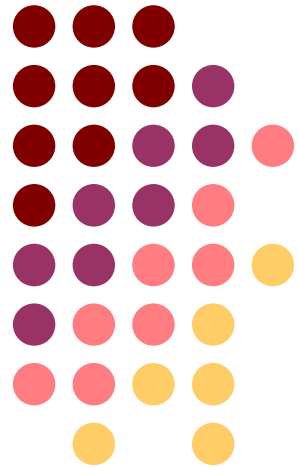
Dados Simbólicos

Universidade Federal de Pernambuco

CIn.ufpe.br



Centro de Informática U·F·P·E

Dados Simbólicos

Dados simbólicos são informações complexas que são expressas por intervalos, conjuntos, frequências distribuições de probabilidade.

Exemplos de situações que necessitam desses tipos de dados

1 - Intervalo para indivíduos (ou objetos de primeira ordem)

Y = tempo gasto com estudos por dia

Y(k) = [0,6] (em horas)

2 - Conjunto para classes de indivíduos (objetos de segunda ordem, objetos agregados)

Y - instituições bancárias existentes em uma cidade

Y(k) = {Bandeirantes, Itau, Bradesco, BB)

onde ω é uma cidade (classe de indivíduos)

3 - Distribuição de frequências - Y = percentual de votos por partido político em um Estado

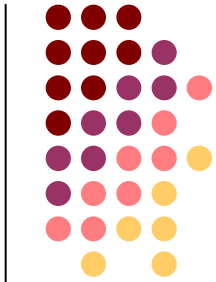
Y(k) = ((ABR;0,50),(PDF;0,30),(DEO;0,20))

4 - Distribuição de probabilidade - Y = níveis de cinza de uma região de uma imagem

Y(k) = Γ(20,30)

Centro de Informática U·F·P·E

Tipos de Variáveis Simbólicas



Notação

Y - uma variável

E - conjunto de entidades

k - uma entidade (ex. uma classe de estudantes, uma empresa)

O - conjunto dos possíveis valores de Y

1 - Variáveis Multi-valorada

Uma variável Y é multi-valores é uma função

$$Y_i : E \rightarrow B = P(O_i)$$
$$y_i(k) \rightarrow U$$

onde $U \subseteq P(O_i)$ conjunto de todos os subconjuntos de O_i. No caso clássico $|y(k)|=1$.

Multi-valor categórico

a) U pode ser um subconjunto categórico.
   Exemplo
   cor (k) = {azul, verde, amarelo}

Centro de Informática U·F·P·E

Multi-valor quantitativo

b) U pode ser um subconjunto finito de números reais.
Exemplo
Y=quantidade de partos realizados em 5 hospitais de uma cidade (k) em uma semana.
y(k)= {100, 200, 150, 220, 180}

Multi-valor ordinal

Se o domínio de *Y* é definido com uma ordenação conforme definido com os dados clássicos.
Na prática a variável tem uma ordenação.

2 - Variáveis Intervalo

Uma variável Y é denominada intervalo se para todo $k \in E$ o subconjunto $y(k) = [a,b]$ onde $a \leq b$.

Exemplo

E= conjunto de empresas
Y= valor gasto com impostos
k= uma empresa
y(k)=[2000,3000]

Dois níveis de paradigmas obtidos por agregação

Exemplo

E={C_1, C_2, ..C_m } conjunto de classes de estudantes

Y = descreve a altura na classe C_i

a) Y - variável multi-valor
$y(C_i)$= {1.50,1.56,1.58,1.70}

b) Y - variável intervalo
$y(C_i)$= [1.50,1.70]

Centro de Informática U·F·P·E

3 - Variáveis Modais

Uma variável modal Y com domínio é uma função

$$Y_i : \quad E \rightarrow B = M(O_i)$$
$$y(k) \rightarrow (U(k), \pi_k)$$

onde

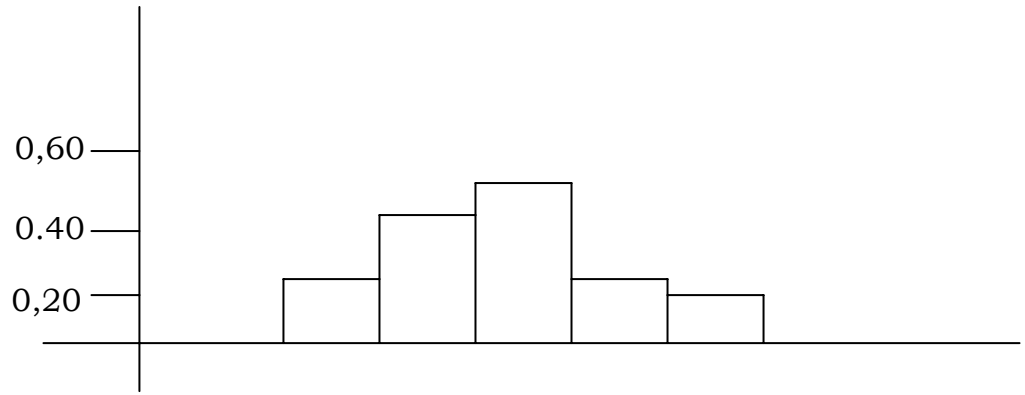
- π_k é uma medida ou uma distribuição (probabilidade, frequência ou peso)
- $U(k) \subset O_i$
- $M(O_i)$ é uma família de medidas π em O_i

Exemplo

C=$\{C_1, \ldots, C_{10}\}$ um conjunto de 10 empresas com 30 funcionários

Y - descreve altura na empresa C_i

1 - Um histograma com os intervalos:



2 - Uma distribuição normal $N(168,48.4)$ com média 168 e variância 48.4

Tabela de dados simbólicos

classe	Marca	Dist. de freq. do tempo de uso	Preco
e_1	{CITR, PEUG}	(A 0.4 B0.3 C0.4)	[12000,30000]
e_2	{BMW,FIAT}	(A 0.2 B0.5 C0.3)	[15000,35000]
e_3	{BMW,FIAT, ALFA}	(A 0.6 B0.2 C0.2)	[8000,45000]
:	:	:	:

Dependências entre Variáveis Simbólicas

Entre variáveis simbólicas são adotados dois tipos de dependências que são expressas através de regras

1 - Dependência Hierárquica

Uma variável Y_i pode tornar-se inaplicável se outra variável Y_j assume valores em um subconjunto $S_j \subset O_j$.

A dependência é expressa pelas regras

$$r: [[y_j \in S_j]] \overset{DH}{\rightarrow} [y_i = NA]] \quad \text{(equivalência lógica)}$$

Centro de Informática U·F·P·E

ou equivalentemente pelas regras

r_1: $[[y_j \in S_j]] \rightarrow [y_i = NA]]$
r_2: $[y_i = NA]] \rightarrow [[y_j \in S_j]]$

Exemplo

Y_1 = Sexo={M,F} e Y_2= Parto={sim,não}

A dependência entre Y_1 e Y_2 é expressa pelas regras
$r_{1:}$ se $[y_1 = M]$ **então** $[y_2 = NA]$
r_2: se $[y_2 = NA]$ **então** $[y_1 = M]$

2 - Dependência lógica

Um subconjunto $S_i \subset O_i$ da variável Y_i pode estar em correspondência com o subconjunto $S_j \subset O_j$ da variável Y_j.

A dependência é expressa pela regra

$$r: [[y_j \in S_j]] \overset{DL}{\rightarrow} [y_j \in S_j]]$$

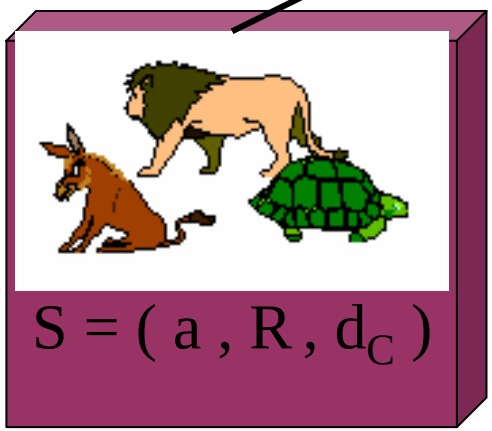
Exemplo

Y_1 = idade[0,90] e Y_2= altura (em cm)[0,4;2,00]

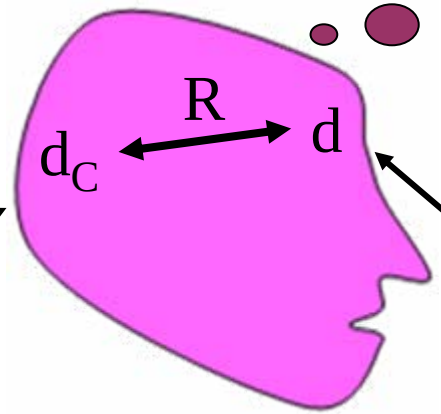
A dependência entre Y_1 e Y_2 é expressa pela regra
r: se $[y_1 =$ [0,10] **então** $[y_2 =$ [0,40;1,60]

Centro de Informática U·F·P·E

Objetos Simbólicos



$S = (a , R , d_C)$



y

k

É um animal(w) = 99% sim

$a(w) = [y(k)\ R\ d_C]$

Centro de Informática U·F·P·E

Objetos Simbólicos Booleanos

Definição

Um objeto simbólico booleano é uma tripla s = (a,R,d) onde R é uma relação entre descrições, $d \in D$ uma descrição e a uma função binária $a_s : E \rightarrow \{0,1\}$ com

$$a_s(k) = \bigwedge_{j=1}^{p} [y_j(k)\, R\, d] = 1 \quad \text{se e somente se} \quad [y_j(k)\, R\, d] = 1$$

$$\forall\ j \in \{1,\ldots,p\}$$

Centro de Informática U·F·P·E

A **extensão** de s é definida como

$$Ext(s) = \{k \in E / a_s(k) = 1\}$$

Exemplo

Considere uma tabela de dados clássicos e s=(a,R,d) um Objeto simbólico onde y (k)= [altura(k),peso(k)], d =[[140,160],[50,60]] e R = $\in$ então

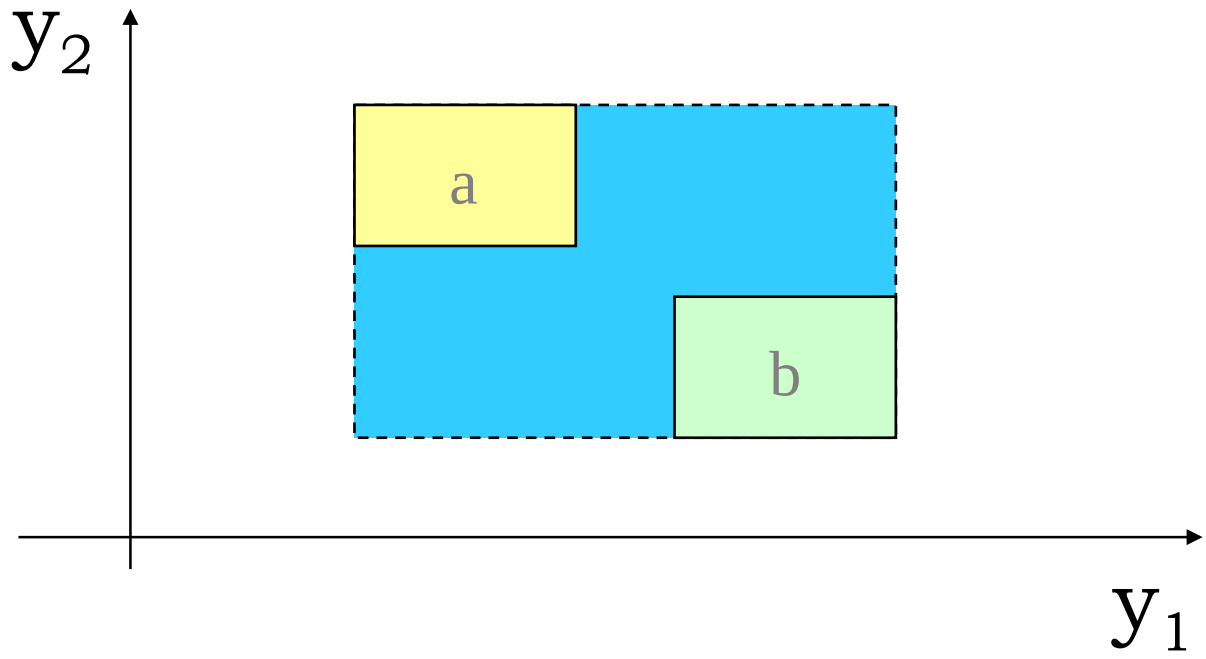
$$a(k) = [altura(k) \in [165,175]] \wedge [peso(k) \in [50,60]]$$

$$Ext(s) = \{\omega_1, \omega_3\}$$

Centro de Informática U·F·P·E

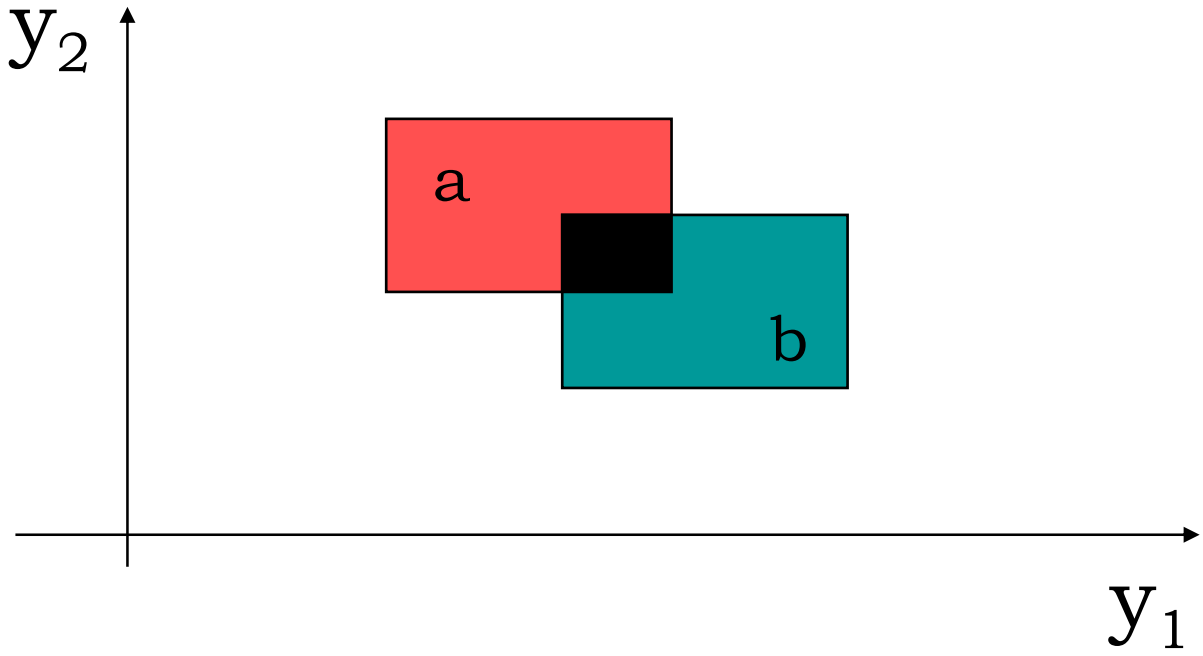
Operadores Simbólicos

Junção



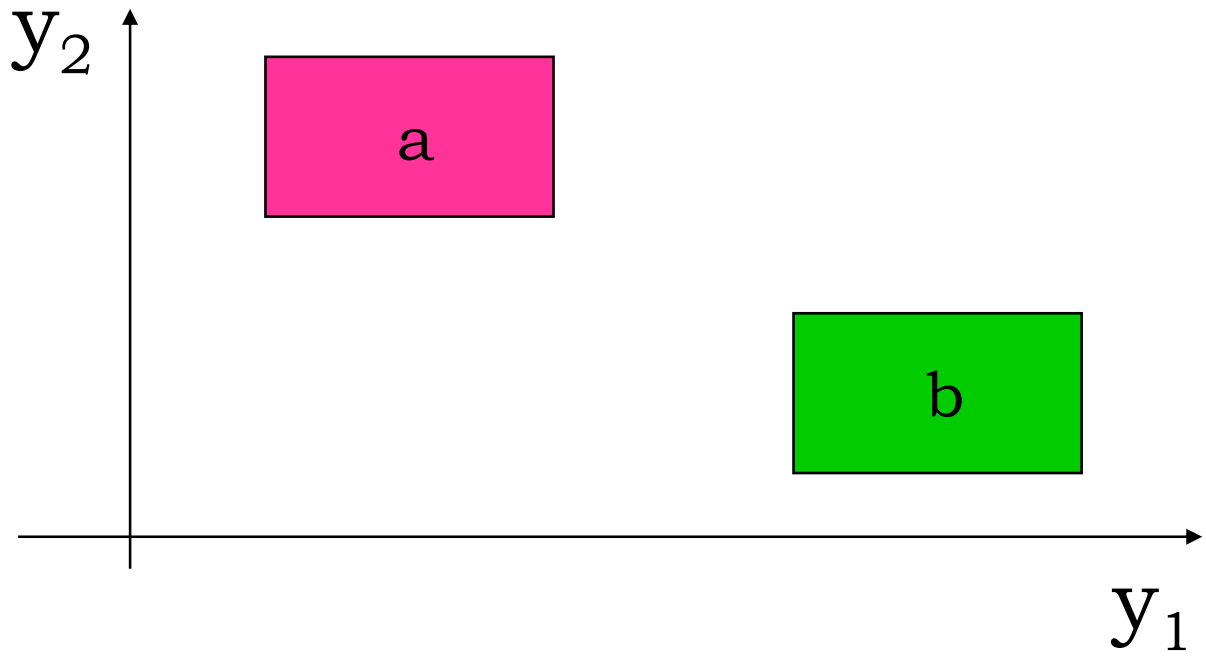
$$a \oplus b = [y_1 \in C_1] \wedge [y_2 \in C_2]$$

Conjunção



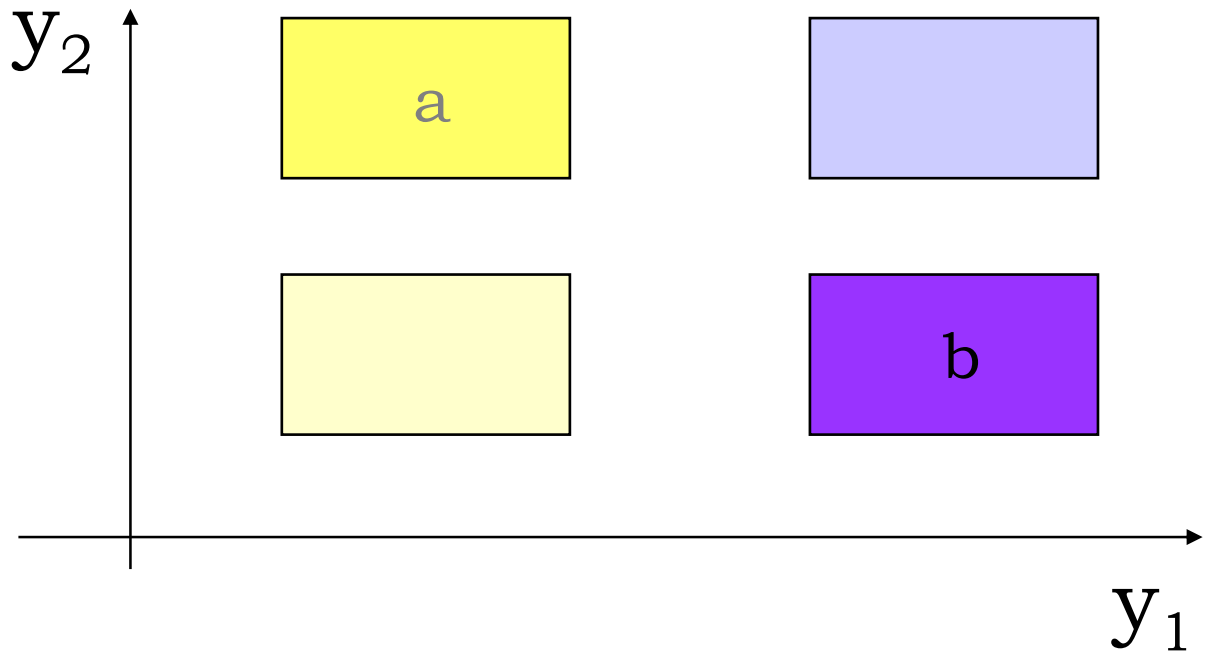
$$a \wedge b = [y_1 \in (A_1 \cap B_1)] \wedge [y_2 \in (A_2 \cap B_2)]$$

Disjunção



$$a \wedge b = \{[y_1 \in A_1] \wedge [y_2 \in A_2]\} \vee \{[y_1 \in B_1] \wedge [y_2 \in B_2]\}$$

União



$$a \cup b = [y_1 \in A_1 \cup B_1] \wedge [y_2 \in A_2 \cup B_2]$$

Potencial de descrição

O potencial de descrição é uma medida positiva definida no objeto que representa o volume (De Carvalho,1995) .

$$\pi(s) = \prod_{j=1}^{p} \mu(d_j)$$

onde $\mu(d_i)$= cardinal de d_i, se d_i é um conjunto
amplitude de d_i, se d_i é um intervalo

Exemplo

Seja o objeto s apresentado anteriormente, então o potencial de s é $\pi(s)$ = (175-165) x (60-50) = 100

Centro de Informática U·F·P·E

Objetos Simbólicos Modais

Definição

Um objeto simbólico modal é uma tripla $s = (a,\Phi,d)$ onde Φ é uma relação de fuzzy e a uma função $a_s : E \rightarrow [0,1]$.

A **extensão** de s é definida como

$$Ext(s) = \{k \in E / a_s(k) > \alpha\}$$

Centro de Informática U·F·P·E

Exemplo

Seja E um conjunto de espécies de animais, k uma classe de animais Y_1 = cor e Y_2 = mês de nascimento. Então O_1={branco,cinza, marrom, preto, azul,...} e O_2={jan,...,dez}

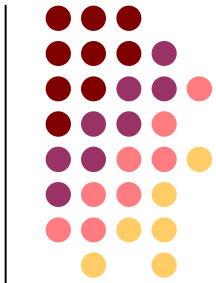
Um objeto s = (a,Φ,d) onde d =[{preto,marrom},{maio, junho}] e Φ é uma relação de fuzzy definida como

$$\Phi(y(k),d) = \frac{|y(k) \cap d|}{|y(k) \cup d|}$$

Então para y(k)=[{branco,marrom},{jun, jul}] Φ(y(k),d)=1/7.

Adotando α=0.5 logo a(k)=0.13 < 0.5 logo k $\not\subset$ Ext(s)

Construindo descrições: O processo de generalização



- Sendo Dc = conjunto das descrições dos indivíduos de C
- Aplicação dos operadores
  - O intervalo G_y (C) = [inf(D_c), sup (D_c)] constitui uma boa generalização de C
- Seja
  - C = {w1, w2, w3} e
  - D_c = {y(w1),y(w2),y(w3)} = y (C)
  - A generalização de C para a variável y é G_y(C)
- Exemplos
  - Seja y uma variável numérica tal que:
    - Y(w1) = 2.5, y(w2)=3.6, y(w3)=7.1
  - Seja D o conjunto de valores incluídos em [1,100]
    - → Gy(C) = [2.5, 7.1] é a generalização de D_c para y

O processo de generalização

- Exemplos
  - Seja y uma variável modal:
    - y = {pequeno, grande} tal que:
    - y(w1) = (1(1/3) , 2(2/3))
      - em que 2(2/3) significa que a freqüência da categoria 2 é 2/3
    - y(w2) = (1(1/2) , 2(1/2))
    - y(w3) = (1(1/4) , 2(3/4))
  - Então Gy(C) = [[[1(1/4)],[1(1/2)]],[[2(1/2)], [2(3/4)]]] é a generalização de D_c para y
    - Isto significa que da categoria 1 o intervalo é [1/4,1/2]
    - Isto significa que da categoria 2 o intervalo é [1/2,3/4]

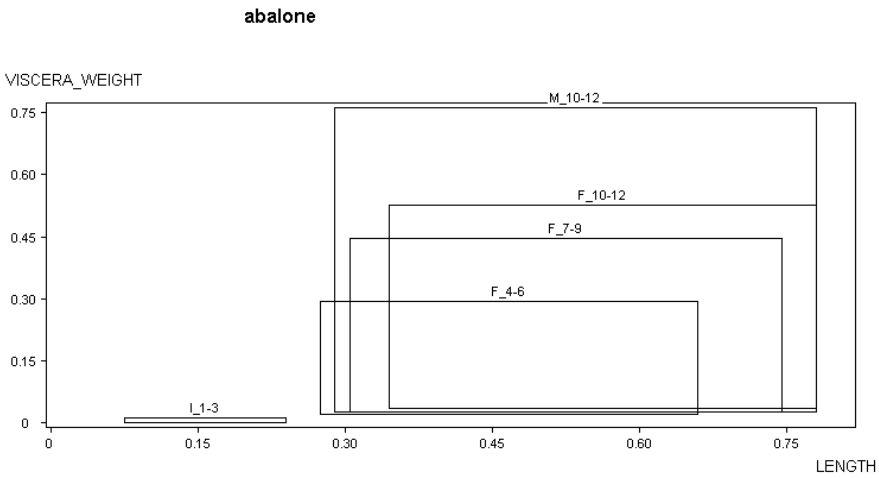
O processo de generalização

- Exemplos
  - Seja y uma variável intervalar tal que:
    - y(w1) = [1.5 , 3.2]
    - y(w2) = [3.6 , 4]
    - y(w3) = [7.1 , 8.4]
  - Então Gy(C) = [1.5, 8.4] é a generalização de D_c para a variável y
  - Seja C = {w1, w2, w3} e y é uma variável categórica não ordenada
  - y(w1) = 2, y(w2) = 2, y(w3) = 1
  - D é o conjunto de ocorrência dos valores 1 e 2
  - G´(y(C)) = [1(1/3), 2(2/3)]
    - → Dc = {y(w1), y(w2), y(w3)} = y(C) para a variável y.

Centro de Informática U·F·P·E

O processo de generalização

- Uso dos operadores
  - Simples
  - Pode gerar outliers
- Estratégia
  - Reduzir os limites para reduzir ocorrência de *outliers*
    - *Teste komogorov Sminorv (distribuição uniforme)*
    - *Algoritmo de agrupamentro*
  - Utilizar freqüências para o caso da variáveis categóricas



Centro de Informática U·F·P·E

O processo de generalização

- O processo se preocupa com pelo menos 2 questões
  - Overgeneralization
    - Ex. age(alunos_renata) = [18,36]
  - Perda de informação estatística, como média, variância, correlações entre variáveis
    - Solução → criação de novas variáveis simbólicas contendo os respectivos valores

Um critério de avaliação da descrição generalizada

Pode-se avaliar a qualidade de uma generalização g_i associada a uma descrição d_i da classe C_i através da medida

$$dens(s_i) = \frac{|Ext(s_i / C_i|}{\pi(d_i)}$$

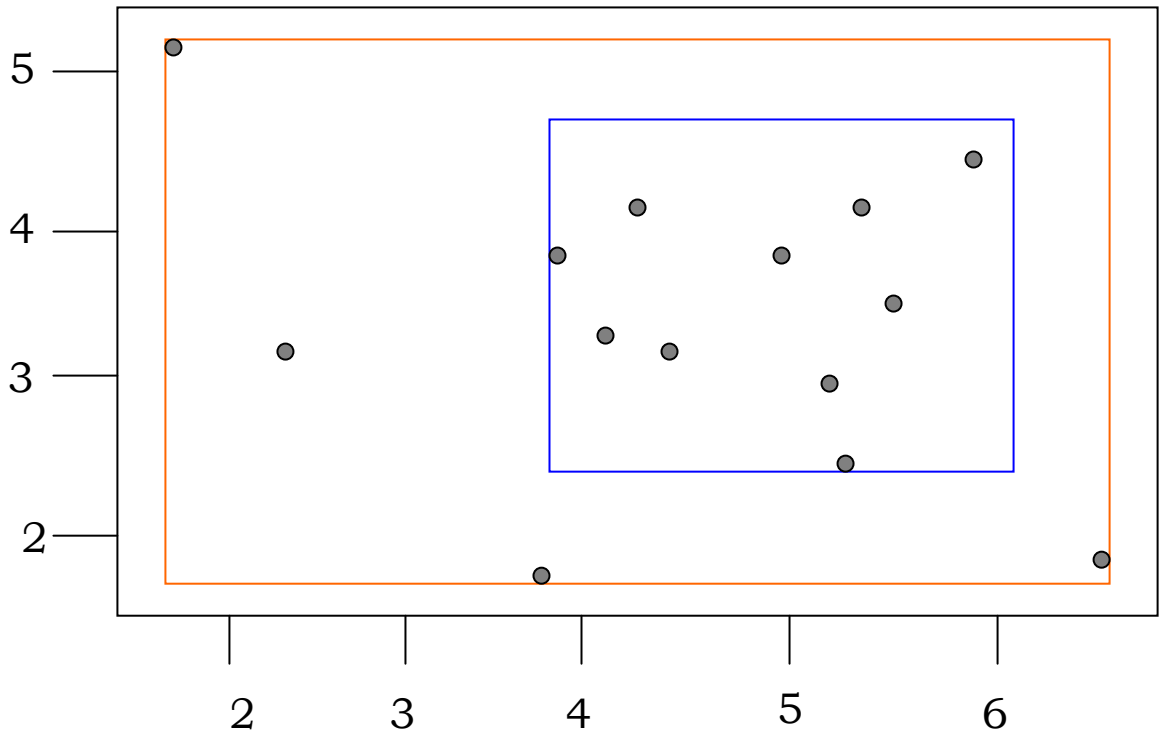
denominada de **densidade do objeto simbólico s_i**

$Ext(s_i/C_i)$ define o conjunto de indivíduos que satisfazem a descrição do objeto s_i

$\pi(d_i)$ é o volume que produz o índice de generalização de d_i

Centro de Informática U·F·P·E

Problema de sobre-generalização



$a_1 = [y_1 \in [1.5, 6.5]] \wedge [y_2 \in [1.5, 5.5]]$ **descrição inicial**

especialização

$a_1 = [y_1 \in [3.8, 6.0]] \wedge [y_2 \in [2.5, 4.5]]$ **descrição final**

Centro de Informática U·F·P·E

Etapa de Especialização

A idéia é encontrar um limiar α-generalização de C_i, que corresponda o melhor trade-off entre a redução do volume do objeto s_i e o conjunto que satisfaz a descrição desse objeto.

Um α-generalização de C_i é um objeto simbólico

$$s_i^{\alpha} = (a_i^{\alpha}, \in, d_i^{\alpha})$$

tal que

- $\left|Ext(s_i^{\alpha} / C_i\right| \geq \alpha \times \left|C_i\right|$
- d_i^{α} tal que

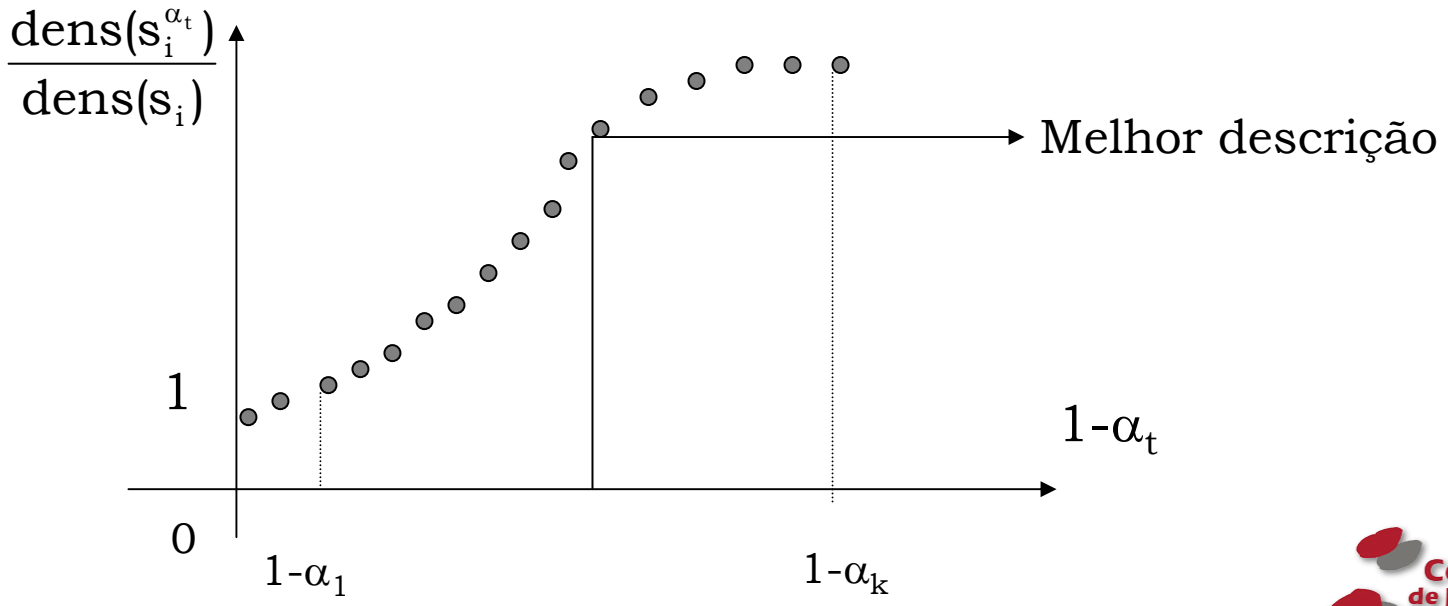
$$vol(d_i^{\alpha}) = \min_{d \in P(O_1)x...xP(O_p)} \left\{ vol(d) / \left| Ext(d / C_i) \right| \geq \alpha \times \left| C_i \right| \right\}$$

Para cada seqüência de $\alpha_1 > \alpha_2 > \ldots > \alpha_t > \ldots > \alpha_k > 0$, pode-se definir o

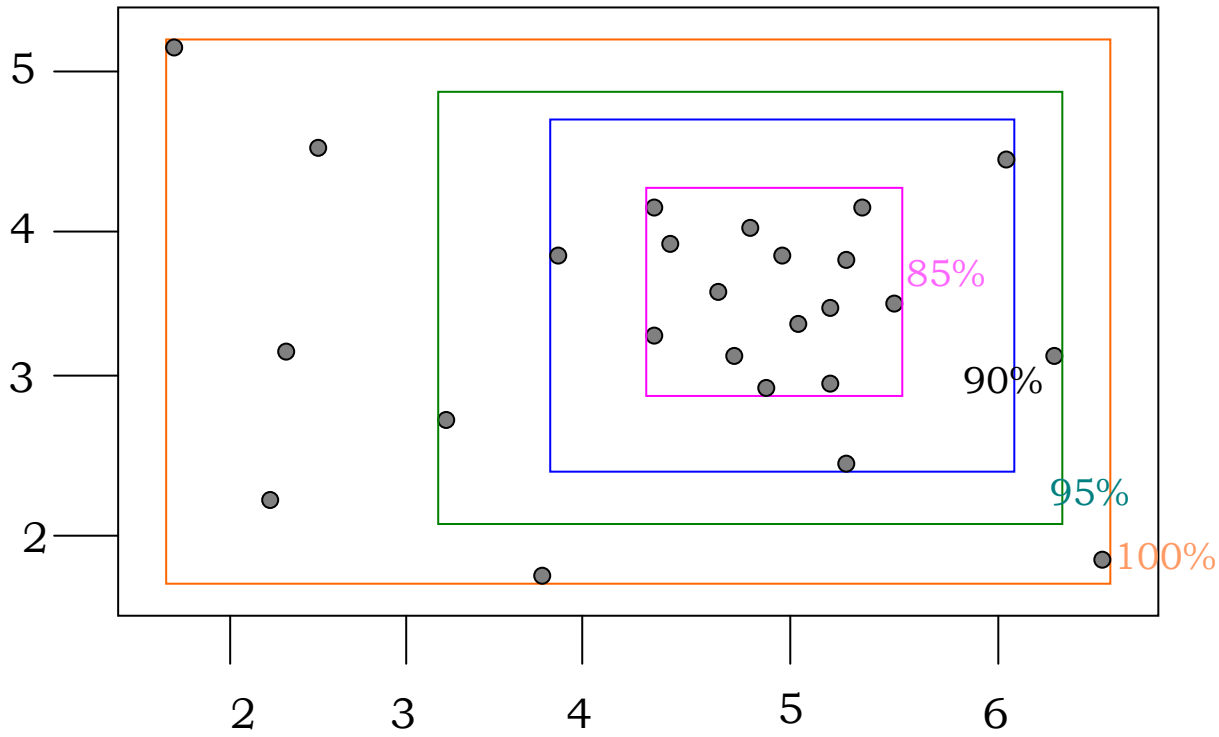
α_t – generalização $s_i^{\alpha_t}$

Cada $s_i^{\alpha_t}$ é avaliado conforme a densidade relativa $\frac{dens(s_i^{\alpha_t})}{dens(s_i)}$

onde s_i corresponde ao objeto completo



Exemplo do resultado do algoritmo de redução (Stephan, 2000)



$a_1 = [y_1 \in [1.5, 6.5]] \wedge [y_2 \in [1.5, 5.5]]$ **descrição inicial**

$a_1 = [y_1 \in [4.5, 5.5]] \wedge [y_2 \in [2.8, 4.3]]$ **descrição final α = 85%**

Centro de Informática U·F·P·E