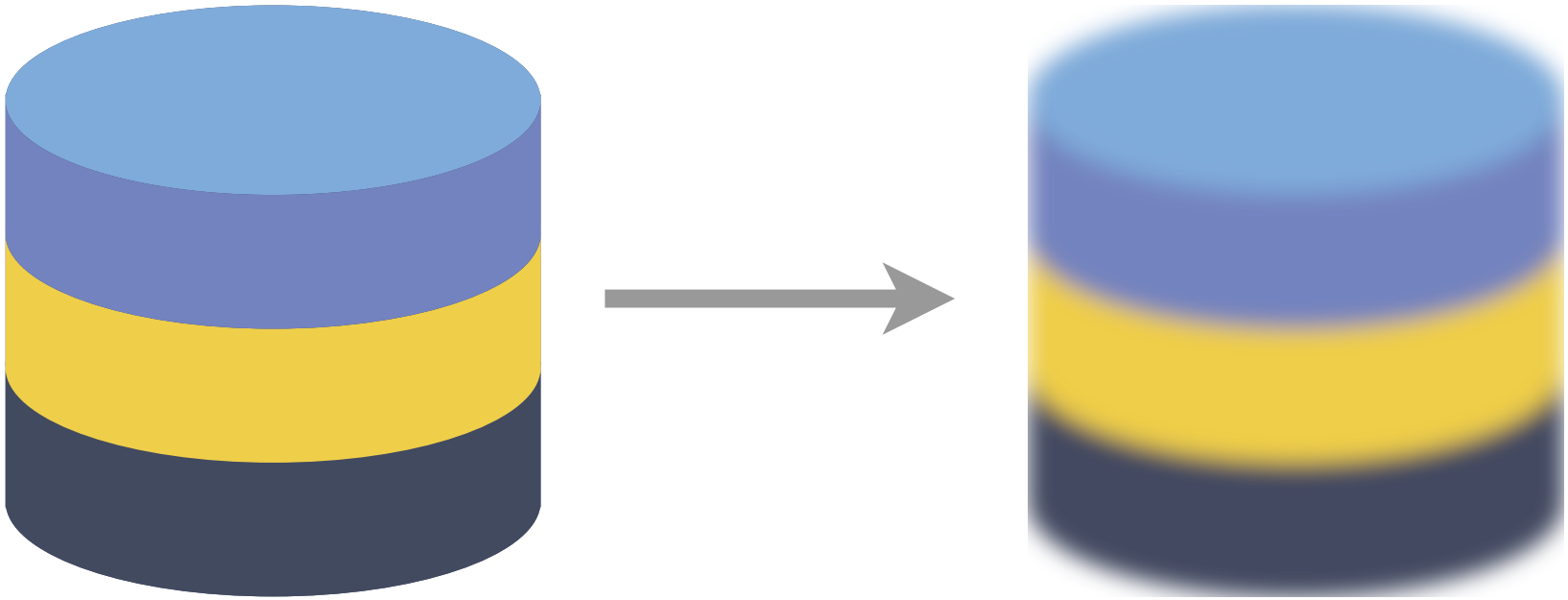
Medidas de Utilidade

Javam Machado

Laboratório de Sistemas e Banco de Dados

Agosto/2019

Anonimização



D D'

Anonimização é sinônimo de:

- Distorção
- Modificação dos dados originais
- Perda de informação
- Menor utilidade para o usuário final
- Privacidade!

- Como quantificar essas perdas/distorções?

Anonimização é sinônimo de:

- Distorção
- Modificação dos dados originais
- Perda de informação
- Menor utilidade para o usuário final
- Privacidade!

- Como quantificar essas perdas/distorções?
  - Por meio de métricas!

Métricas

Classificação das métricas

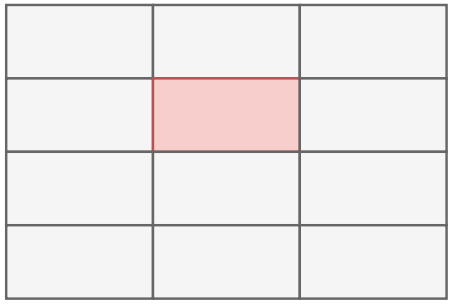
De uso geral

- Utilizadas no cenário em que o publicador dos dados não tem conhecimento prévio sobre a área de aplicação dos dados
- Ou quando não há objetivos específicos pré-definidos para o uso desses dados
- Devem reter a maior quantidade de informação sempre que possível

De finalidade específica

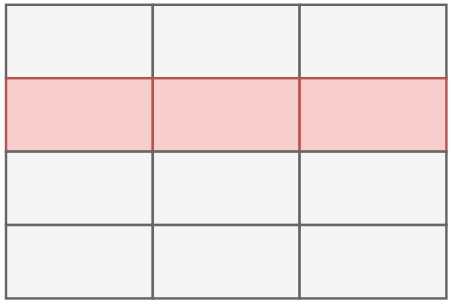
- Atende a demandas exclusivas de usuários
- Finalidade dos dados deve ser conhecida no momento da publicação
- Ex. técnica de classificação: certos atributos não devem ser anonimizados

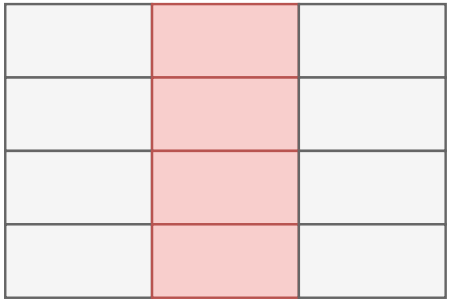
Métricas de uso geral



Orientada a células

Orientada a registros





Orientada a atributos

Orientada a células

Orientada a células

- Precisão
  - Penaliza cada instância de um valor de atributo que é generalizado ou suprimido
  - Quanto maior a precisão, maior a utilidade dos dados
  - $Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$
  - D: conjunto de dados
  - N_a: número de atributos semi-identificadores
  - h: altura da hierarquia de generalização de valor do atributo Ai após anonimização
  - $|HGV_{A_i}|$: altura máxima da hierarquia

Orientada a células – Precisão

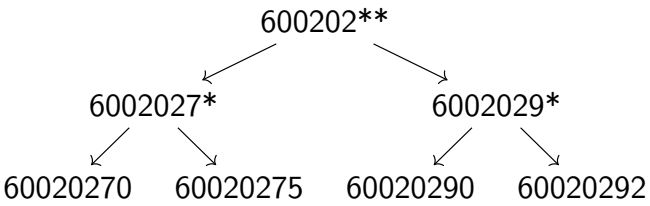


Figura: HGV_c

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
25	60020275
27	60020290
29	60020292

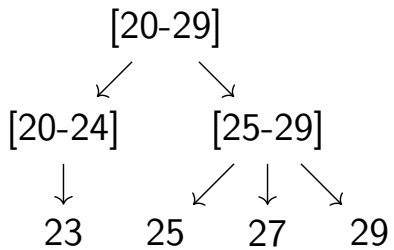


Figura: HGV_i

Orientada a células – Precisão

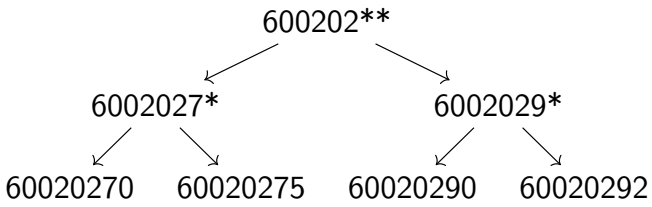


Figura: HGV_c

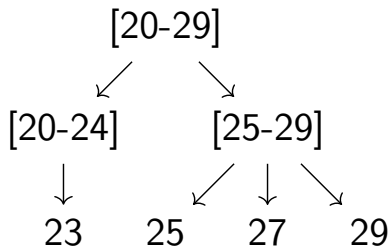


Figura: HGV_i

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
25	60020275
27	60020290
29	60020292

- $|D|=4$
- $h_i=1$
- $|HGV_c|=2$

- $N_a=2$
- $h_c=0$
- $|HGV_i|=2$

Precisão??

Orientada a células – Precisão

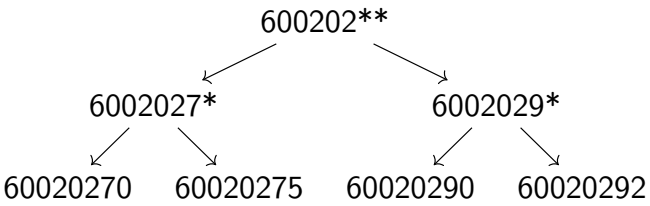


Figura: HGV_c

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
$[25-29]$	60020275
$[25-29]$	60020290
$[25-29]$	60020292

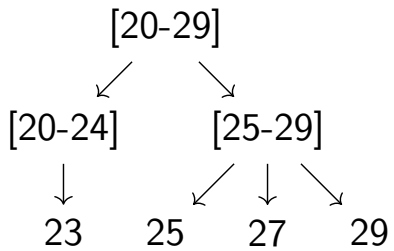


Figura: HGV_i

Orientada a células – Precisão

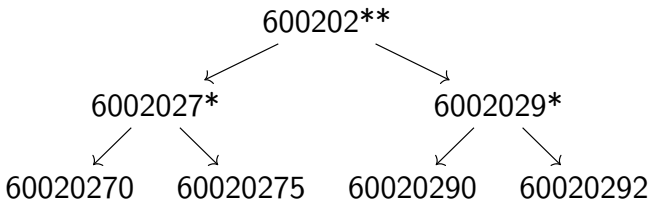


Figura: HGV_c

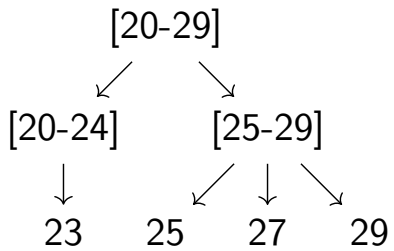


Figura: HGV_i

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
$[25-29]$	60020275
$[25-29]$	60020290
$[25-29]$	60020292

- $|D|=4$
- $h_i=1$
- $|HGV_c|=2$

- $N_a=2$
- $h_c=0$
- $|HGV_i|=2$

Precisão??

Orientada a células – Precisão

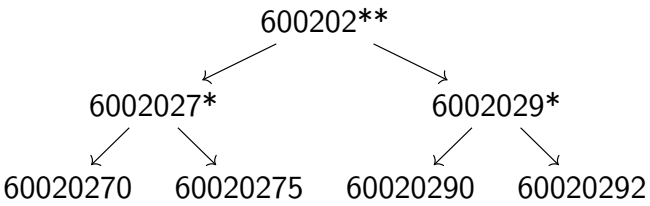


Figura: HGV_c

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	600202**
$[25-29]$	600202**
$[25-29]$	600202**
$[25-29]$	600202**

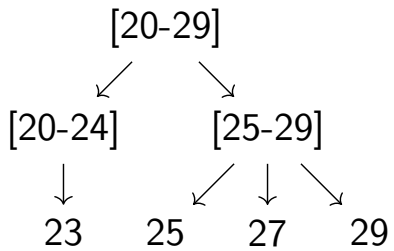


Figura: HGV_i

Orientada a células – Precisão

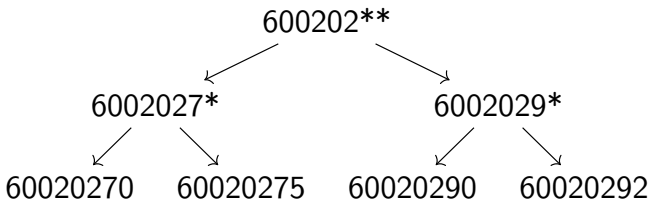


Figura: HGV_c

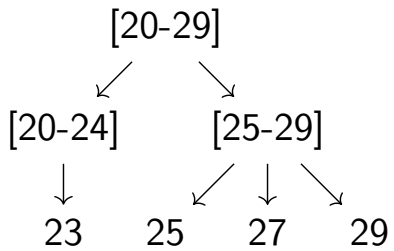


Figura: HGV_i

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D| \frac{h}{|HGV_{A_i}|}}}{|D| \times |N_a|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	600202**
$[25-29]$	600202**
$[25-29]$	600202**
$[25-29]$	600202**

- $|D|=4$
- $h_i=2$
- $|HGV_c|=2$

- $N_a=2$
- $h_c=2$
- $|HGV_i|=2$

Precisão??

Orientada a células – ILoss

- ILoss
  - Captura a fração de nós folhas que são generalizados
  - Quanto menor ILoss, maior a utilidade dos dados
  - $ILoss(V_g) = \frac{|V_g|-1}{|D_a|}$
  - $ILoss(r) = \sum_{V_g \in r} (W_i \times ILoss(V_g))$
  - $ILoss(D) = \frac{\sum_{r \in D} ILoss(r)}{|D|}$
  - V_g: nó na HGD, $|V_g|$: nº folhas na sub-árvore de V_g
  - $|D_A|$: número de valores no domínio (total de folhas)
  - W_i: peso (penalidade definida pelo usuário)
  - D: conjunto de dados

Orientada a células – ILoss

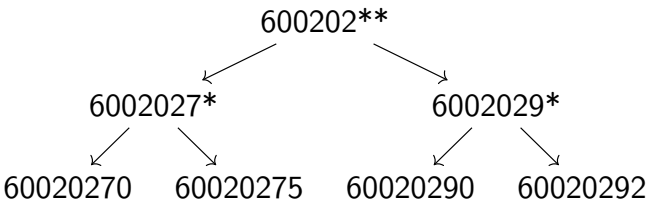


Figura: HGV_c

$$ILoss(V_g) = \frac{|V_g| - 1}{|D_a|}$$

$$ILoss(r) = \sum_{V_g \in r} (W_i \times ILoss(V_g))$$

$$ILoss(D) = \frac{\sum_{r \in D} ILoss(r)}{|D|}$$

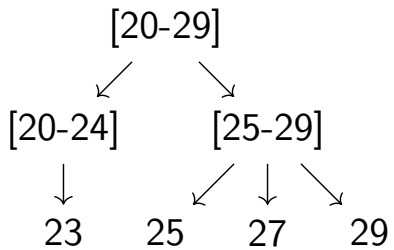


Figura: HGV_i

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
25	60020275
27	60020290
29	60020292

Orientada a células – ILoss

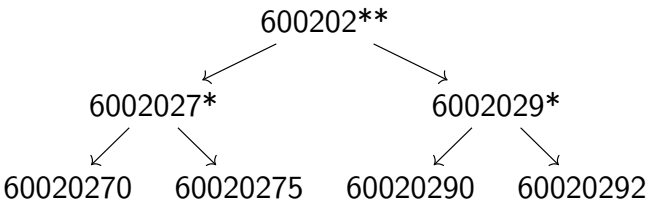


Figura: HGV_c

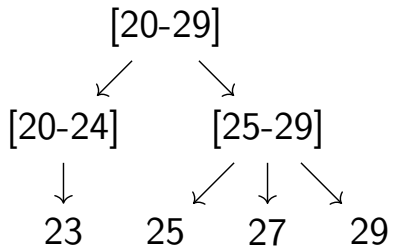


Figura: HGV_i

$$ILoss(V_g) = \frac{|V_g| - 1}{|D_a|}$$

$$ILoss(r) = \sum_{V_g \in r} (W_i \times ILoss(V_g))$$

$$ILoss(D) = \frac{\sum_{r \in D} ILoss(r)}{|D|}$$

Idade	CEP
23	60020270
25	60020275
27	60020290
29	60020292

$\Rightarrow$

Idade	CEP
$[20-24]$	60020270
25	60020275
27	60020290
29	60020292

$|V_g| = 2$

$|D_A| = 7$

$|W_i| = 2$ **ILoss??**

Orientada a registros

Orientada a registros – Classes de Equivalência

- Considere uma série de atributos $A = \{A_1, \ldots, A_n\}$ em D
- Uma classe de equivalência (E) é um conjunto de todos os registros em D que contém valores idênticos para os atributos em A

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*

$$E_1 = \{[20-24], \quad M, 6002027*\}$$
$$E_2 = \{[25-29], \quad F, 6002029*\}$$

Orientada a registros – Tamanho médio das Classes de Equivalência

- Mede o quão bem uma classe de equivalência se aproxima do melhor caso
- Objetivo: reduzir a média normalizada do tamanho das partições

$$C_{avg} = \frac{\left(\frac{totalRegistros}{totalClassesEq}\right)}{k}$$

- k: número mínimo de registros indistinguíveis em uma classe de equivalência

Orientada a registros – Tamanho médio das Classes de Equivalência

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*

$$C_{avg} = \frac{\left(\frac{totalRegistros}{totalClassesEq}\right)}{k}$$

$$C_{avg} = ??$$

Orientada a registros – Tamanho médio das Classes de Equivalência

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*

$$C_{avg} = \frac{\left(\frac{totalRegistros}{totalClassesEq}\right)}{k}$$

$$C_{avg} = ??$$

Orientada a registros – Tamanho médio das Classes de Equivalência

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[30-34]$	F	6002029*
$[30-34]$	F	6002029*

$$C_{avg} = \frac{\left(\frac{totalRegistros}{totalClassesEq}\right)}{k}$$

$$C_{avg} = ??$$

Orientada a registros – Discernibilidade

- Penalidade determinada pelo tamanho da classe de equivalência de um registro r
- Se um registro pertence a uma classe equivalente de tamanho s, a penalidade para o registro é s
- Se uma tupla é suprimida, então é atribuída uma penalidade de valor $|D|$

$$C_{DM} = \sum_{classesEq} |E|^2$$

Orientada a registros – Discernibilidade

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*

$$C_{DM} = ??$$

Orientada a registros – Discernibilidade

Idade	Gênero	CEP
$[20-24]$	M	6002027*
$[20-24]$	M	6002027*
$[25-29]$	F	6002029*
$[25-29]$	F	6002029*
$[30-34]$	F	6002029*
$[30-34]$	F	6002029*

$$C_{DM} = ??$$

Métricas de uso geral

Métricas de uso geral – Altura

- Quantifica a perda de informação como a soma dos níveis de generalização aplicados a todos os valores dos atributos.

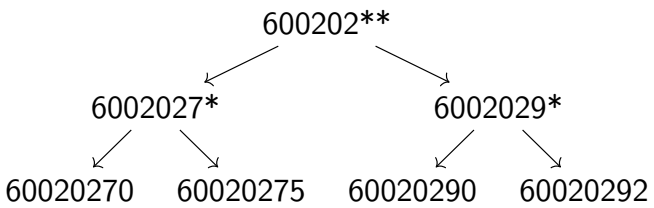


Figura: HGV_c

Idade	CEP
$[20-29]$	600202**
$[20-29]$	600202**
$[20-29]$	600202**
$[20-29]$	600202**

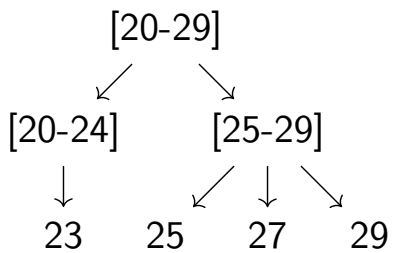


Figura: HGV_i

Métricas de uso geral – Entropia

- Termodinâmica: medida do grau de **irreversibilidade** de um determinado sistema.
- Privacidade: mede a **incerteza** sobre um conjunto de dados.

Métricas de uso geral – Entropia

Idade
23
25
27
29
28
34
32
37
38
39

$\Rightarrow$

Idade
$[21-25]$
$[21-25]$
$[26-30]$
$[26-30]$
$[26-30]$
$[31-35]$
$[31-35]$
$[36-40]$
$[36-40]$
$[36-40]$

$H = \alpha$

$\Rightarrow$

Idade
$[21-30]$
$[21-30]$
$[21-30]$
$[21-30]$
$[21-30]$
$[31-40]$
$[31-40]$
$[31-40]$
$[31-40]$
$[31-40]$

$H = \beta$

$\alpha \leq \beta$