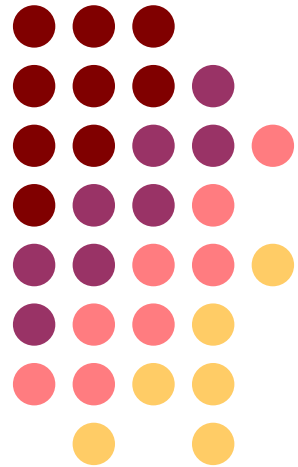
Symbolic Data Analysis

Universidade Federal de Pernambuco

CIn.ufpe.br



Centro de Informática U·F·P·E

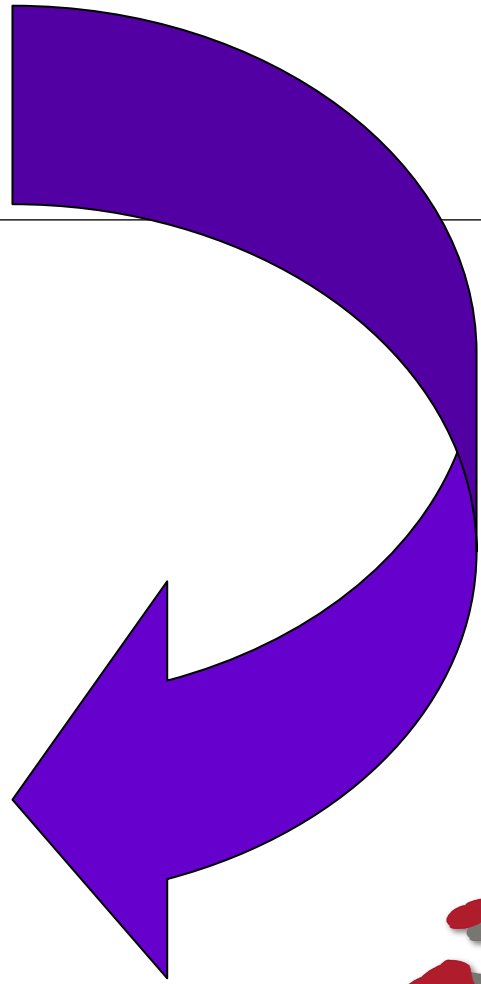
MINERAÇÃO DE DADOS

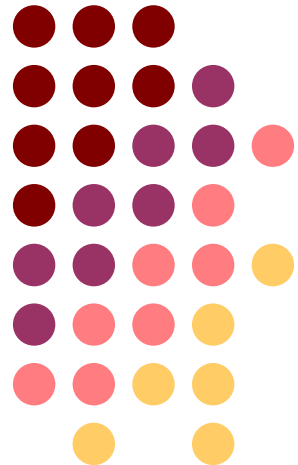
Data Minig

Análise de Dados Simbólicos

MINERAÇÃO DE CONHECIMENTOS

Knowledge Mining





Centro de Informática U·F·P·E

Análise de Dados Simbólicos

- Surgiu em 1988 (E. Diday) e recebe influência de três grandes áreas: Análise Exploratória de Dados, Inteligência Artificial e Taxonomia Numérica.
- Nova abordagem na área de Descoberta de Conhecimento (KDD) que visa estender as técnicas estatísticas e os métodos da análise exploratória de dados para dados mais complexos chamados de Dados Simbólicos.

Análise de Dados Simbólicos

- Novas estruturas de dados
  - Células multivaloradas;
  - Intervalos numéricos;
  - Distribuições empíricas ou de probabilidade.
- Soda's Project – Software p/ análise de dados simbólicos
  - http:/ /www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm
- Livros:
  - Analysis of Symbolic Data, H.-H Bock and E. Diday, Springer-Verlag, 2000
  - Symbolic Data Analysis: Conceptual Statistics and Data Mining.y, L. Billard and E. Diday, John Wiley, 2007
  - Symbolic Data Analysis and The SODAS Software E. Diday and Monique Noirhomme-Fraiture, John Wiley, 2008

Centro de Informática U·F·P·E

Tabela de Dados Simbólicos

Produto	*Preço*	*Cidade*	*Cor*
P1	[5; 9]	Londres	{(0.1) R ; (0.3) G ; (0.6) B }
P2	[12 ; 15]	{ Paris ; Londres }	{(0.4) Y, (0.6) G}
P3	[3 ; 9]	{Bruxelas, Paris}	{ (0.3) W ; (0.7) B }
P4	[1 ; 8]	{Lisboa, Madri}	{ (0.5) W ; (0.5) B }

- **Cidade = Variável Multivalorada**
- **Cor= Variável Multivalorada Ponderada (Modal)**
- **Preço = Variável Intervalar**

Centro de Informática U·F·P·E

Tabela de Dados do tipo Intervalo

Temperaturas Mensais, Mínimio e Máximo, registradas em 60 estações metereológicas da China

Estações	Temperatura mensais ([min : max]) – ano 1988				
	Januay	February		November	December
AnQing	[1.8,7.1]	[2.1,7.2]		[7.8,17.9]	[4.3,11.8]
	...	...	...	...	...
ZhoJiang	[2.7,8.4]	[2.7,8.7]		[8.2;20]	[5.1,13.3]

Alguns Métodos para Dados Simbólicos

- Análise Componente Principal e Análise Fatorial.
- Estatísticas Descritivas.
- Análise de Cluster (classificação não supervisionada.
- Análise Discriminante (classificação supervisionada).
- Modelos de Regressão
- Redes Neurais MLP

A Idéia Básica

- **Unidades de primeira ordem** (como um cavalo ou uma pessoa qualquer), cada qual correspondendo a um único indivíduo do mundo;
- **Unidades de segunda ordem** (como o cavalo ou a pessoa, de forma geral), correspondendo a uma classe de indivíduos do mundo.

Exemplo de Indivíduos de Primeira Ordem

Indivíduo	Classes	Variáveis Descritivas das Unidades		
ID	*Região*	*Qtd de Camas*	*Qtd de Salas de Jantar*	*Classe Social*
1	**Norte**	**2**	**1**	**1**
2	**Norte**	**2**	**1**	**3**
3	**Norte**	**1**	**3**	**3**
4	**Leste**	**1**	**3**	**3**
5	**Leste**	**2**	**2**	**1**
6	**Leste**	**1**	**2**	**3**

Centro de Informática U·F·P·E

Obtendo os Indivíduos de Segunda Ordem

Classes	Variáveis Descritivas das Unidades		
Regiões	*Qtd de Camas*	*Qtd de Salas de Jantar*	*Classe Social*
Norte	2	1	1
Norte	2	1	3
Norte	1	3	3
Leste	1	3	3
Leste	2	2	1
Leste	1	2	3

Classes	Variáveis Descritivas das Unidades		
Regiões	*Qtd de Camas*	*Qtd de Salas de Jantar*	*Classe Social*
Norte	(1/3) 1, (2/3) 2	(2/3) 1, (1/3) 3	(1/3) 1, (2/3) 3
Leste	(2/3) 1, (1/3) 2	(2/3) 2, (1/3) 3	(1/3) 1, (2/3) 3

Centro de Informática U·F·P·E

Exemplo de Indivíduos de Primeira Ordem – outro exemplo

- Em uma ilha há 600 pássaros juntos: 400 Swallows, 100 Ostriches e 100 Penguins

Birds	*Species*	*Flying*	*Size*
1	**Penguin**	**No**	**80**
.	**...**	**...**	**...**
.	**...**	**...**	**...**
.	**...**	**...**	**...**
599	**Swallow**	**Yes**	**70**
600	**Ostrich**	**No**	**125**

Exemplo de Indivíduos de Primeira Ordem – outro exemplo

- Swallow bird

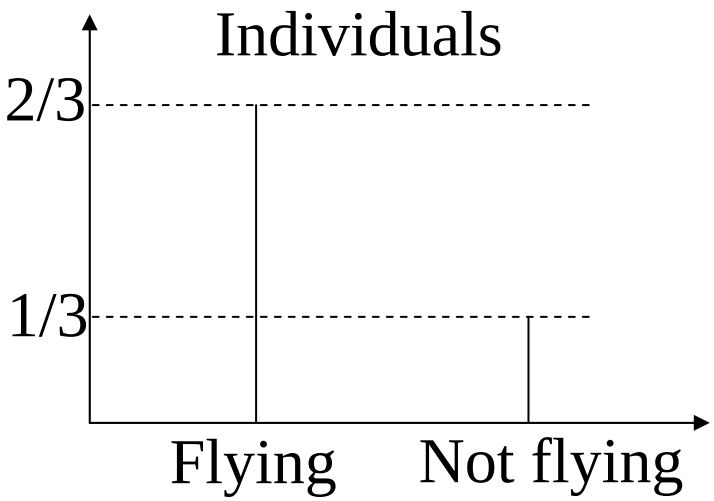
- Ostrich

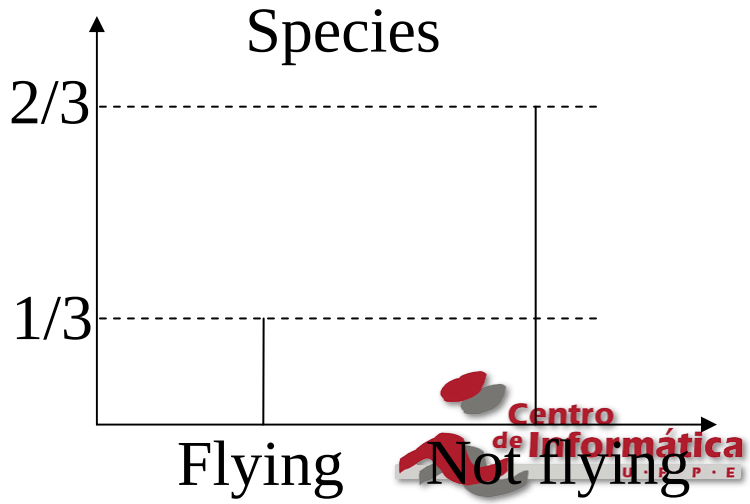
- Penguin

Centro de Informática U·F·P·E

Obtendo os Indivíduos de Segunda Ordem

Species	*Flying*	*Size*	*Migration*
Swallow	**{Yes}**	**[60,85]**	**{ (0.1) N ; (0.9) Y }**
Ostrich	**{No}**	**[85,160]**	**{ (1.0) N ; (0.0) Y }**
Penguin	**{No}**	**[70,95]**	**{ (0.0) N ; (1.0) Y }**





Base de dados descrevendo jogadores de futebol

Player	*Team*	*Age*	*Weight*	*Height*	*Nationality*
Fernandes	**Spain**	**29**	**85**	**1.84**	**Spanish**
Rodrígues	**Spain**	**23**	**90**	**1.92**	**Brazilian**
Mballo	**France**	**25**	**82**	**1.9**	**Senegalese**
Zidane	**France**	**27**	**78**	**1.85**	**French**

Team	*AGE*	*Weigh*	*Height*	*Nationality*	*Number of goals at the wolrd cup 1998*
Spain	**[23,29]**	**[85,90]**	**[1.84,1.92]**	**(0.5 Sp, 0.5 Br)**	**18**
France	**[21,28]**	**[85,90]**	**[1.84,1.92]**	**(0.5 Fr, 0.5 Se)**	**24**

Centro de Informática U·F·P·E

Tabela de Dados Simbólicos

- As células podem conter dados complexos
  - Valores numéricos
    - Ex. height (Tom) = 1.80
  - Intervalos
    - Ex. age (Spain) = [23,29]
  - Categóricos
    - Ex. Nationality (Deco) = {brasileira}
    - Ex. Nationality (Spain) = {brasilian, spanish, french}
  - Modal
    - Ex. Nationality (Spain) = {(0.1)brasilian, (0.8)spanish, (0.1)french}

Mais um exemplo de Tabela de Dados Simbólicos

Produto	*Altura*	*Cidade*	*Cor*
P1	3.5	Londres	{ R , G , B }
P2	[3 , 8]	{ Paris , Londres }	
P3	{ P , M , G , GG }		{ (0.3) W , (0.7) B }
P4	[(1/3) [2,3] , (2/3) [4,5]]		

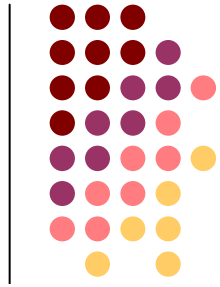
- Outros tipos de dados simbólicos
  - Taxonomia (induced rules)
  - Dependência hierárquica (mother-daughter variable)
  - Dependência lógica

cor = “fria” ← “violeta”, “azul” e “verde”

“flying” (mother variable), “speed of flying” (daughter variable), ”

If age(w) <= 2 months → (height < 80 cm)

Fontes de Dados Simbólicos



- A partir de variáveis categóricas
  - Como tipo do empregado
  - Obtido por clusterização de grandes massas de dados
- De bancos de dados
  - Consultas originando novas variáveis (*queries*)
- Do conhecimento do especialista
- De dados confidenciais
  - Para esconder informações privadas.
- De dados estocásticos
  - Distribuição de probabilidade
- De séries temporais
  - Descrevendo intervalos de tempo

O processo de generalização

- “O processo de generalização é aplicado a um conjunto de indivíduos para produzir uma descrição simbólica”
- Exemplo para descrição da espécie “Swallow”:
  - d = ({yes}, [60,85],[90% yes, 10% no])
- Dados simbólicos
  - Cor das espécies na tabela de dados simbólicos
  - Swallows → [branco, preto]
  - Penguins → {branco, preto}

Generalizando dados simbólicos a partir de dados fuzzy, imprecisos ou conjuntivos

- Dados fuzzy
  - Ex. variável numérica *altura (homem)* = 1,60 m
  - Pode ser associada ao valor
    - “pequeno” com peso 0,9
    - “médio” com peso 0,1
    - “alto” com peso 0,0
- Dados imprecisos
  - Ocorrem quando não é possível obter uma medida exata
  - Ex: é possível dizer que uma árvore tem 10m ± 1
  - Significa que a altura da árvore está no intervalo [9,11]
- Dados conjuntivos
  - Ocorrem quando muitas categorias aparecem simultaneamente.
  - Ex. uma maçã pode ser vermelha e verde, ou amarela, as três cores ao mesmo tempo

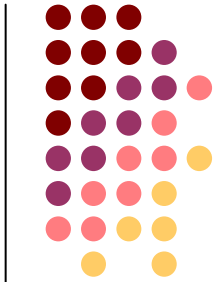
Centro de Informática U·F·P·E

Incerteza e dados simbólicos

- Ex1. Vamos considerar que não sabemos a altura do tenista Rafael Nadal;
- Considerando que ao jogar com o tenista Roger Federer, o Nadal parece ser apenas um pouco mais baixo, e sabendo que Federer mede 1.85m.
- É plausível considerar que
  - Altura (Nadal) = [1.80, 1.85] (incerteza)
- Ex2. Considerando a altura dos dez melhores tenistas ranqueados da ATP, podemos considerar a altura como dado simbólico:
  - Altura (top 10) = [1.80, 1.85] (variabilidade)
- Os dois intervalos são iguais mas representam semânticas completamente diferentes

Centro de Informática U·F·P·E

Dados simbólicos a partir de dados estruturados



- Dados estruturados ocorrem quando há variáveis do tipo mother/daughter ou taxonômicas ou ainda em tabelas associadas em um SGDB

- É possível unir tabelas com poucas variáveis em comum usando dados simbólicos

Centro de Informática U·F·P·E

Dados simbólicos a partir de dados estruturados

Escola	*Cidade*	*Nr alunos*	*Tipo*	*Nível*
Jaurès	Paris	320	Public	1
Condorcet	Paris	450	Public	3
Chevreul	Lyon	200	Public	2
St Helene	Lyon	380	Private	3
St Sernin	Toulouse	290	Public	1
St Hilare	Toulouse	210	Private	2

Descrição clássica de Escolas

Cidade	*Nr Alunos*	*Tipo*	*Nível*
Paris	[320,450]	(100) public	{1,3}
Lyon	[200,380]	(50%) public, (50%) private	{2,3}
Toulouse	[210,290]	(50%) public, (50%) private	{1,2}

Descrição simbólica de cidades pela variável escola - generalização

Centro de Informática U·F·P·E

Dados simbólicos a partir de dados estruturados

Hospital	*Cidade*	*Nr Leitos*	*Código da especialidade*
Lariboisiere	Paris	750	5
St Louis	Paris	1200	3
Herriot	Lyon	650	3
Besgenettes	Lyon	720	2
Purpan	Toulouse	520	6
Marchant	Toulouse	450	2

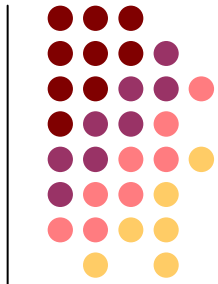
Descrição clássica de Hospitais

Cidade	*Nr Leitos*	*Código da especialidade*
Paris	[750,1200]	{3,5}
Lyon	[650,720]	{2,3}
Toulouse	[450,520]	{2,6}

Descrição simbólica de cidades pela variável hospital - generalização

Centro de Informática U·F·P·E

Dados simbólicos a partir de dados estruturados



Cidade	*Nr Alunos*	*Tipo*	*Nível*
Paris	[320,450]	(100) public	{1,3}
Lyon	[200,380]	(50%) public, (50%) private	{2,3}
Toulouse	[210,290]	(50%) public, (50%) private	{1,2}

Cidade	*Nr Leitos*	*Código da especialidade*
Paris	[750,1200]	{3,5}
Lyon	[650,720]	{2,3}
Toulouse	[450,520]	{2,6}

Cidade	*Nr alunos*	*Tipo*	*Nível*	*Nr Leitos*	*Código da especialidade*
Paris	[320,450]	(100) public	{1,3}	[750,1200]	{3,5}
Lyon	[200,380]	(50%) public, (50%) private	{2,3}	[650,720]	{2,3}
Toulouse	[210,290]	(50%) public, (50%) private	{1,2}	[450,520]	{2,6}

Centro de Informática U·F·P·E

Os quatro tipos de estatísticas e "*data mining*"

Análise Clássica **Análise Simbólica**

Dados Clássicos

Caso 1

Caso 2

Cidade	*Min. Alunos*	*Máx. alunos*	*Public*	*Private*	*Nível 1*	*Nível 2*	*Nível 3*
Paris	**320**	**450**	**100**	**0**	**1**	**0**	**1**
Lyon	**200**	**380**	**50**	**50**	**0**	**1**	**1**
Toulouse	**210**	**290**	**50**	**50**	**1**	**1**	**0**

Análise de dados simbólicos sobre dados simbólicos.

P… d… de tabelas de dados clássicos

…ormação de dados simbólicos …blicação de estatística clássica. Muita informação é perdida

Centro de Informática U·F·P·E

SODAS – Symbolic Official Data Analysis System

- Protótipo disponível gratuitamente
  - http://www.info.fundp.ac.be/asso/
  - Funcionalidades
  - Construção de tabelas de dados simbólicos a partir de BD's tradicionais
  - Descrição de regras e hierarquias
  - Análise dos dados através de métodos de análise de dados simbólicos
    - Estatística descritiva
    - Análise Fatorial
    - Agrupamento
    - Árvore de Decisão
    - ...

Centro de Informática U·F·P·E

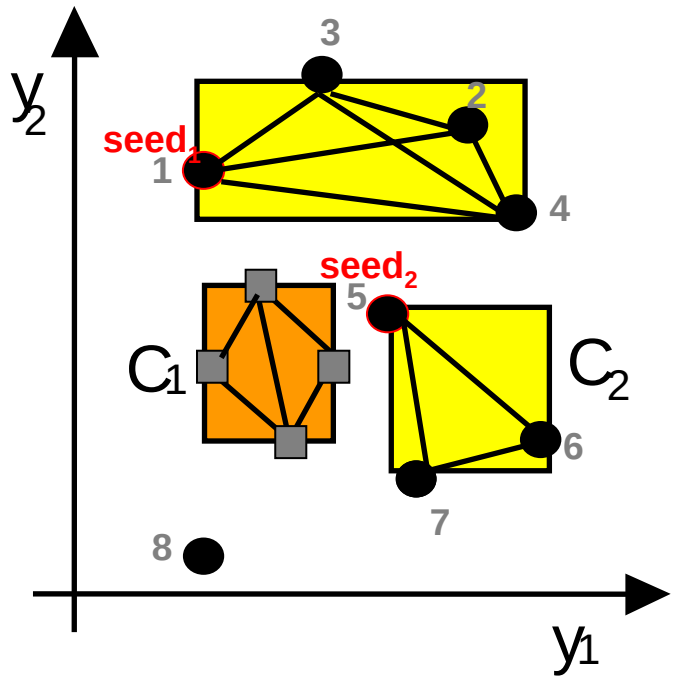
CLASSIFICADOR SIMBÓLICO APLICADO A IMAGENS SAR

Etapa de Aprendizagem

Entrada: Dados usuais

segmento	níveis de cinza	
	media (y_1)	desvio padrão (y_2)
seg_1	44.50	12.50
seg_2	83.60	3.60
seg_3	120.30	6.45
•	•	•
•	•	•

Aproximação do Grafo de Vizinhos Mútuos



Saída: Dados Simbólicos que descrevem as regiões (grupos de segmentos)

Grupo	Sub-Grupo	níveis de cinza	
		media (y_1)	desvio padrão (y_2)
Região 1	G_{11}	[37.35,57.70]	[0,20,5.62]
Região 2	G_{21}	[132.56,160.79]	[0.73,6.84]
	G_{22}	[167.12,196.67]	[1.30,10.66]
•	•	•	•
•	•	•	•

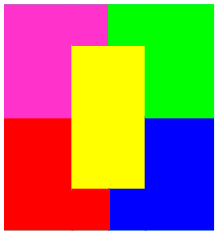
Etapa de Alocação: Funções de Proximidade

1 - Distância (De Carvalho et al (1998)) $$d_r(s,s') = \frac{1}{p}\left[\sum_{i=1}^{p} \left\{\Phi_{jk\gamma}\left(s_i, s_i^{'}\right)\right\}^r\right]^{1/r}$$

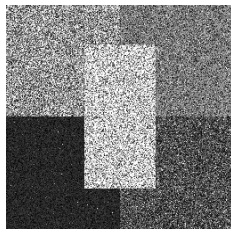
2 - Palumbo et al (1996) $$q_{ss'} = \frac{\pi(s \oplus s') - \pi(s)}{\pi(s)}, \pi(s) \neq 0$$

Centro de Informática U·F·P·E

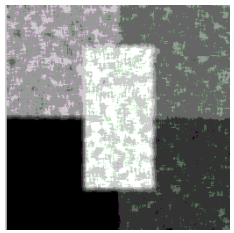
EXPERIMENTO MONTE CARLO



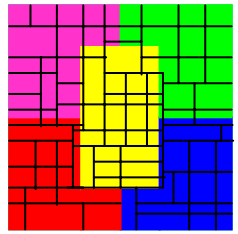
PHANTOM



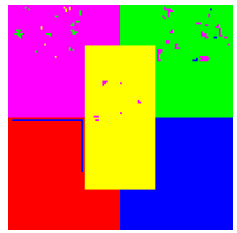
MULTIPLICATIVE MODEL
(Frery et al 1997))



LEE FILTER



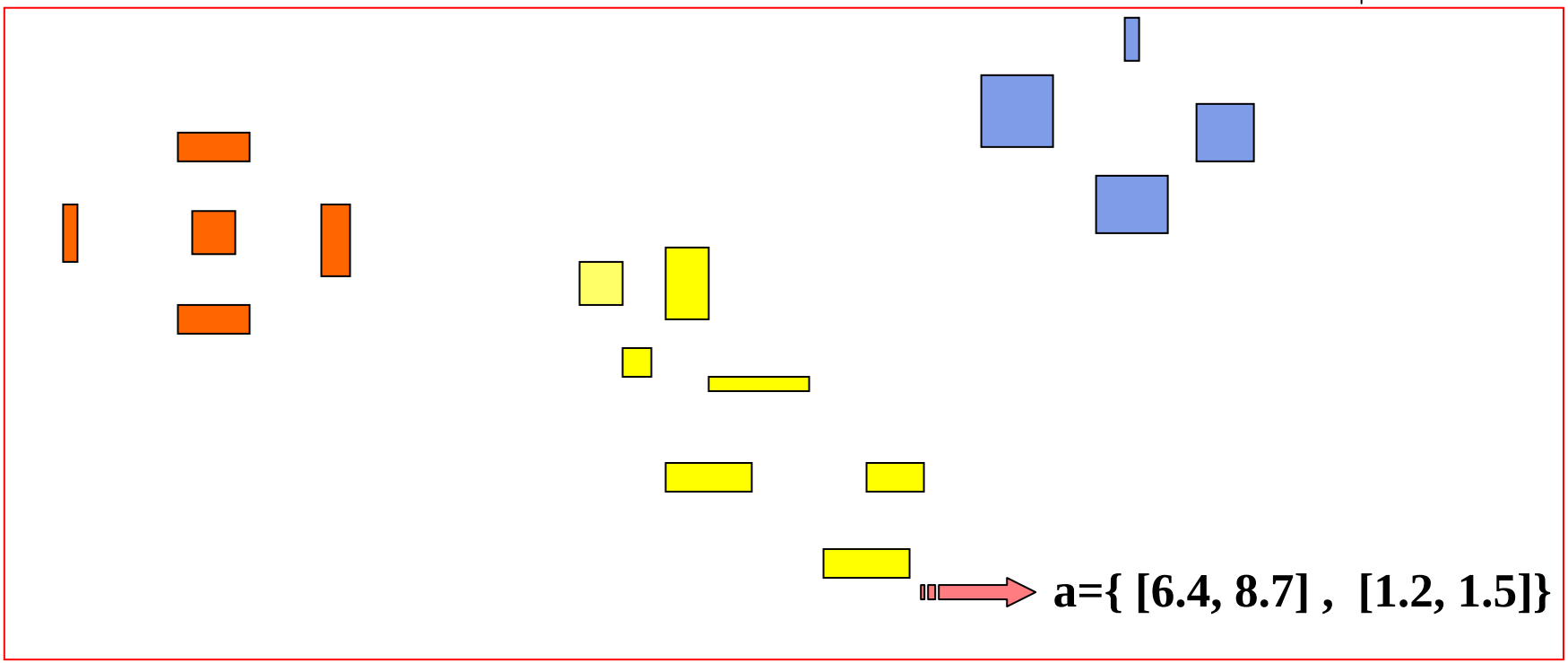
SEGMENTATION REGION GROWTH



SYMBOLIC CLASSIFIER

Centro de Informática U·F·P·E

Análise de Cluster

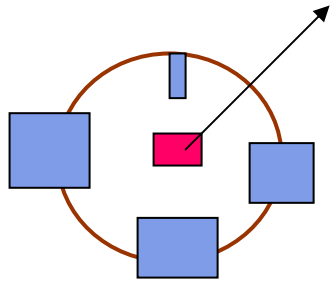


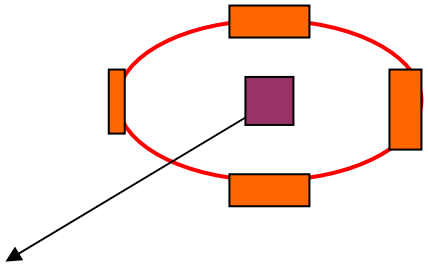
O algoritmo de nuvens dinâmicas visa encontrar uma partição e um conjunto de representantes das classes otimizando um critério de ajustamento entre classes e seus representantes. **As distâncias adaptativas** permitem encontrar clusters de formas e tamanhos diferentes.

Centro de Informática U·F·P·E

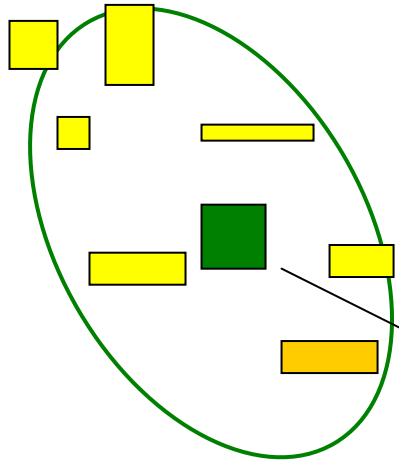
Saída: A partição e as descrições das classes

C_3: ([18.4, 19.6] , [9.7, 10.5])





C_1: ([3.4, 4.3] , [6.7, 7.5])



protótipo

C_2: ([13.4, 15.3] , [3.7, 5.5])

Critério a ser otimizado:

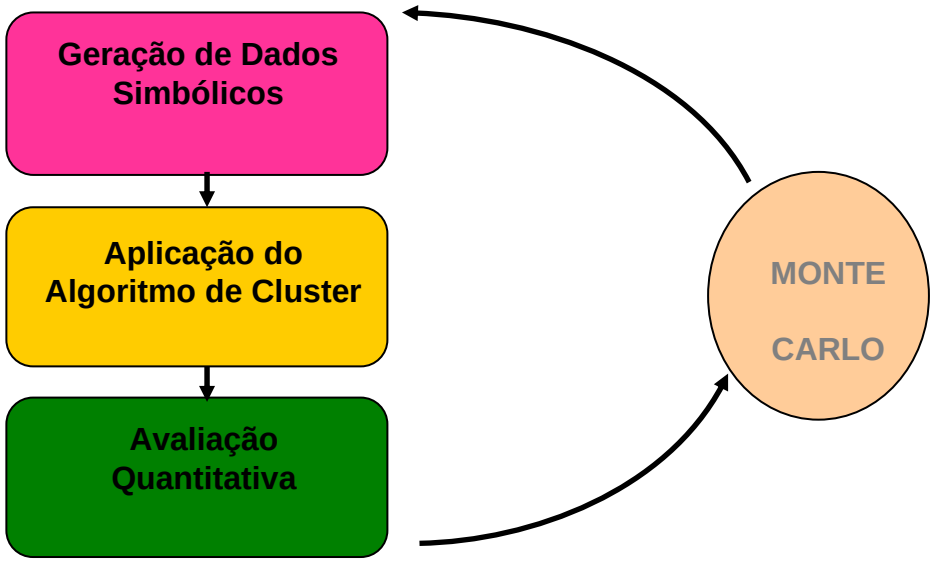
$$W = \sum_{i=1}^{k} \sum_{x \in C_i} d_i(x, G_i)$$

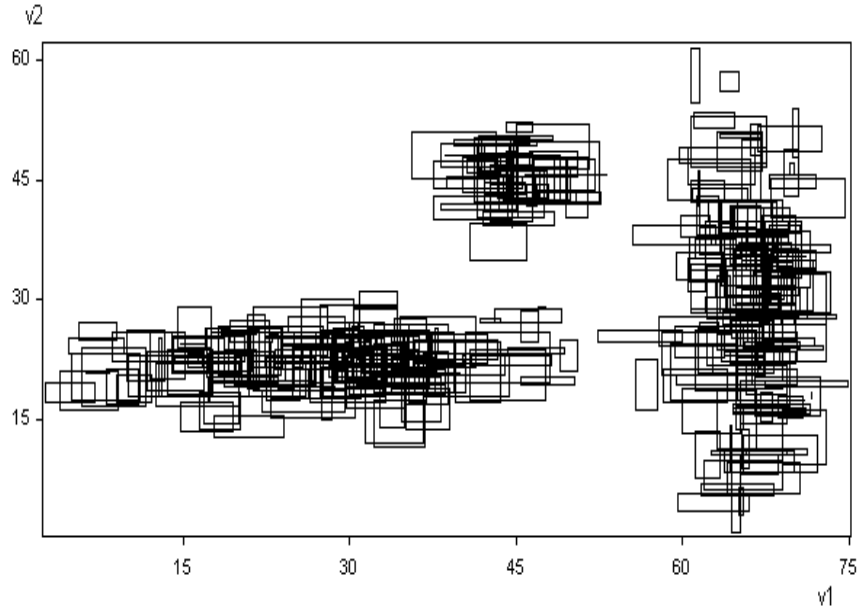
Distâncias Adaptativas

- City-Block
- Euclideana
- Chebyshev
- Mahalanobis

Centro de Informática U·F·P·E

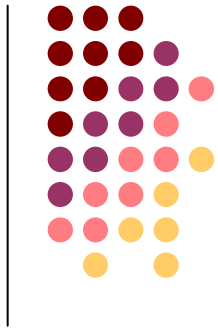
Método Monte Carlo





Centro de Informática U·F·P·E

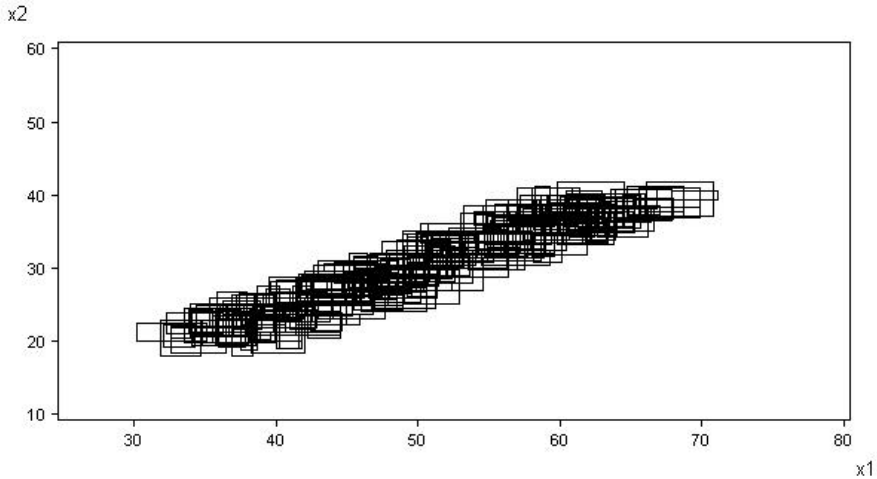
Modelos de Regressão para Dados Simbólicos Tipo-Intervalo

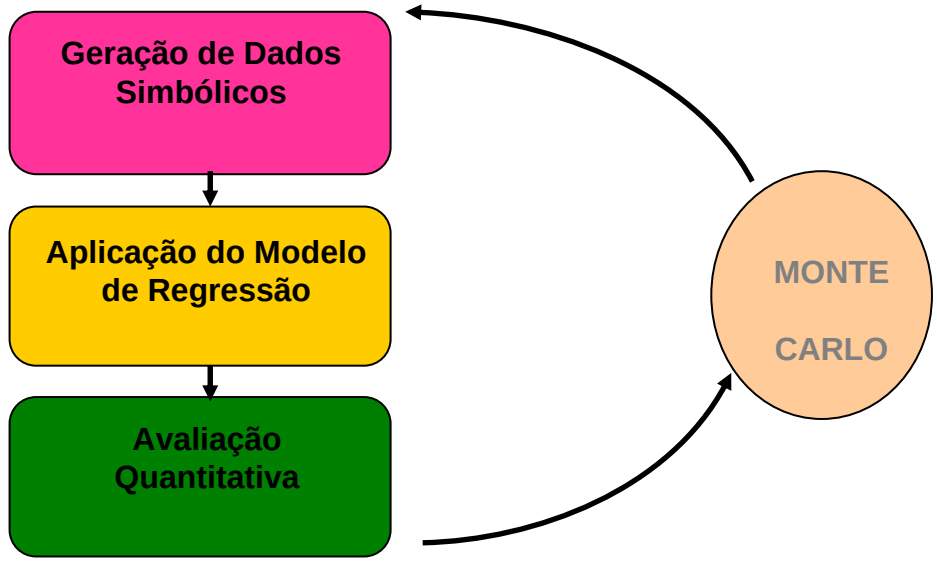


e	Pulso (Y)	Pressão Sistólica (X_1)	Pressão Diastólica (X_2)
1	[44-68]	[90-100]	[50-70]
2	[60-72]	[90-130]	[70-90]
3	[56-90]	[140-180]	[90-100]
4	[70-112]	[110-142]	[80-108]
5	[54-72]	[90-100]	[50-70]
6	[70-100]	[130-160]	[80-110]
7	[63-75]	[60-100]	[140-150]
8	[72-100]	[130-160]	[76-90]
9	[76-98]	[110-190]	[70-110]
10	[86-96]	[138-180]	[90-110]
11	[86-100]	[110-150]	[78-100]

Centro de Informática U·F·P·E

Problema de regressão: dados intervalares simulados





Analisando Gestões Administrativas Municipais

- Motivação:
  - Agrupar, por meio de uma abordagem simbólica, cidades que apresentem gestões administrativas similares.
  - Traçar um mapa administrativo de cada município identificando, mais detalhadamente, seus defeitos e suas virtudes
  - Utilizar a opinião de seus habitantes sobre alguns serviços públicos municipais

Centro de Informática U·F·P·E

As cidades

BJD	STL	GRN	ARC	BEZ
PLT	AEL	IGA	SCC	RCF
JDG	OLI	PET	CAM	CPN
CAR	CSA	VSA	SLM	GOI
ARA	GVT	IPJ	PES	ESC

As Variáveis de Interesse

Serviços Públicos Municipais (variáveis)	Codificação
Limpeza de ruas e avenidas	**Limpeza**
Recolhimento do lixo domiciliar	**Lixo**
Iluminação pública	**Iluminação**
Ensino municipal	**Educação**
Assistência à população pobre	**Ass_pobre**
Conservação de praças, parques e jardins	**Conserv**
Assistência médica municipal	**Saúde**
Promoção de festas populares	**Festa**
Abastecimento de água	**Agua**
Rede de esgoto/saneamento	**Saneam**
Segurança	**Segurança**
Pavimentação de ruas e avenidas	**Paviment**
Apoio à geração de empregos	**Emprego**
Conservação de estradas	**Estrada**
Trânsito	**Transito**
Avaliação Administrativa	**Aval**

16 variáveis

Centro de Informática U·F·P·E

Os Dados

- Escala de Avaliação de Desempenho das Variáveis de Interesse

Escore	1	2	3	4	5
Categoria	Péssimo	Ruim	Regular	Bom	Ótimo

O BD clássico

ID	Município*	Limpeza	Lixo	....	Transito	Aval
1	BJD	Regular	Bom	....	Ruim	Ruim
2	BJD	Bom	Otimo	....	Bom	Bom
....	....	....	....	....	....	....
231	BJD	Bom	Otimo	....	Bom	Bom
232	STL	Pessimo	Bom	....	Ruim	Regular
....	....	....	....	....	....	....
5.241	ESC	Ruim	Bom	....	Bom	Péssimo

89.097 células

Centro de Informática U·F·P·E

O BD simbólico

	Limpeza	..	Trânsito	Saldo
JDG	Péssi(0.29), Ruim(0.15), Regul(0.26), Bom(0.26), Ótimo(0.03)	..	Péssi(0.22), Ruim(0.17), Regul(0.37), Bom(0.23), Ótimo(0.02)	n
RFC	Péssi(0.13), Ruim(0.10), Regul(0.34), Bom(0.35), Ótimo(0.09)	..	Péssi(0.15), Ruim(0.12), Regul(0.34), Bom(0.37), Ótimo(0.03)	n
OLI	Péssi(0.31), Ruim(0.13), Regul(0.27), Bom(0.24), Ótimo(0.05)	..	Péssi(0.16), Ruim(0.09), Regul(0.34), Bom(0.38), Ótimo(0.03)	n
PLT	Péssi(0.14), Ruim(0.07), Regul(0.24), Bom(0.41), Ótimo(0.14)	..	Péssi(0.12), Ruim(0.06), Regul(0.47), Bom(0.33), Ótimo(0.02)	p
CAR	Péssi(0.07), Ruim(0.04), Regul(0.24), Bom(0.49), Ótimo(0.15)	..	Péssi(0.19), Ruim(0.06), Regul(0.23), Bom(0.48), Ótimo(0.04)	n
....	....	..	....	....
ESC	Péssi(0.21), Ruim(0.06), Regul(0.31), Bom(0.32), Ótimo(0.09)	..	Péssi(0.21), Ruim(0.14), Regul(0.22), Bom(0.41), Ótimo(0.02)	n

400 células

Centro de Informática U·F·P·E

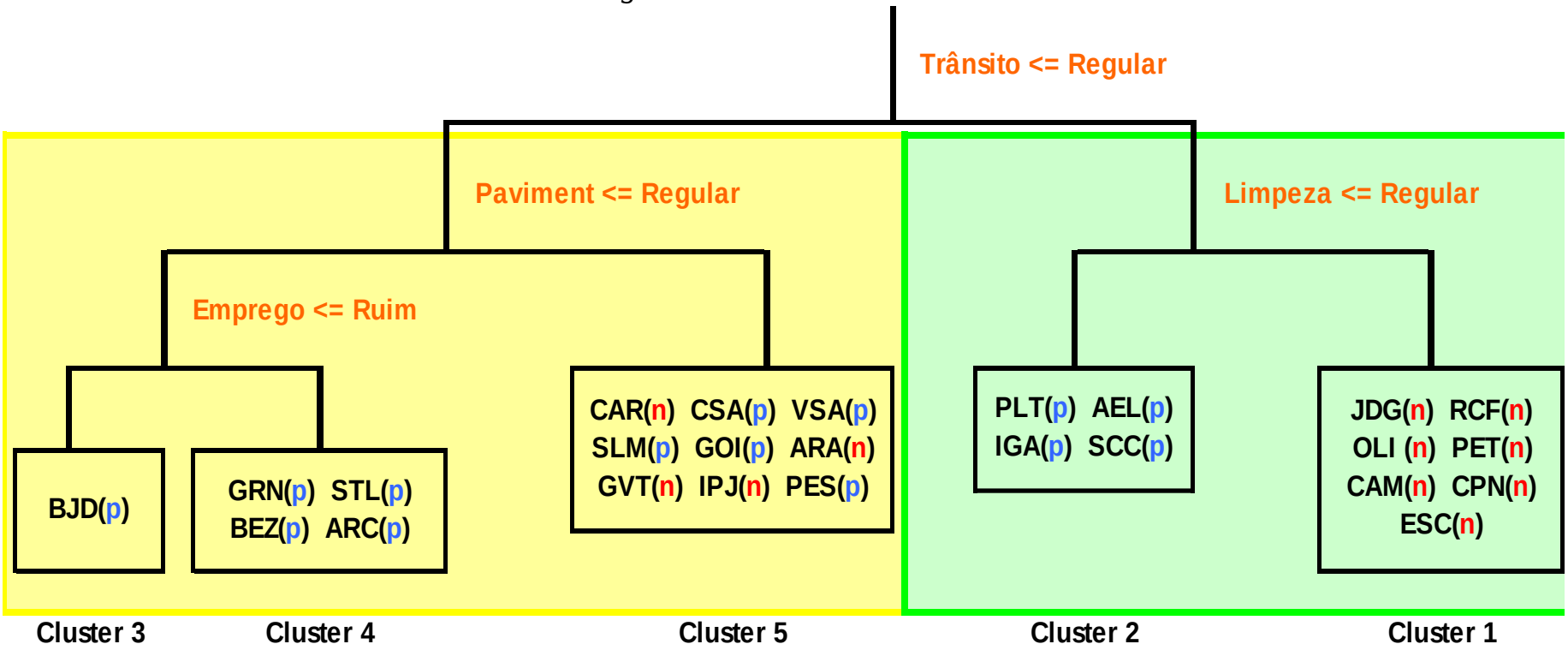
Os Dados

Novas Variáveis Simbólicas

- Saldo = {n, p}
  - Avaliação Administrativa
  - Se [(%Ótimo + %Bom) - (%Ruim + %Péssimo)] > %Regular
    - então Saldo = p (gestão adm. positiva)
    - cc. Saldo = n (gestão adm. negativa)
- Variáveis Multivaloradas Ponderadas (Serviços Públicos)
  - Limpeza, Lixo, Iluminação, Educação, Assistência à pobreza, Conservação de praças, Saúde, Festas, Água, Saneamento, Segurança, Pavimentação, Emprego, Estrada e Trânsito.

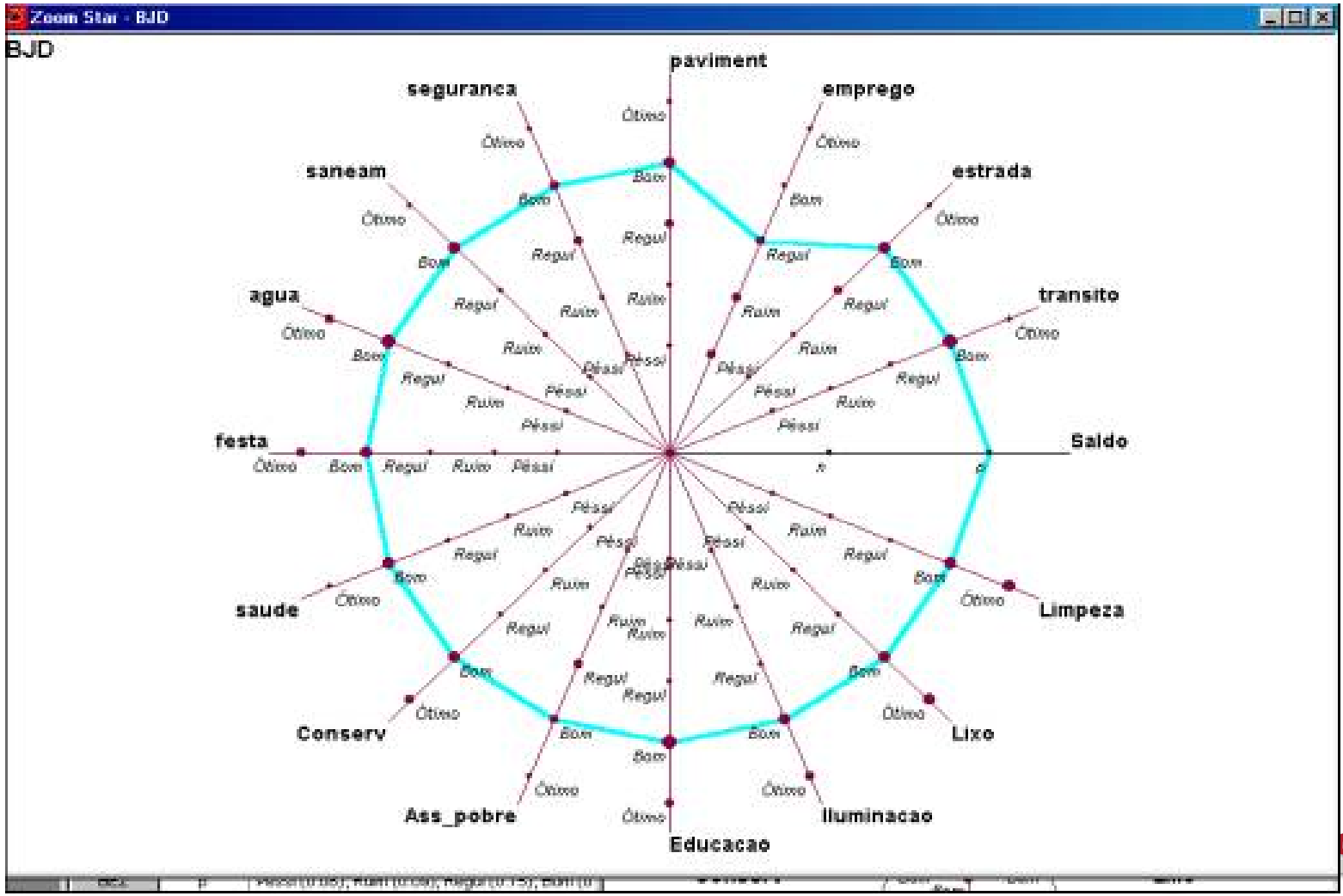
Métodos Utilizados : A árvore de Cluster

Figura 1: A Árvore de Cluster

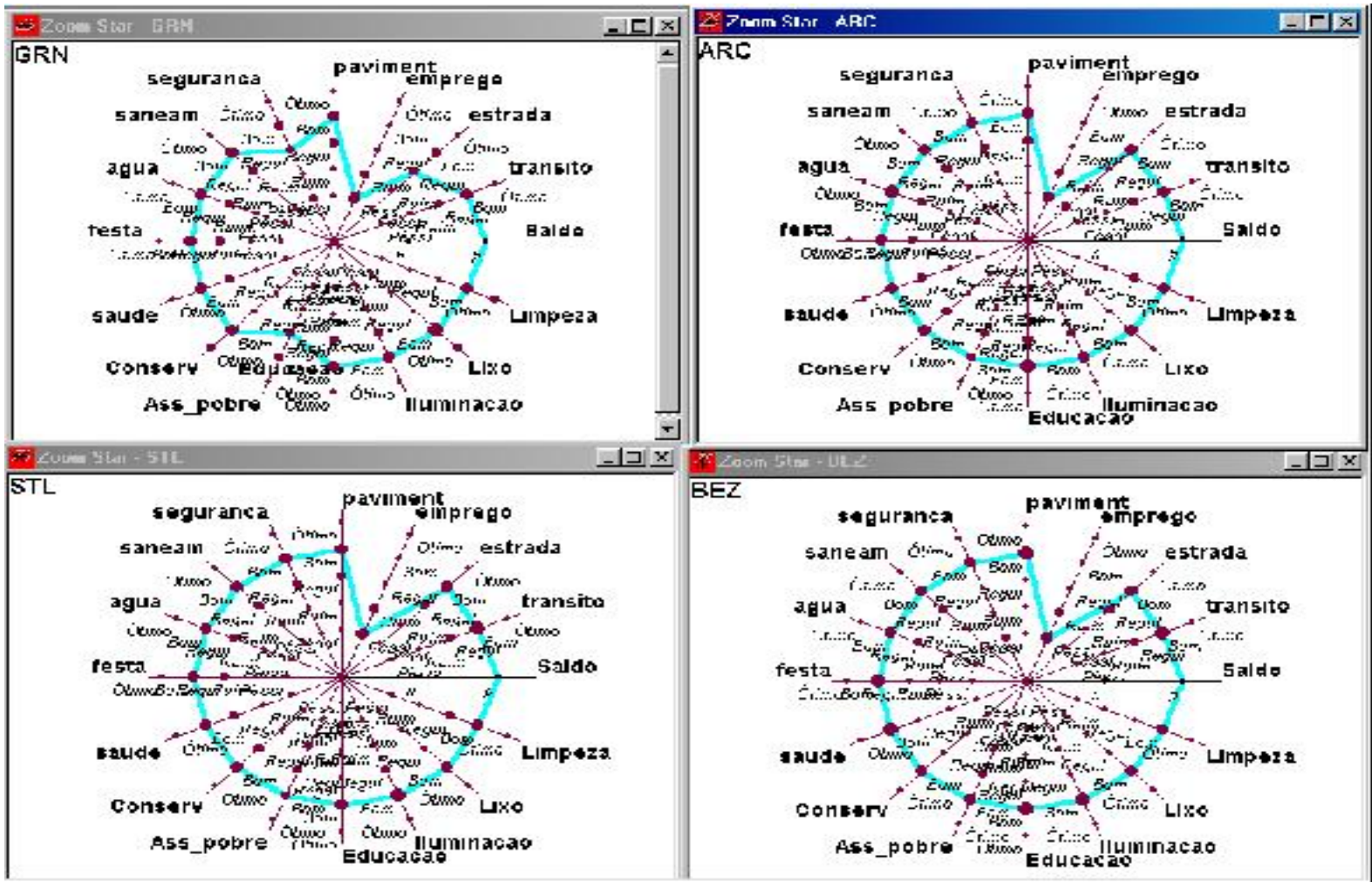


Métodos Utilizados:

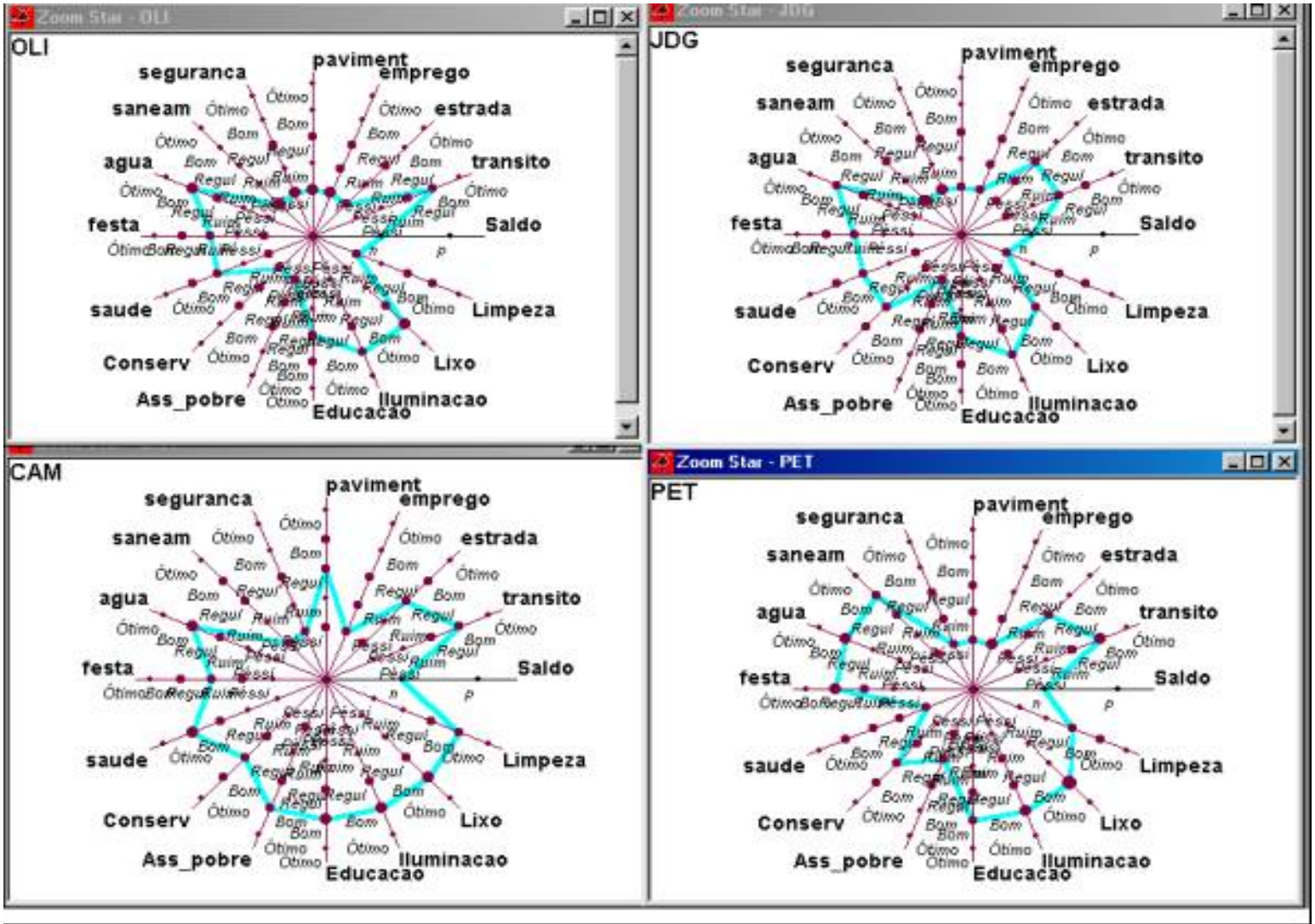
Zoom Stars 2D — cluster 3



Métodos Utilizados :
Zoom Stars 2D — cluster 4



Zoom Stars 2D — cluster 1



Conclusões

- Utilizando a **Análise de Cluster** foi possível identificar padrões de gestões administrativas;
- Através dos gráficos **Zoom Stars 2D** identificamos os pontos positivos e negativos de cada gestão;
- As estrelas maiores e mais uniformes implicam em gestões positivas. Já estrelas menores e/ou com diversas deformidades caracterizam gestões negativas.

Centro de Informática U·F·P·E

Conclusões

- Uma gestão administrativa negativa foi caracterizada pelas cidades que obtiveram uma classificação *Regular, Ruim ou Péssima* em *Trânsito* e *Limpeza de Ruas ou Avenidas;*
- Uma gestão administrativa positiva foi caracterizada pelas cidades que obtiveram uma classificação *Boa* ou *Ótima* em *Trânsito* e *Pavimentação de ruas ou avenidas;*
- A gestão administrativa da cidade BJD mereceu destaque em virtude da classificação *Regular* na variável *Apoio a geração de empregos*.