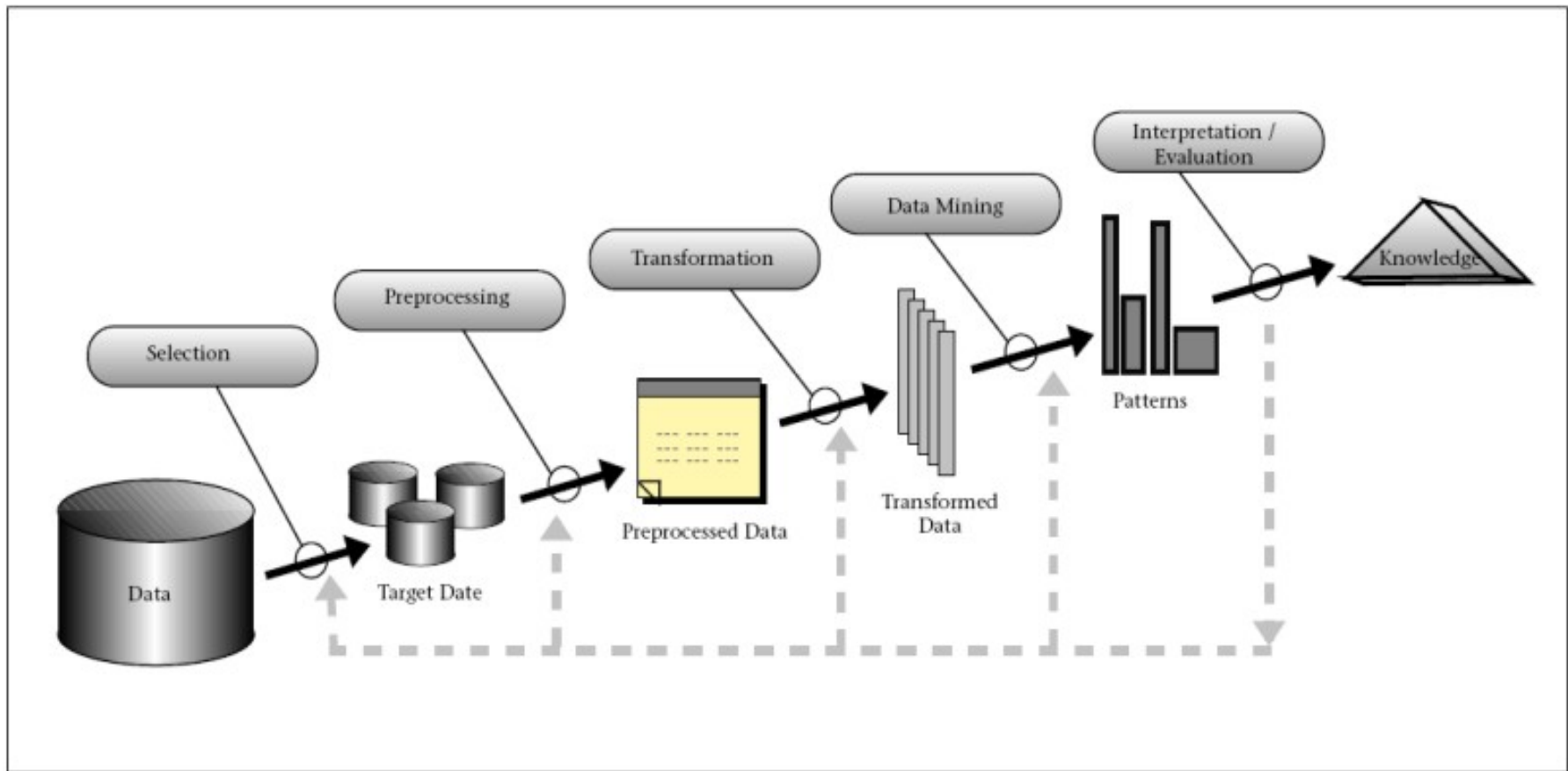


Mineração de Dados

Principais abordagens

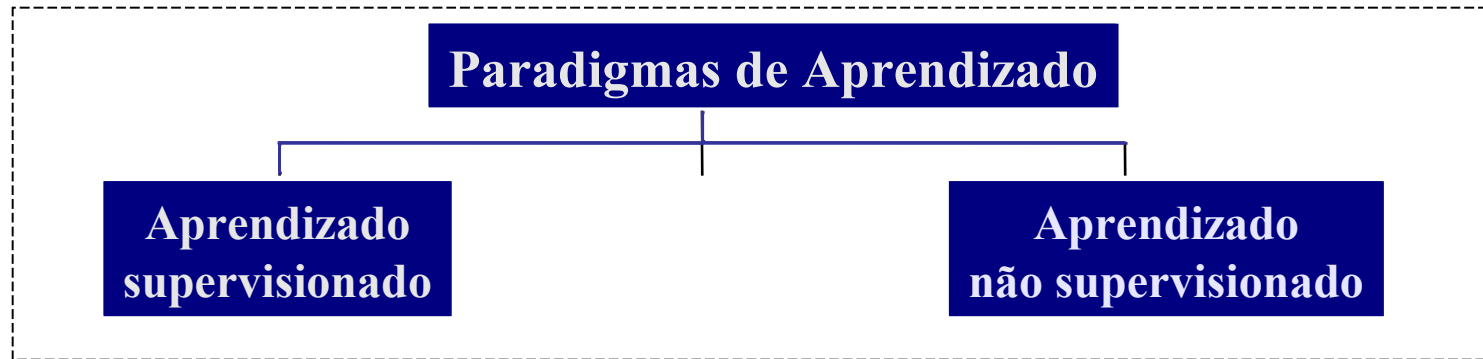
prof^a Carine G. Webber

Processo de Descoberta de Conhecimento



Fayyad, 1996

Paradigmas Algorítmicos



Tipos de Aprendizado

4

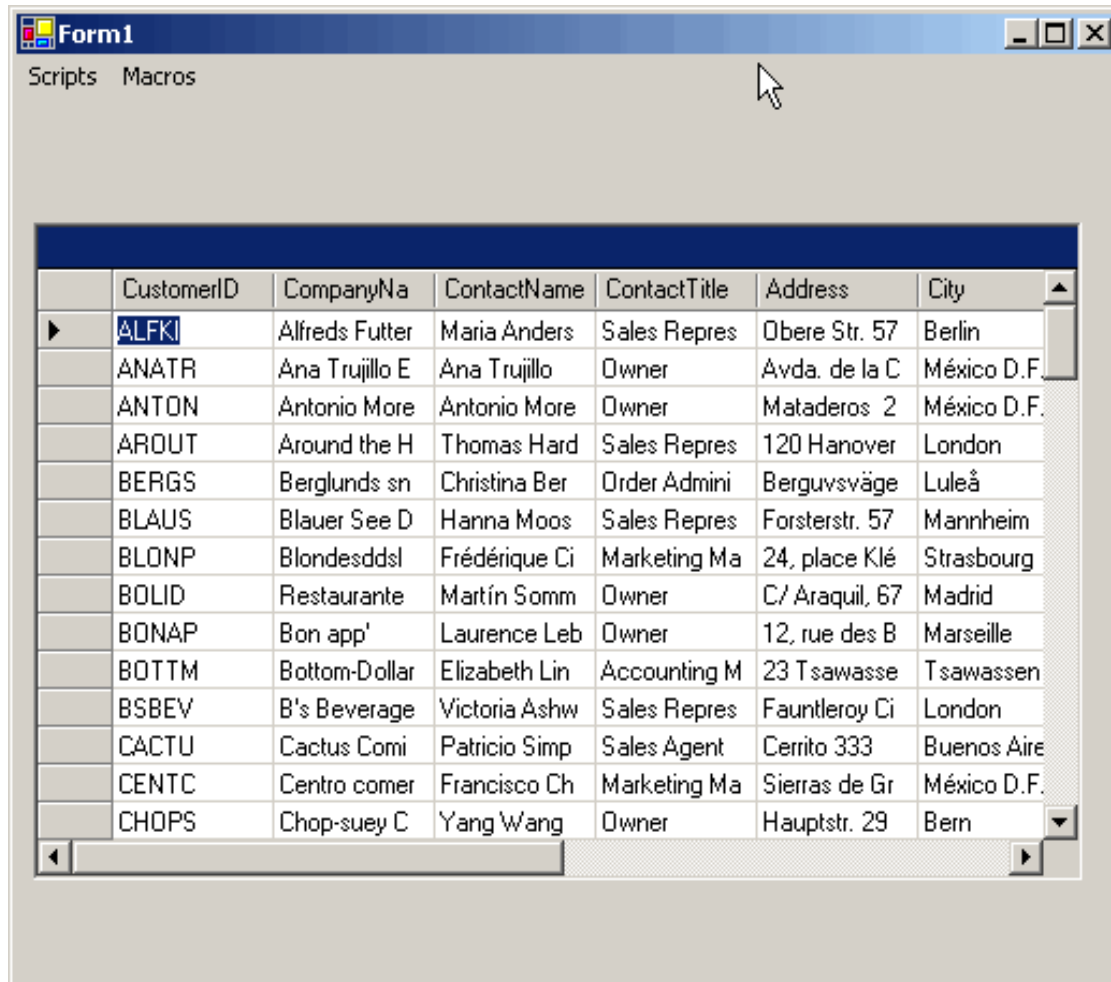
Supervisionado

- Resolve problemas de classificação de dados
- Ocorre a partir de exemplos previamente classificados
- Modelo dos dados é conhecido

Não supervisionado

- Resolve problemas de agrupamento de dados similares
- As categorias estão implícitas e subjacentes aos dados
- Modelo dos dados é desconhecido

Exemplo: dataset de clientes



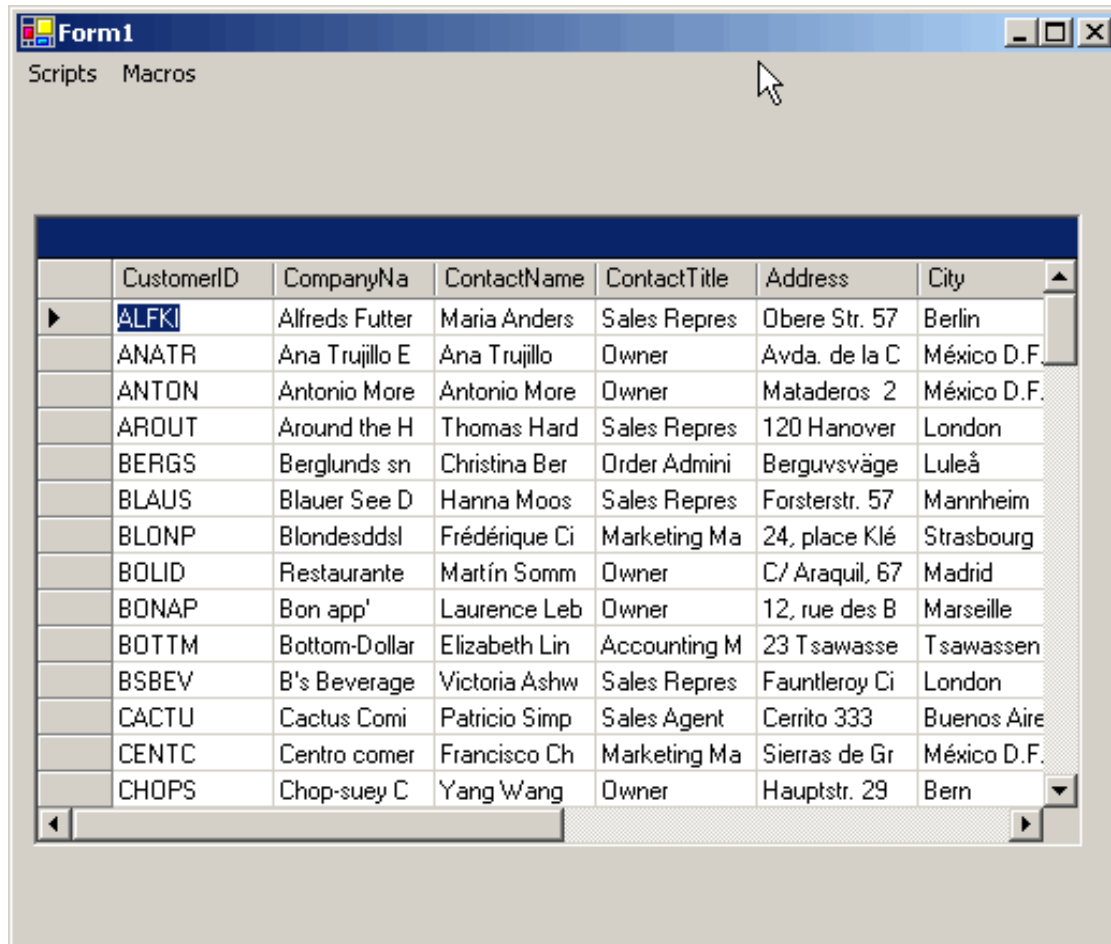
	CustomerID	CompanyNa	ContactName	ContactTitle	Address	City
▶	ALFKI	Alfreds Futter	Maria Anders	Sales Repres	Obere Str. 57	Berlin
	ANATR	Ana Trujillo E	Ana Trujillo	Owner	Avda. de la C	México D.F.
	ANTON	Antonio More	Antonio More	Owner	Mataderos 2	México D.F.
	AROUT	Around the H	Thomas Hard	Sales Repres	120 Hanover	London
	BERGS	Berglunds sn	Christina Ber	Order Admini	Berguvsväge	Luleå
	BLAUS	Blauer See D	Hanna Moos	Sales Repres	Forsterstr. 57	Mannheim
	BLONP	Blondesddsl	Frédérique Ci	Marketing Ma	24, place Klé	Strasbourg
	BOLID	Restaurante	Martín Somm	Owner	C/ Araquil, 67	Madrid
	BONAP	Bon app'	Laurence Leb	Owner	12, rue des B	Marseille
	BOTTM	Bottom-Dollar	Elizabeth Lin	Accounting M	23 T sawasse	T sawassen
	BSBEV	B's Beverage	Victoria Ashw	Sales Repres	Fauntleroy Ci	London
	CACTU	Cactus Comi	Patricio Simp	Sales Agent	Cerrito 333	Buenos Aire
	CENTC	Centro comer	Francisco Ch	Marketing Ma	Sierras de Gr	México D.F.
	CHOPS	Chop-suey C	Yang Wang	Owner	Hauptstr. 29	Bern

Cenário não-supervisionado

Como podemos particionar
nossos clientes?

Queremos descobrir
categorias de clientes:
Geografia, Pedidos,
Pagamentos, Inovação, ...

Exemplo: dataset de clientes

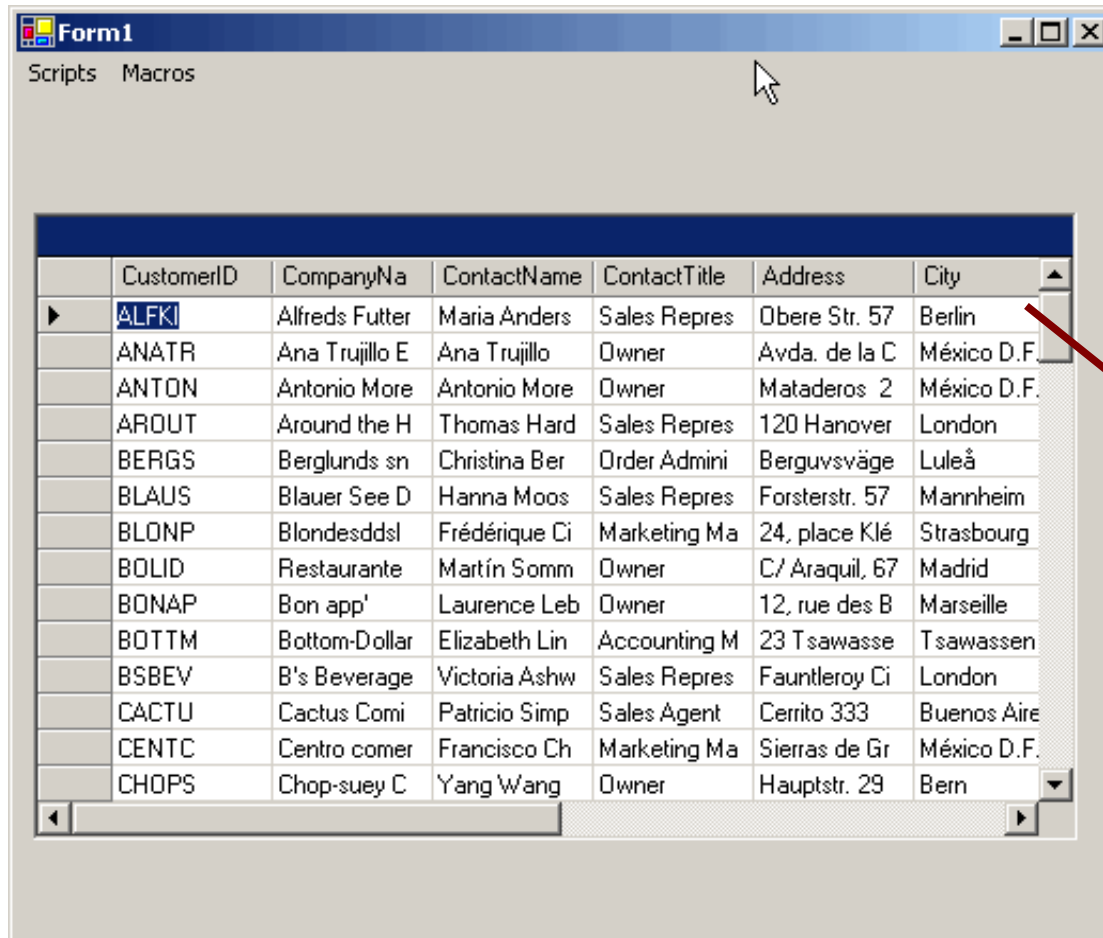


CustomerID	CompanyNa	ContactName	ContactTitle	Address	City
ALFKI	Alfreds Futter	Maria Anders	Sales Repres	Obere Str. 57	Berlin
ANATR	Ana Trujillo E	Ana Trujillo	Owner	Avda. de la C	México D.F.
ANTON	Antonio More	Antonio More	Owner	Mataderos 2	México D.F.
AROUT	Around the H	Thomas Hard	Sales Repres	120 Hanover	London
BERGS	Berglunds sn	Christina Ber	Order Admini	Berguvsväge	Luleå
BLAUS	Blauer See D	Hanna Moos	Sales Repres	Forsterstr. 57	Mannheim
BLONP	Blondesddsl	Frédérique Ci	Marketing Ma	24, place Klé	Strasbourg
BOLID	Restaurante	Martín Somm	Owner	C/ Araquil, 67	Madrid
BONAP	Bon app'	Laurence Leb	Owner	12, rue des B	Marseille
BOTTM	Bottom-Dollar	Elizabeth Lin	Accounting M	23 T sawasse	T sawassen
BSBEV	B's Beverage	Victoria Ashw	Sales Repres	Fauntleroy Ci	London
CACTU	Cactus Comi	Patricio Simp	Sales Agent	Cerrito 333	Buenos Aire
CENTC	Centro comer	Francisco Ch	Marketing Ma	Sierras de Gr	México D.F.
CHOPS	Chop-suey C	Yang Wang	Owner	Hauptstr. 29	Bern

Perfil Inovador

Perfil Conservador

Exemplo: dataset de clientes



Form1

Scripts Macros

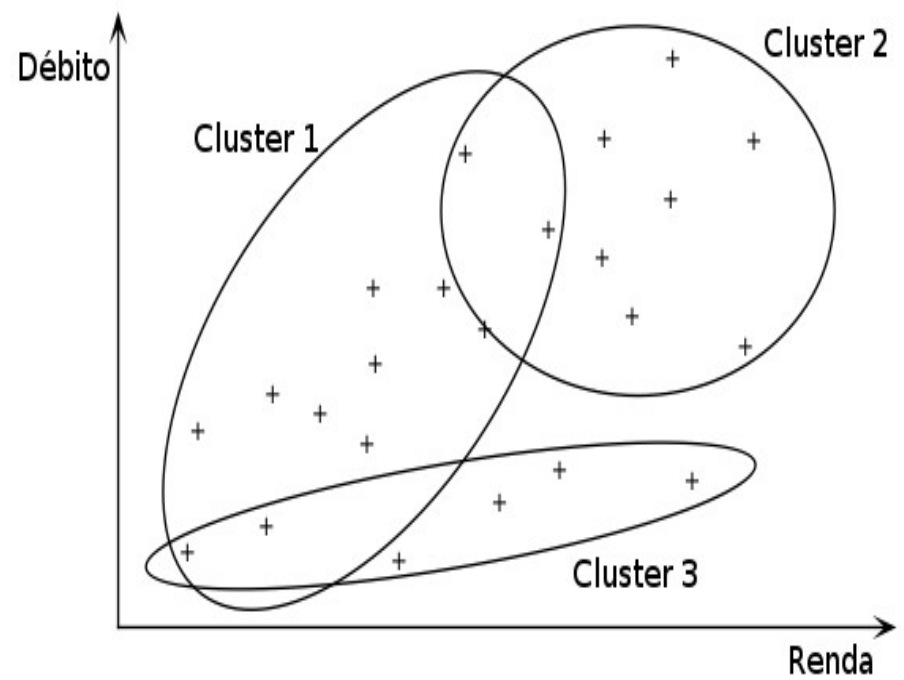
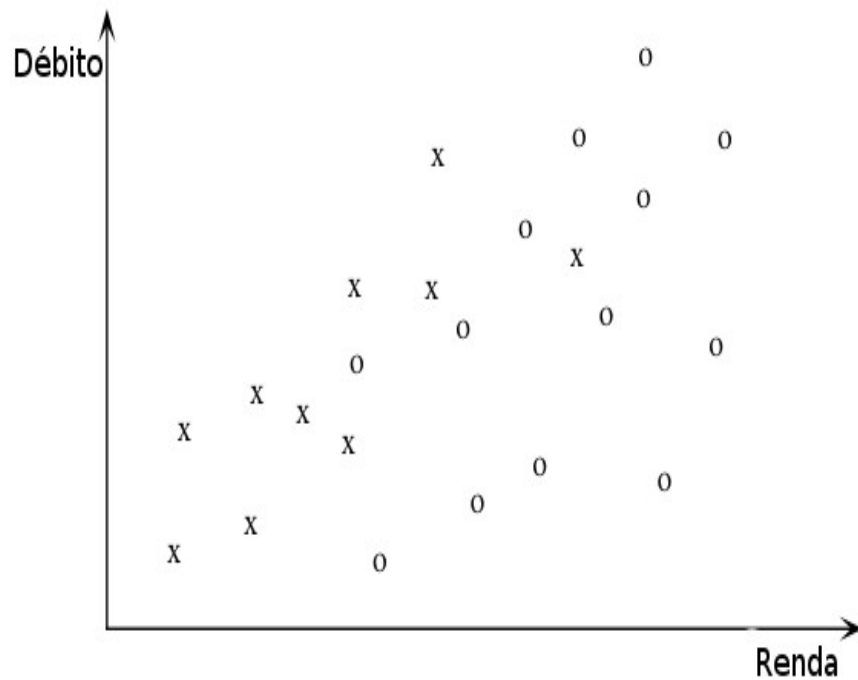
	CustomerID	CompanyNa	ContactName	ContactTitle	Address	City
▶	ALFKI	Alfreds Futter	Maria Anders	Sales Repres	Obere Str. 57	Berlin
	ANATR	Ana Trujillo E	Ana Trujillo	Owner	Avda. de la C	México D.F.
	ANTON	Antonio More	Antonio More	Owner	Mataderos 2	México D.F.
	AROUT	Around the H	Thomas Hard	Sales Repres	120 Hanover	London
	BERGS	Berglunds sn	Christina Ber	Order Admini	Berguvsväge	Luleå
	BLAUS	Blauer See D	Hanna Moos	Sales Repres	Forsterstr. 57	Mannheim
	BLONP	Blondesddsl	Frédérique Ci	Marketing Ma	24, place Klé	Strasbourg
	BOLID	Restaurante	Martín Somm	Owner	C/ Araquil, 67	Madrid
	BONAP	Bon app'	Laurence Leb	Owner	12, rue des B	Marseille
	BOTTM	Bottom-Dollar	Elizabeth Lin	Accounting M	23 T sawasse	T sawassen
	BSBEV	B's Beverage	Victoria Ashw	Sales Repres	Fauntleroy Ci	London
	CACTU	Cactus Comi	Patricio Simp	Sales Agent	Cerrito 333	Buenos Aire
	CENTC	Centro comer	Francisco Ch	Marketing Ma	Sierras de Gr	México D.F.
	CHOPS	Chop-suey C	Yang Wang	Owner	Hauptstr. 29	Bern

Cenário supervisionado

Perfis previamente definidos

Perfil Inovador ou
Perfil Conservador

Agrupamento ou Clustering



Técnicas de Clustering

Se aplicam quando:

- Deseja-se agrupar um conjunto de instâncias não classificadas segundo critérios de similaridade.
- Os clusters (grupos) refletem algum mecanismo que funciona no domínio das instâncias e faz com que **algumas instâncias sejam mais parecidas entre si** do que com o restante delas.
- Utiliza-se técnicas diferentes das de **classificação** e **associação** utilizadas no aprendizado supervisionado.

Clustering

Técnica de análise de dados baseada na criação de classes através da partição do banco de dados em sub-conjuntos de instâncias. Esta partição é realizada considerando-se a similaridade entre os valores dos atributos das instâncias.

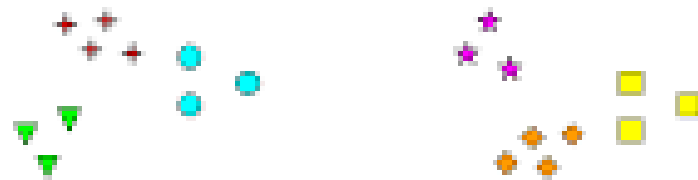
Técnicas de agrupamento:

- **Algoritmos particionais:** criam partições do conjunto de dados usando algum critério.
- **Algoritmos hierárquicos:** criam grupos através da decomposição hierárquica do conjunto de dados.
- **Algoritmos baseados em densidade:** criam grupos a partir da densidade dos objetos em relação a sua localização espacial.

Algoritmos Particionais



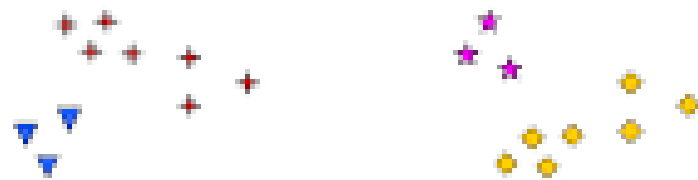
Quantos clusters existem?



Seis clusters ?

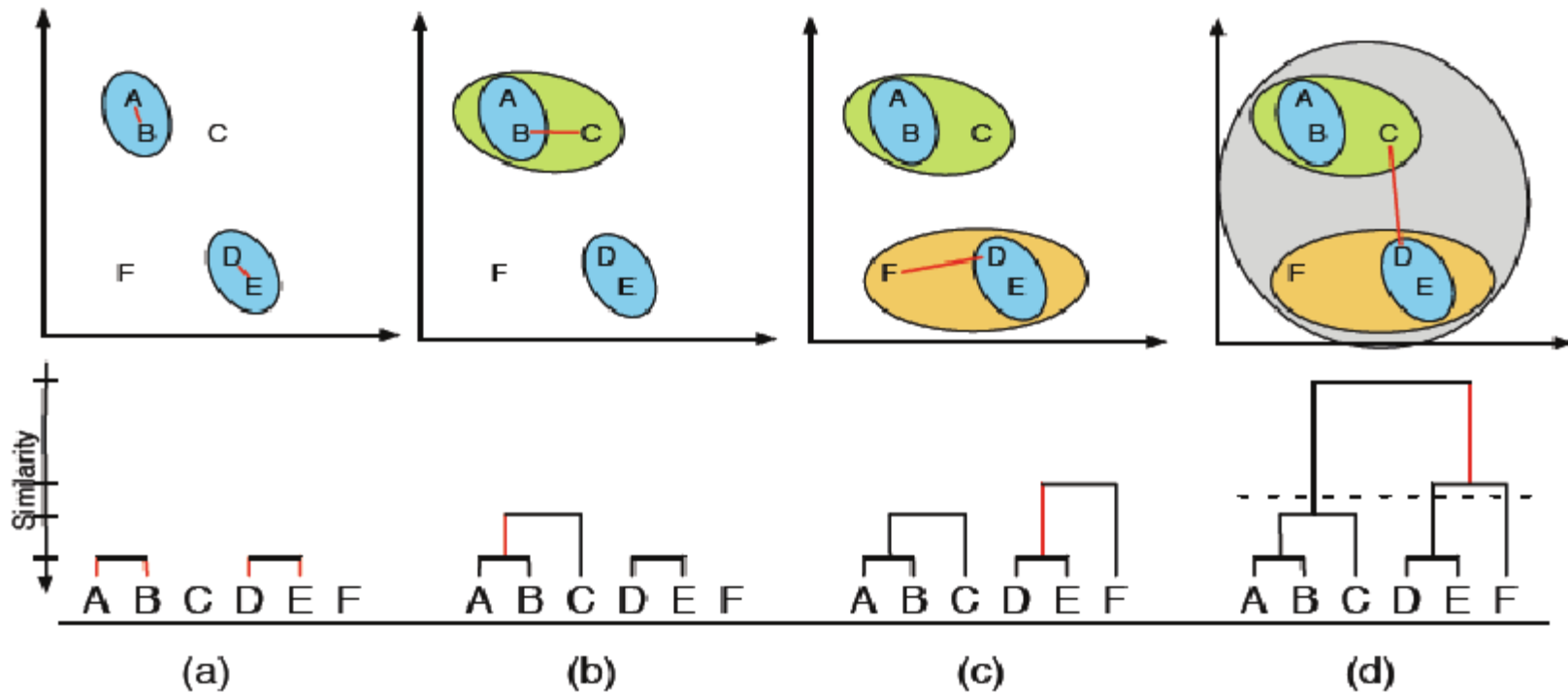


Dois clusters?

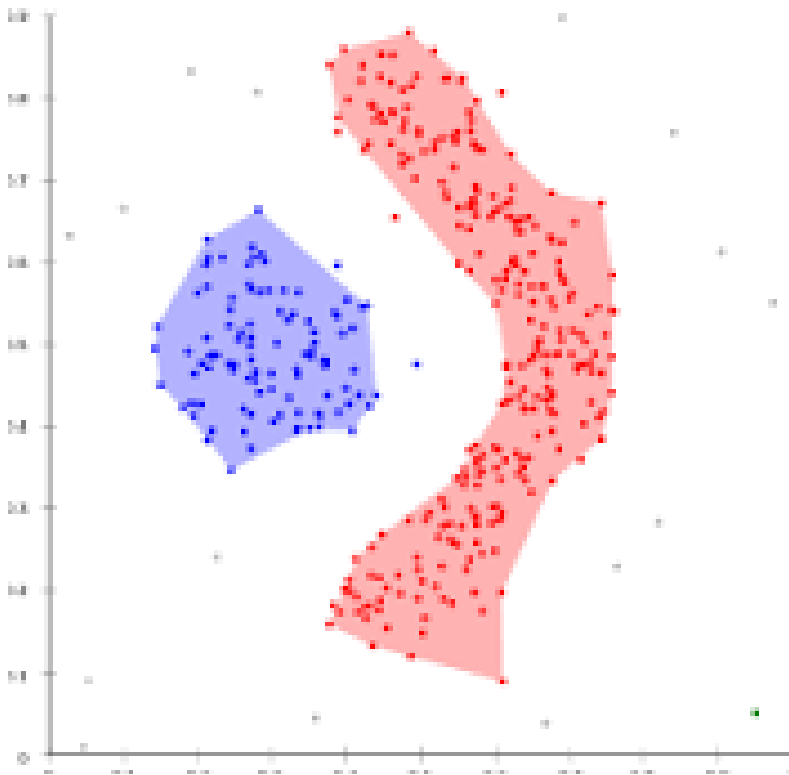


Quatro clusters ?

Algoritmos Hierárquicos



Algoritmos baseados em Densidade



A densidade é determinada pelo número de pontos localizados em uma região ou dentro de um raio.

O algoritmo localiza regiões de alta densidade separadas por regiões de baixa densidade.

Exemplos de técnicas



- **K-Means:** agrupa as instâncias em função da sua distância Euclidiana em relação a um centróide.

- **EM:** agrupa as instâncias através da sua similaridade probabilística com um grupo de dados.

Medidas de Distância

São utilizadas para calcular similaridades entre objetos, podendo ser baseadas em uma única dimensão ou em múltiplas dimensões.

- **Distância Euclidiana: é a distância entre dois objetos em um espaço multidimensional.**
- **Distância Euclidiana Quadrática: dá maior peso a objetos mais distantes.**
- **Distância de Manhattan (City-block): distância absoluta entre dois objetos.**
- **Distância de Chebychev: é a distância máxima entre dois objetos.**

Medidas de Distâncias

Assuma os vetores p, q, r e s no plano xy

p=(x1,y1) q=(x2,y2)							
exemplo	x	y					
p=(x1,y1)	1	2					
q=(x2,y2)	3	4					
exemplo	x	y					
r=(x1,y1)	10	2					
s=(x2,y2)	3	6					

distância euclidiana quadrática(p,q) = 8

distância euclidiana (p,q) = 2,8284271247

distância Manhatan (p,q) = 4

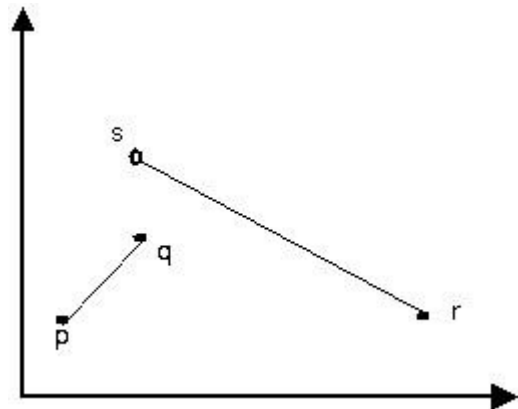
distância Chebyshev(p,q)= 2

distância euclidiana quadrática(r,s) = 65

distância euclidiana (r,s) = 8,0622577483

distância Manhatan (r,s) = 11

distância Chebyshev (r,s)= 7



Normalização de atributos

Consiste em alterar o valor de todos os atributos para que fiquem dentro da faixa entre 0 e 1.

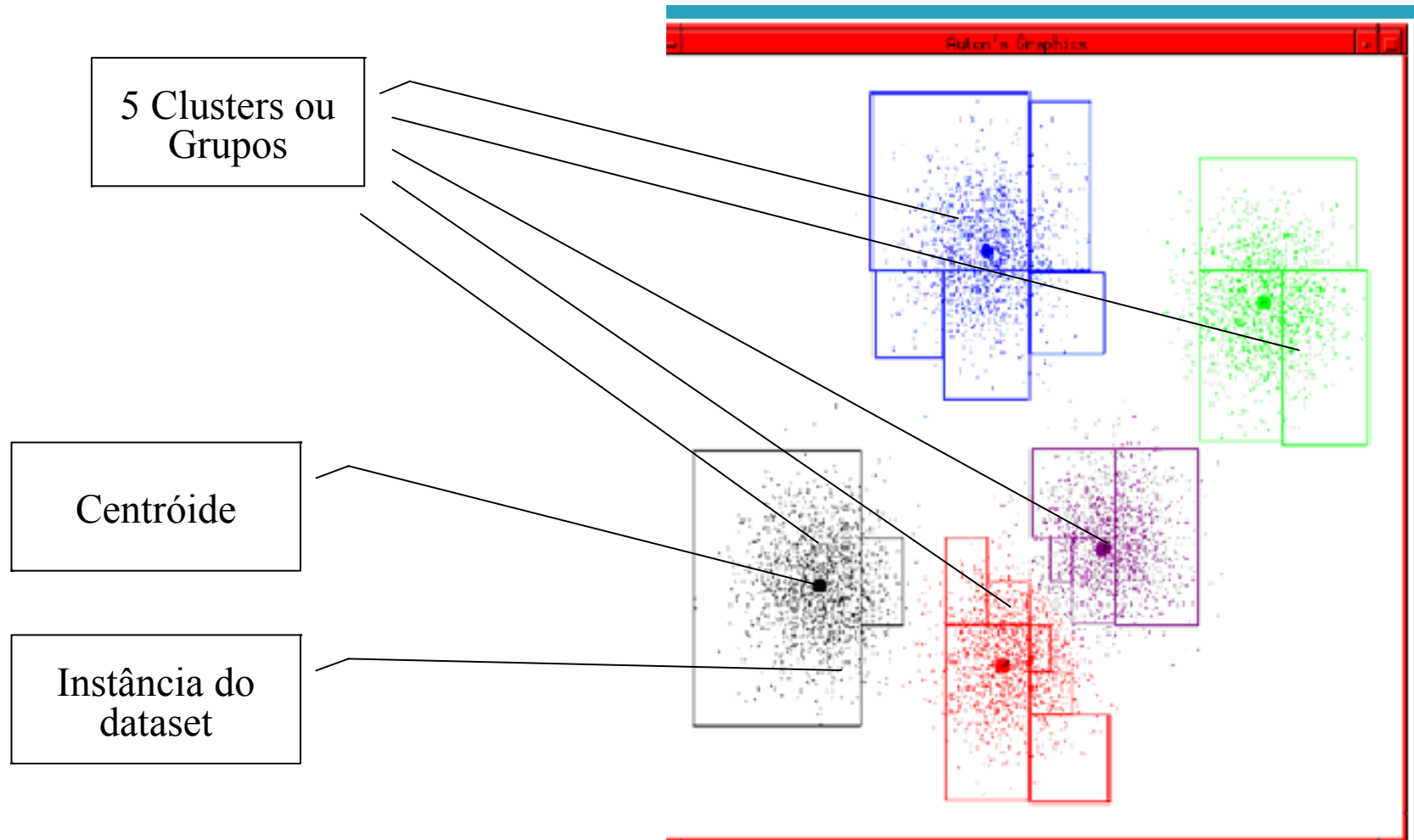
Têm a função de minimizar a diferença na escala dos valores dos atributos.

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

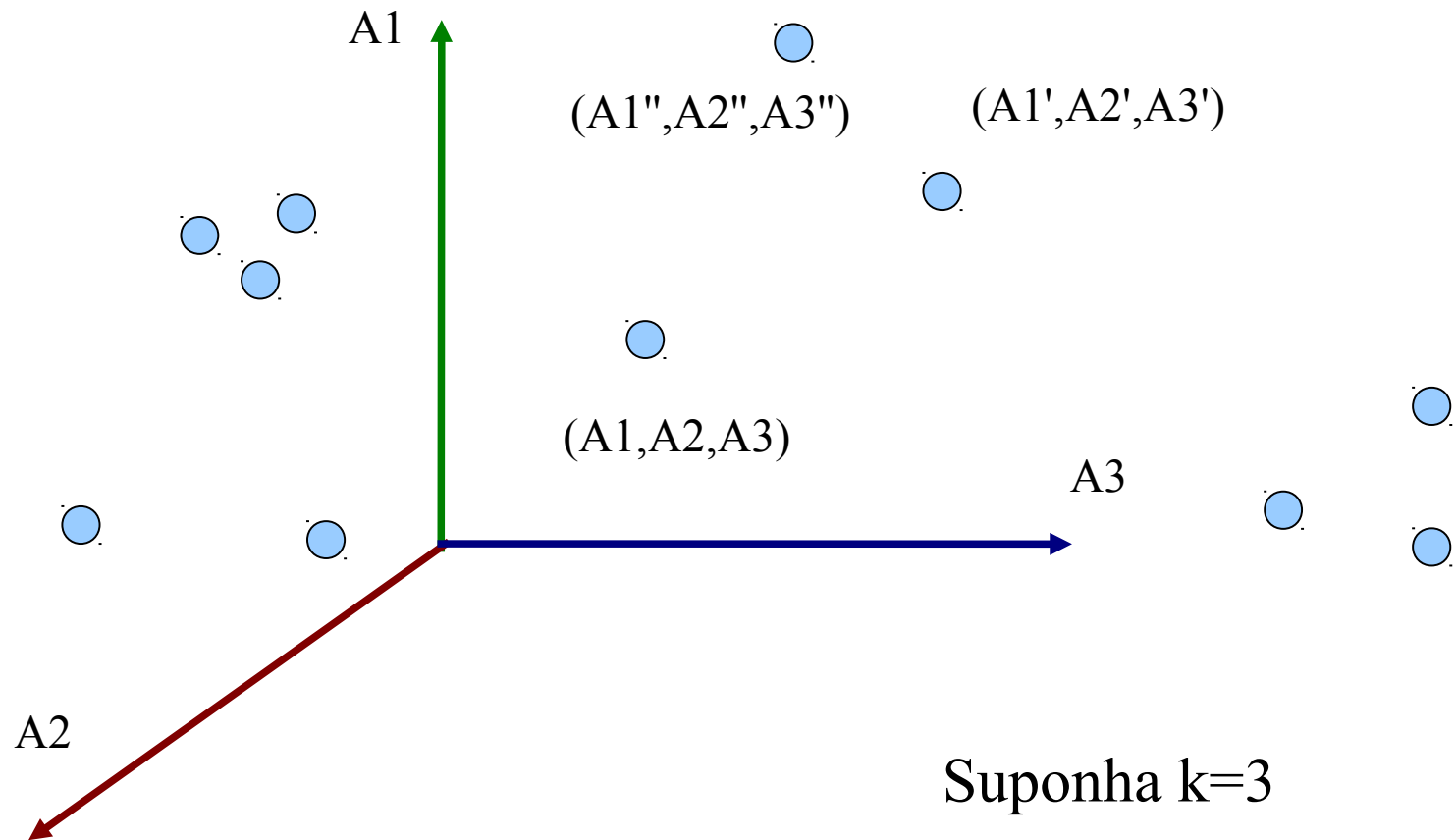
$$a_i = \frac{v_i}{t}$$

onde v_i é o valor atual do atributo i , e t é o valor total do atributo.

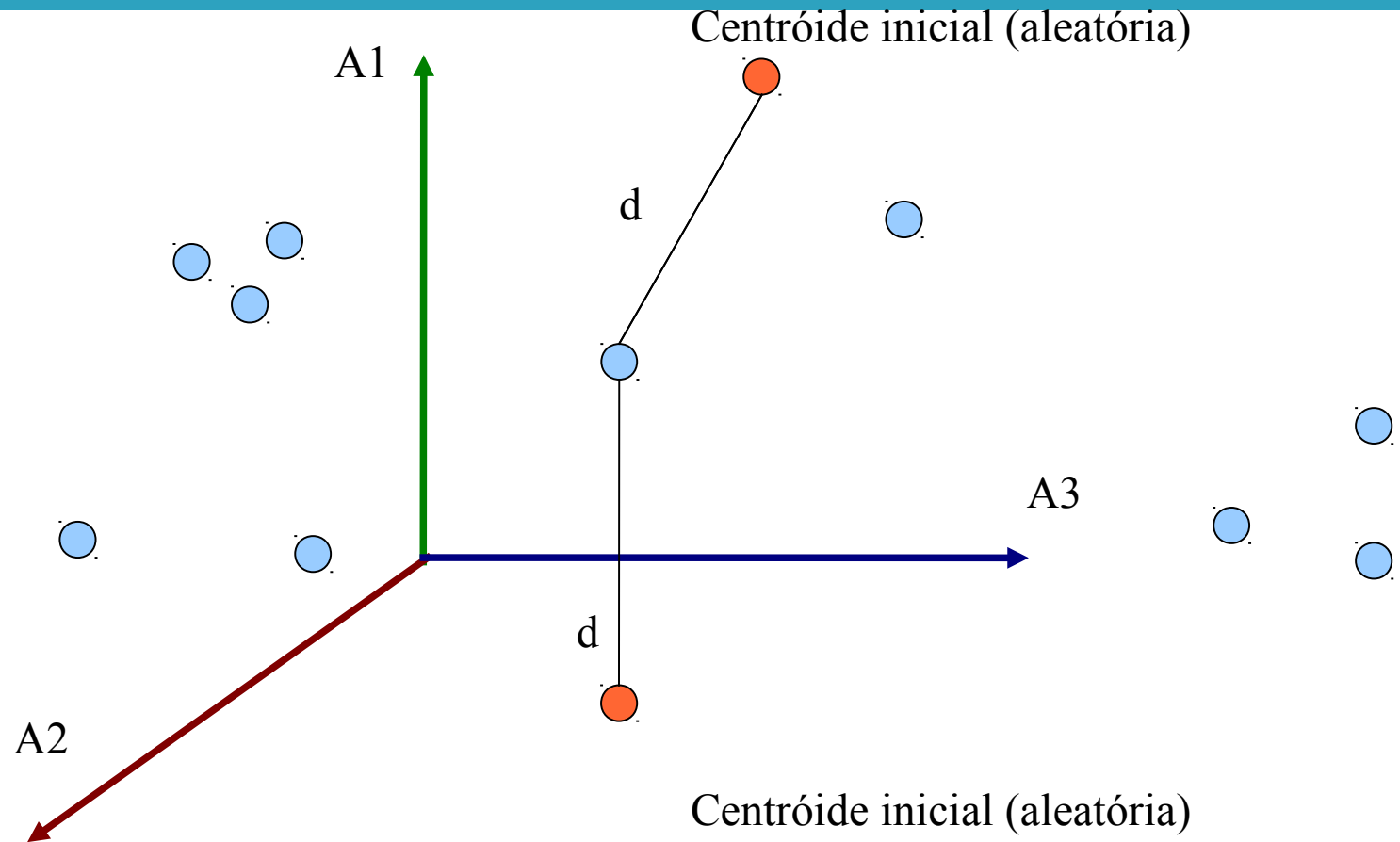
Algoritmo K-means



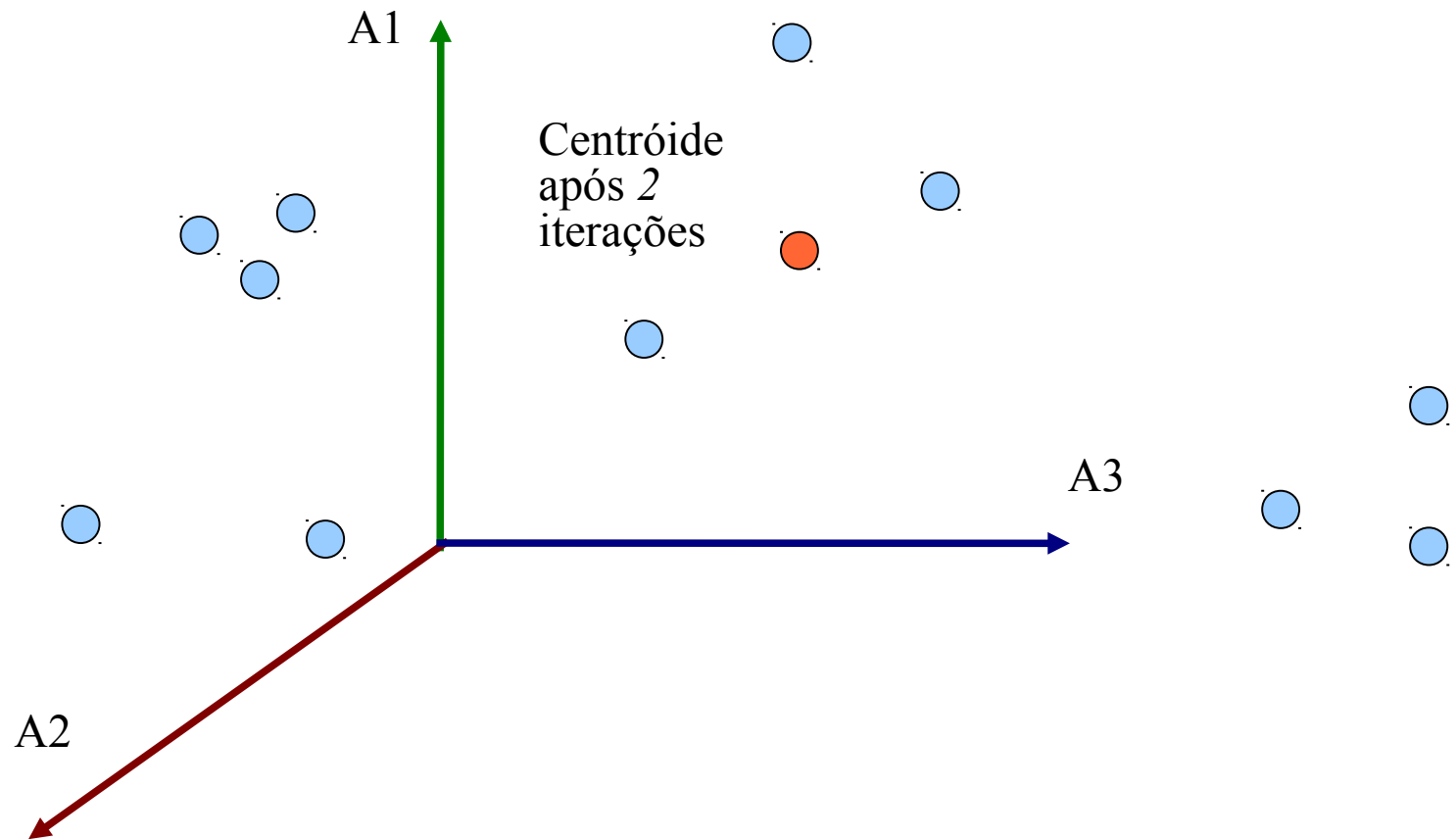
Algoritmo *k-means* (*k-médias*)



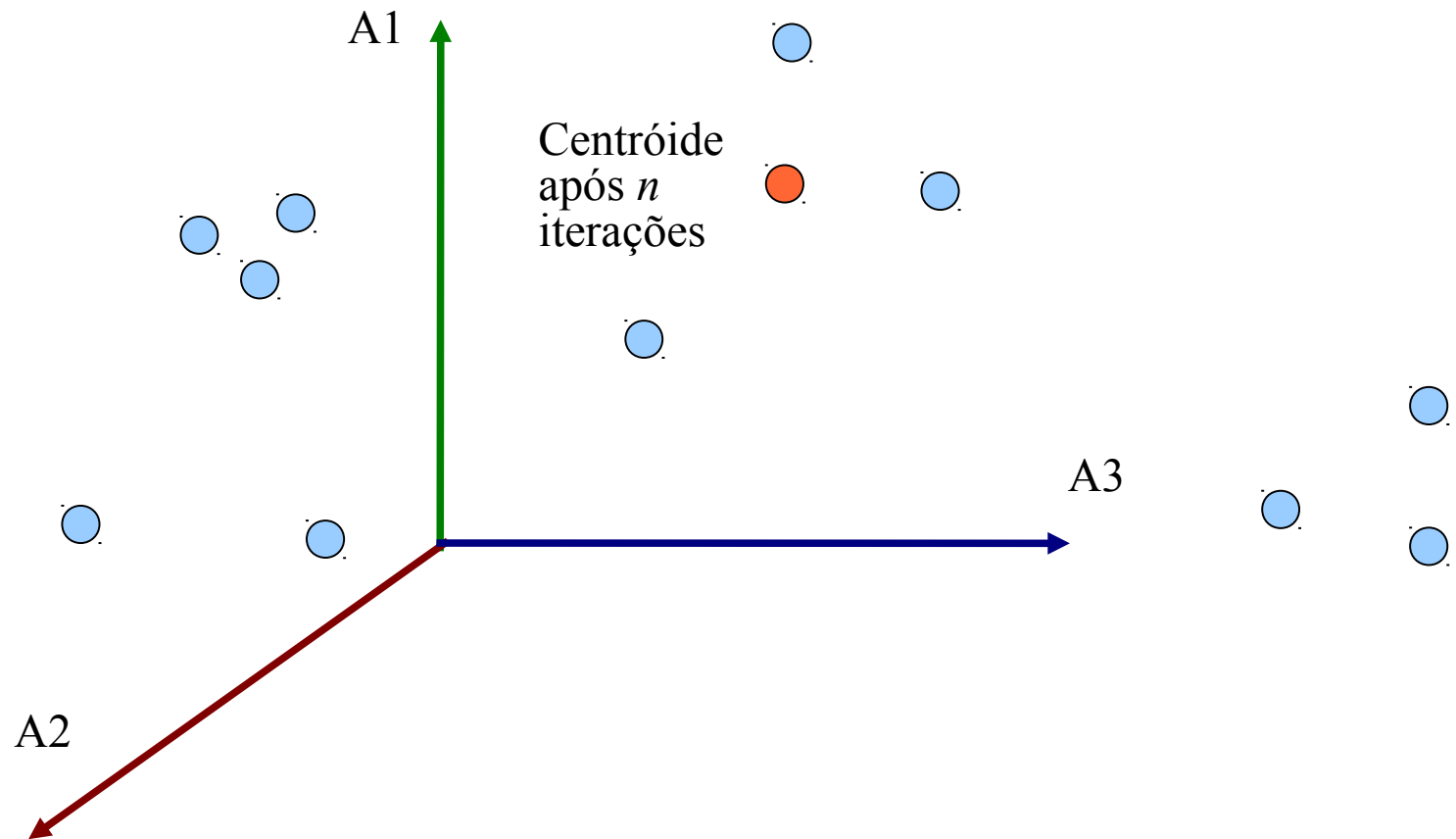
Algoritmo K-means



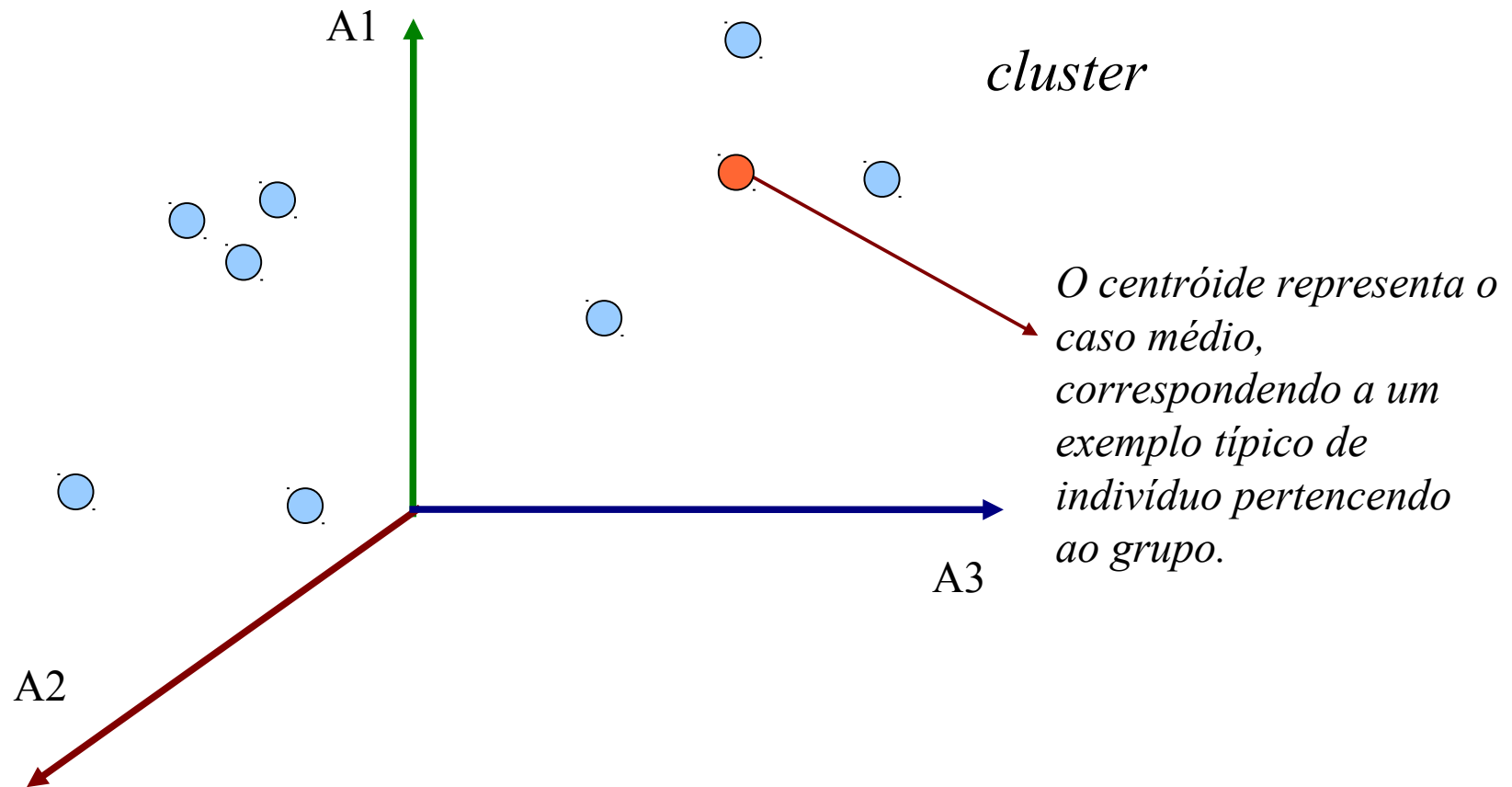
Algoritmo K-means



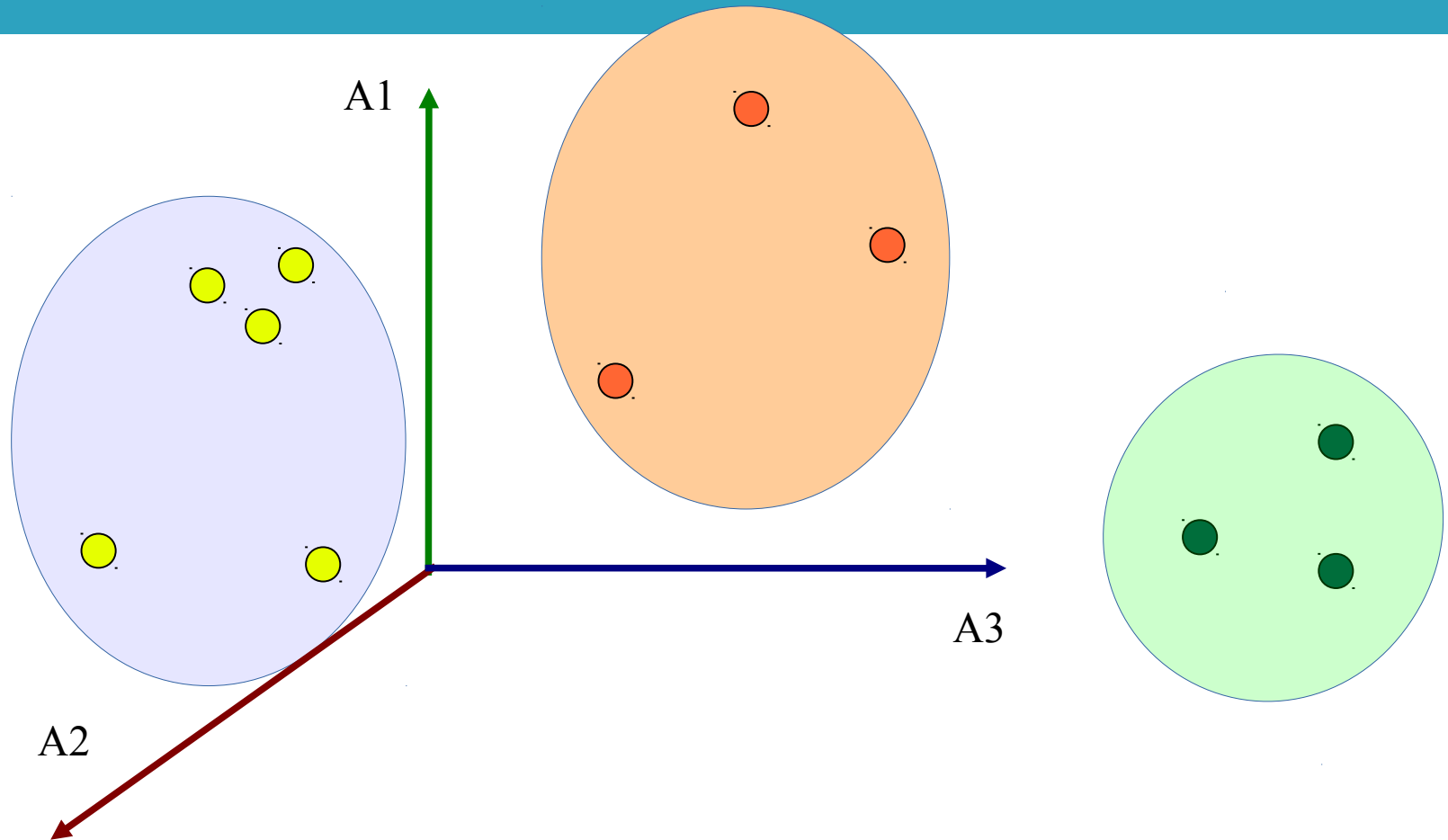
Algoritmo K-means



Algoritmo K-means



Formação de três *Clusters*



Algoritmo *k-means*

Seja K o número esperado de clusters:

Passo 1 - k pontos são criados como centróides

Passo 2- Cada instância é atribuída para a centróide mais próxima (utilizando uma métrica de distância).

Passo 3 - Para cada cluster, calcula-se um ponto médio (parte *means* do algoritmo).

Passo 4 - Estas tornam-se as novas centróides de cada cluster.

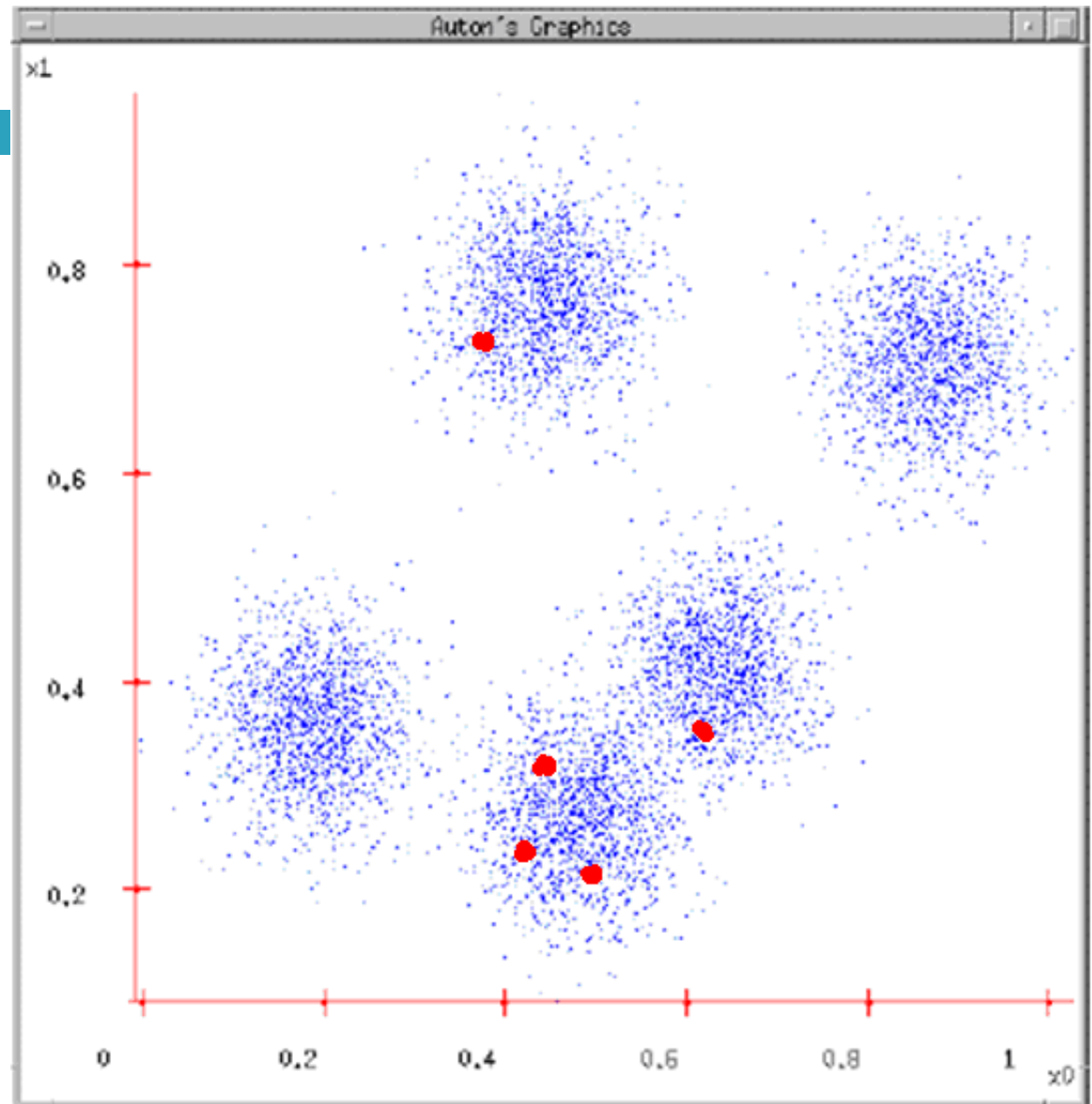
Passo 5- Repete-se este processo até que um único ponto de centróide seja atribuído a cada cluster (convergência).

Algoritmo K - Means

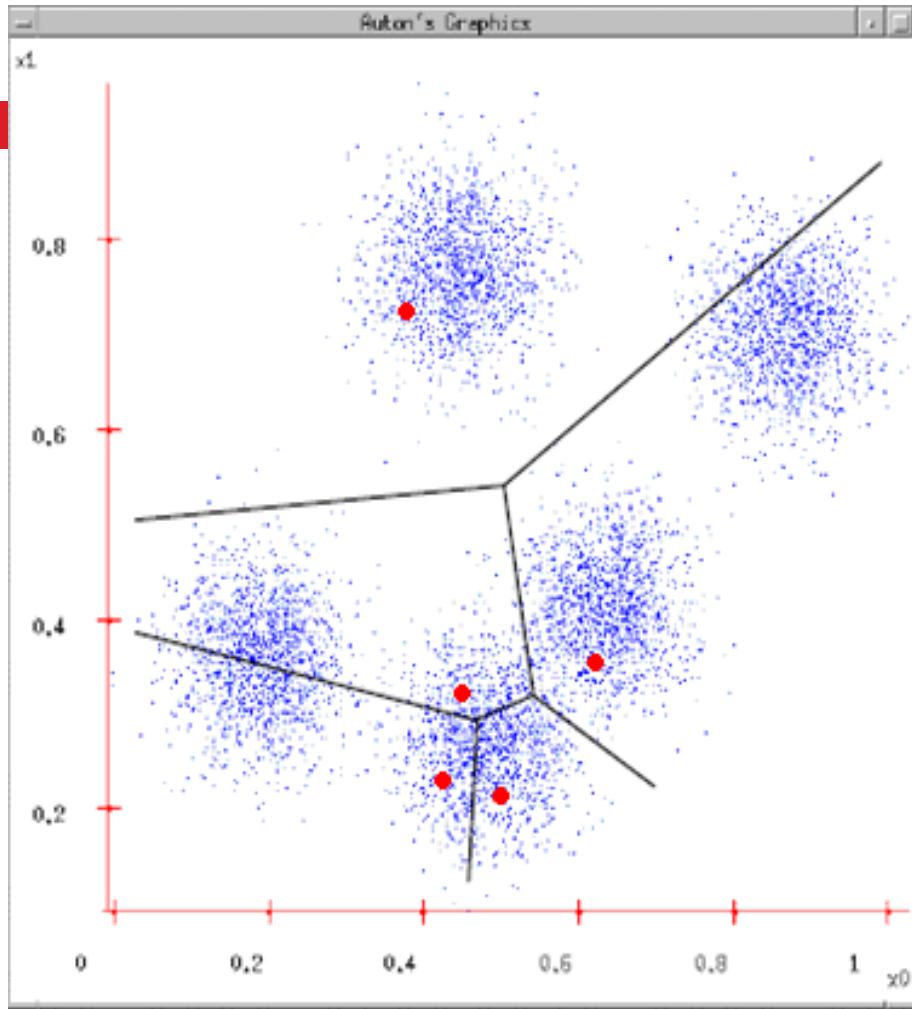
Inicialização

K-Means: agrupa as instâncias (pontos) em função da sua distância Euclidiana em relação a um centróide.

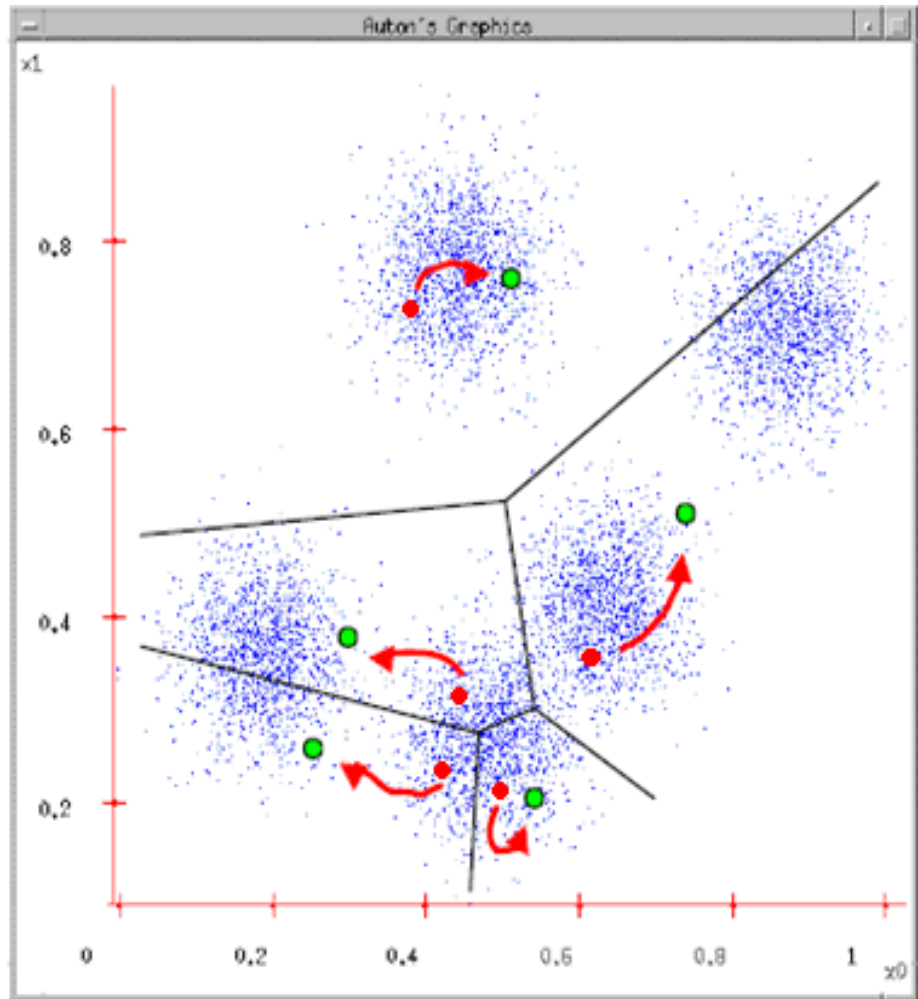
Centróide: pontos centrais dos grupos. Calculados aleatoriamente na primeira execução.



Algoritmo K - Means

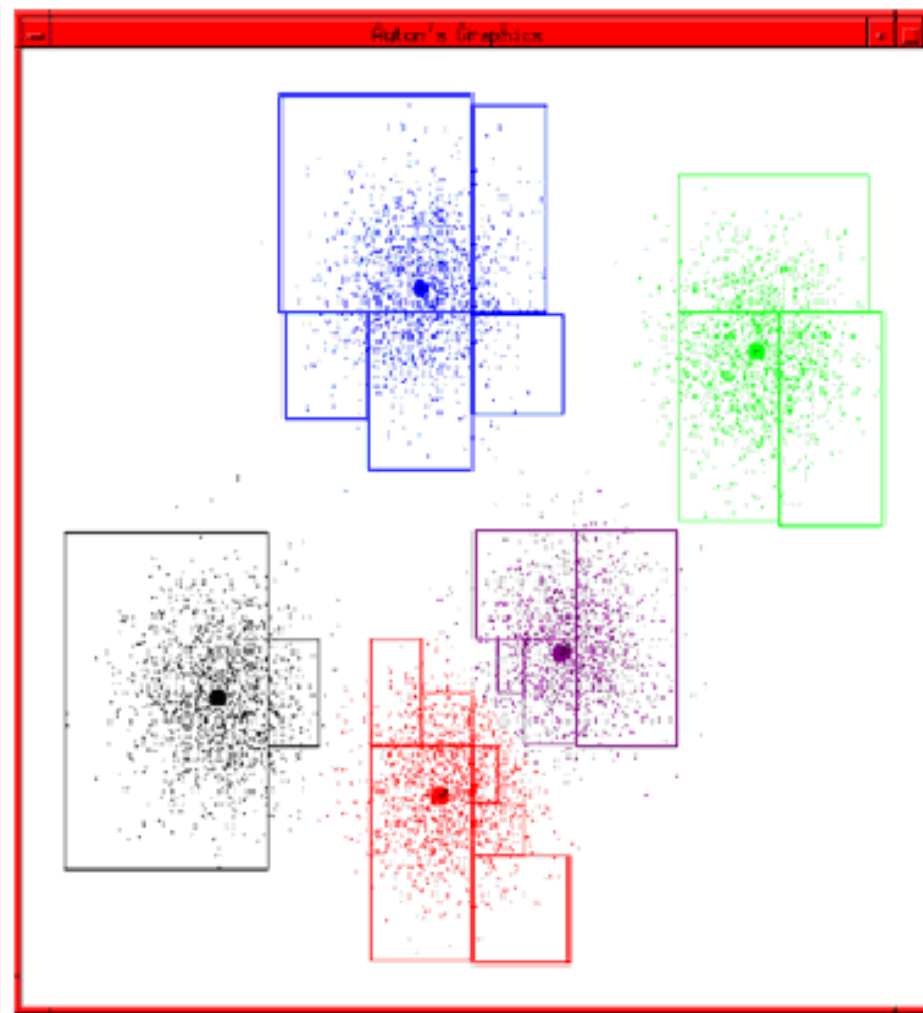
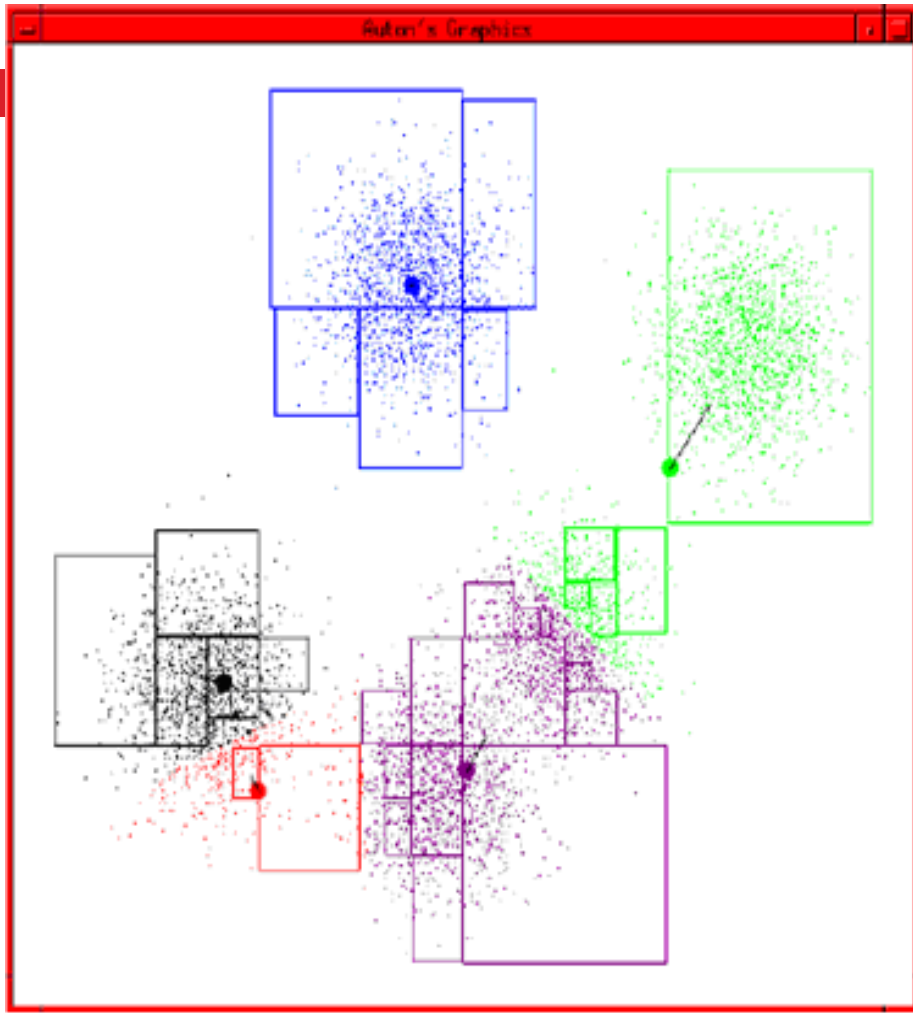


Definidos grupos iniciais em função da sua distância Euclidiana em relação a um centróide.



Recalculados os novos centróides (pontos médios).

Algoritmo K - Means



Etapas da análise de dados



1. Definição do domínio de conhecimento;
2. Seleção dos dados : construção do dataset;
3. Preparação e transformação: operações básicas realizados sobre os dados para remoção de ruídos e dados desnecessários. Busca a qualidade dos dados.
4. Aplicação de algoritmos
 - a) Realização de testes preliminares para determinação da técnica de aprendizagem a ser utilizada;
 - b) Construção de cenários em conjunto com o especialista;
 - c) Avaliação dos resultados pelo especialista.

Trabalhos desenvolvidos



- 1) **Clustering para construção de modelos de alunos** em processos de aprendizagem (Webber, 1999-2003)
- 2) **Clustering de processos de usinagem** (Rocha, 2005; Guerra, 2005)
- 3) **Clustering em ordens de produção paradas ou em atraso** (Kilder, 2007)
- 4) **Clustering em dados de vendas de artigos esportivos** (Bassotto, 2008)
- 5) **Clustering em bases de dados do Datasus** (Almeida, 2005; Salvadori, 2006; Todeschini, 2008; Schmitz e Webber, 2008, 2009, 2010)
- 6) **Clustering para formação de grupos de alunos** (Webber & Prado Lima, 2010)

Weka Explorer

Preprocess | Classify | **Cluster** | Associate | Select attributes | Visualize

Clusterer

Choose **EM** -I 100 -N -1 -S 100 -M 1.0E-6

Cluster mode

- ☒ Use training set
- ☐ Supplied test set
- ☐ Percentage split %
- ☐ Classes to clusters evaluation
-
- ☒ Store clusters for visualization

Result list (right-click for options)

21:54:58 - EM

Clusterer output


```

Normal Distribution. Mean = 6.9426 StdDev = 0.498
Attribute: sepalwidth
Normal Distribution. Mean = 3.1103 StdDev = 0.2952
Attribute: petallength
Normal Distribution. Mean = 5.8559 StdDev = 0.4626
Attribute: petalwidth
Normal Distribution. Mean = 2.1495 StdDev = 0.232
Attribute: class
Discrete Estimator. Counts = 1 1.02 31.04 (Total = 33.06)


Cluster: 3 Prior probability: 0.1446




Attribute: sepallength
Normal Distribution. Mean = 6.1304 StdDev = 0.2943
Attribute: sepalwidth
Normal Distribution. Mean = 2.8088 StdDev = 0.2361
Attribute: petallength
Normal Distribution. Mean = 5.0993 StdDev = 0.2462
Attribute: petalwidth
Normal Distribution. Mean = 1.8254 StdDev = 0.2152
Attribute: class
Discrete Estimator. Counts = 1 3.87 19.86 (Total = 24.73)
Clustered Instances
0      48 ( 32%)
1      50 ( 33%)
2      29 ( 19%)
3      23 ( 15%)
Log likelihood: -2.03504
    
```

Status

OK  x 0

AGRUP – INTERFACE PARA AGRUPAMENTO DE DADOS


Agrupamento de Dados

Sair
Ajudar

Seleção
Processamento
Visualização

Abrir Arquivo
Editar

Relação

Relação
broca_etapa_ra.arff

Instâncias
1221

Atributos
12

Atributos

Nr.	Nome
0	Instance_number
1	broca
2	etapa
3	furo
4	profundidade
5	medicao
6	ra
7	rq
8	rz
9	rmax
10	sm
11	Cluster

Estadísticas

Nome: ra
Sem valor: 0 (0%)
Distintos: 353
Tipo: Numérico
único: 102 (8%)

Estatística	Valores
Mínimo	0.4
Máximo	11.6
Média	4.01
Desvio Padrão	1.814

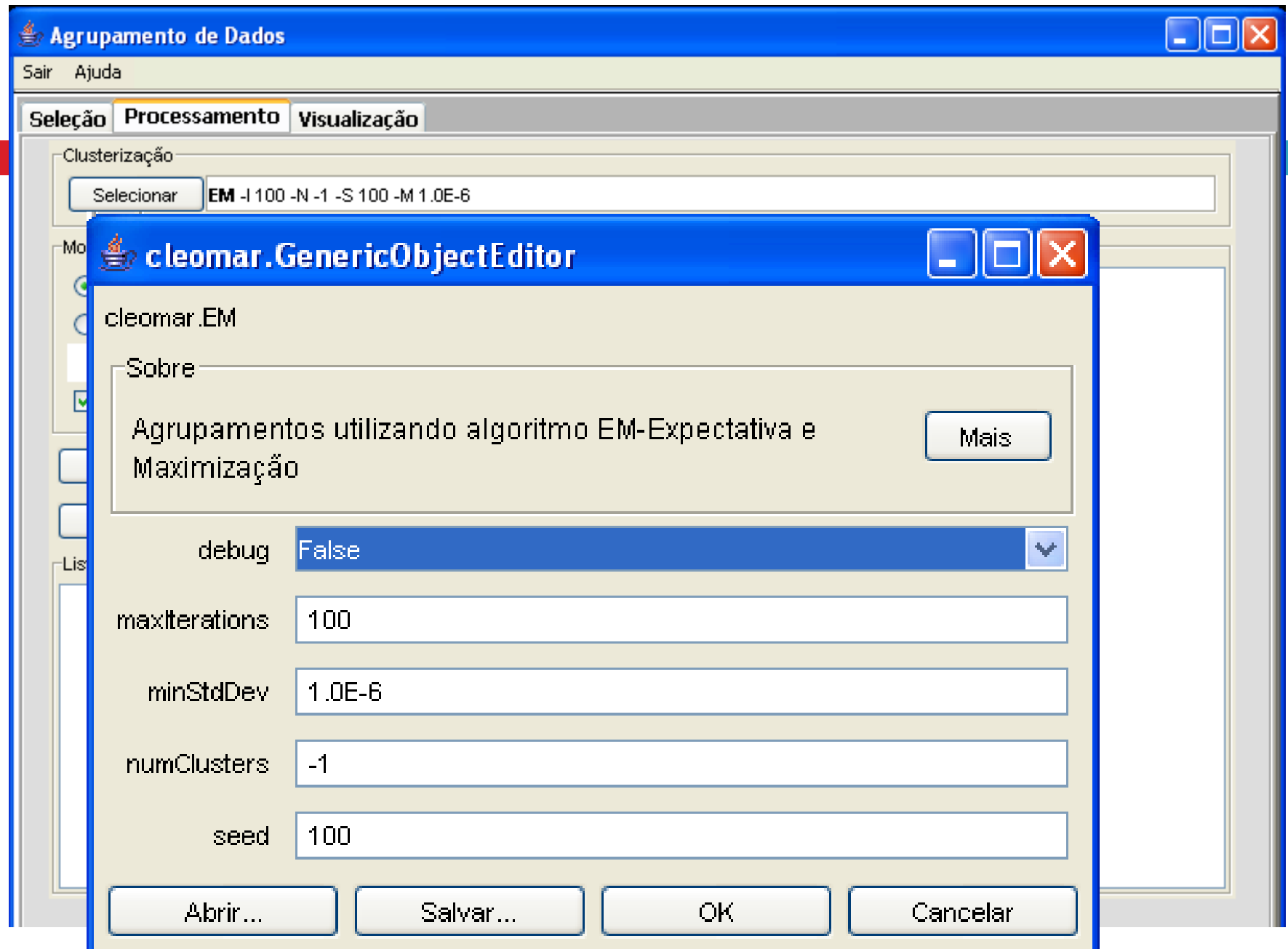
AGRUP – INTERFACE PARA AGRUPAMENTO DE DADOS

Visualizador												
Relation: bruto30_clustered												
No.	Instance_number Numeric	broca Nominal	etapa Nominal	furo Numeric	profundidade Nominal	medicao Nominal	ra Numeric	rq Numeric	rz Numeric	rmax Numeric	sm Numeric	Cluster Nominal
1	0.0	a	fluido	1.0	fim	primeira	3.82	4.8	17.8	21.4	158.0	cluster5
2	1.0	a	fluido	1.0	inicio	primeira	3.02	3.92	15.9	23.0	106.0	cluster5
3	2.0	a	fluido	1.0	meio	primeira	2.56	3.2	14.5	17.6	108.0	cluster5
4	3.0	a	fluido	1.0	fim	segunda	3.4	4.3	17.9	21.7	132.0	cluster5
5	4.0	a	fluido	1.0	inicio	segunda	2.44	3.14	11.2	13.4	122.0	cluster5
6	5.0	a	fluido	1.0	meio	segunda	3.26	4.08	17.1	18.7	181.0	cluster5
7	6.0	a	fluido	1.0	fim	terceira	3.36	4.22	17.0	22.4	129.0	cluster5
8	7.0	a	fluido	1.0	inicio	terceira	2.94	3.7	14.7	19.1	154.0	cluster5
9	8.0	a	fluido	1.0	meio	terceira	2.42	3.24	14.9	18.5	143.0	cluster5
10	9.0	a	fluido	1.0	fim	primeira	2.44	3.2	13.6	16.7	123.0	cluster5
11	10.0	a	fluido	1.0	inicio	primeira	2.22	2.86	11.8	14.9	103.0	cluster5
12	11.0	a	fluido	1.0	meio	primeira	2.7	3.46	14.6	19.6	118.0	cluster5
13	12.0	a	fluido	1.0	fim	segunda	2.7	3.54	15.3	20.5	115.0	cluster5
14	13.0	a	fluido	1.0	inicio	segunda	2.58	3.42	15.1	18.2	120.0	cluster5
15	14.0	a	fluido	1.0	meio	segunda	2.7	3.44	14.4	18.2	99.0	cluster5
16	15.0	a	fluido	1.0	fim	terceira	2.8	3.64	16.2	20.7	112.0	cluster5
17	16.0	a	fluido	1.0	inicio	terceira	2.9	3.66	15.6	16.1	107.0	cluster5
18	17.0	a	fluido	1.0	meio	terceira	2.68	3.4	14.8	18.4	135.0	cluster5
19	18.0	a	fluido	2.0	fim	primeira	3.8	4.7	18.4	25.6	140.0	cluster5
20	19.0	a	fluido	2.0	inicio	primeira	2.72	3.4	13.6	17.6	136.0	cluster5
21	20.0	a	fluido	2.0	meio	primeira	2.82	3.58	16.2	19.9	142.0	cluster5
22	21.0	a	fluido	2.0	fim	segunda	3.78	4.94	21.2	27.4	139.0	cluster5
23	22.0	a	fluido	2.0	inicio	segunda	2.72	3.56	14.2	20.3	137.0	cluster5
24	23.0	a	fluido	2.0	meio	segunda	2.44	3.12	13.3	16.7	150.0	cluster5

CTRL <A> para selecionar todas instâncias CTRL <C> para copiar

Desfazer OK Cancelar

AGRUP – INTERFACE PARA AGRUPAMENTO DE DADOS



AGRUP – INTERFACE PARA AGRUPAMENTO DE DADOS

The screenshot displays the 'Agrupamento de Dados' (Data Clustering) application window. The 'Processamento' (Processing) tab is active, showing the 'Clusterização' (Clustering) section. The 'Seleção' (Selection) button is highlighted, and the 'EM' algorithm is selected with parameters '-I 100 -N 6 -S 100 -M 1.0E-6'. The 'Modo Clusterização' (Clustering Mode) section shows 'Classes para avaliação de clusters' (Classes for cluster evaluation) selected. A list of variables is shown, with '(Nom) profundidade' selected. The 'Saída da Clusterização' (Clustering Output) section displays the 'Log likelihood: -2.65905' and the 'Class attribute: profundidade'. The 'Classes to Clusters' table is shown, with columns for cluster number, class number, and class name. The 'Select items' dialog box is open, showing a list of variables: 'Instance_number', 'broca', 'etapa', 'furo', 'profundidade', 'medicao', 'ra', and 'rq'. The 'Select' button is highlighted.

Agrupamento de Dados

Sair Ajuda

Seleção **Processamento** **Visualização**

Clusterização

Selecionar **EM** -I 100 -N 6 -S 100 -M 1.0E-6

Modo Clusterização

☐ Usar conjunto treinado

☒ Classes para avaliação de clusters

(Nom) profundidade

(Nom) profundidade

(Nom) medicao

(Num) ra

(Num) rq

(Num) rz

(Num) rmax

(Num) sm

(Nom) Cluster

09:19:39 - cleomar.EM

Saída da Clusterização

4 43 (4%)

5 143 (12%)

Log likelihood: -2.65905

Class attribute: profundidade

Classes to Clusters:

0	1	2	3	4	5	<-- assigned to cluster
70	7	6	248	15	65	inicio
66	18	13	256	14	44	meio
98	17	13	223	14	34	fim

Cluster 0 <-- fim

Cluster 1 <-- No class

Cluster 2 <-- No class

Cluster 3 <-- meio

Cluster 4 <-- No class

Cluster 5 <-- inicio

Select items

Instance_number

broca

etapa

furo

profundidade

medicao

ra

rq

Select Cancel

Clustering em bases de dados do Datasus



Exemplo de cenário analisado utilizando a base de dados do SIM
(Sistema de Informações de Mortalidade):

- Óbitos não fetais
- Causa do óbito: hepatite
- Abrangência: Rio Grande do Sul
- Ano: 2007
- Faixa etária: maiores de vinte anos

Clustering em bases de dados do Datasus

Exemplo de campos do SIM

- 1) CAUSABAS: determina a causa básica do óbito conforme a Classificação Internacional de Doença (CID): B18 (Hepatite Viral Crônica), B180 (Hepatite Viral Crônica B com Agente Delta), B181 (Hepatite Crônica Viral sem Agente Delta), B182 (Hepatite Viral Crônica C), B188 (Outras Hepatite Crônicas Virais), B189 (Hepatite Viral Crônica NE), B19 (Hepatite Viral NE), B190 (Hepatite Viral NE com Coma), B199 (Hepatite Viral NE sem Coma).
- 2) TIPOBITO: determina se o óbito é fetal ou não fetal, foi utilizado o filtro de óbito não fetal, ou seja, o campo deve estar preenchido com o numeral 2.
- 3) IDADE: determina a idade, este campo para estar dentro das especificações desejadas deve conter um valor maior ou igual a 420.

Clustering em bases de dados do Datasus

Campos selecionado para análise nos Clusters :

1)Idade

2)Sexo

3)Raça / Cor: 1 indica a raça Branca, 2 raça Preta, 3 raça Amarela, 4 raça Parda e o 5 raça Indígena.

4)Estado Civil: 1 indica solteiros, 2 casados, 3 viúvos, 4 separados judicialmente, 5 união consensual e 9 ignorado.

5)Escolaridade: anos de estudo concluídos: 1 representa nenhum ano concluído, 2 representa indivíduos com 1 a 3 anos, 3 representa 4 a 7 anos, 4 para 8 a 11 anos, 5 para 12 e mais anos, e o 9 para número de anos ignorados.

Clustering em bases de dados do Datasus

O melhor resultado foi obtido a partir da formação de dois clusters.

O cluster 1 possui a maioria dos indivíduos (74%) com as seguintes características:

- Pessoas do sexo masculino, com idade entre 36 e 75 anos, da raça branca, casadas e com escolaridade entre 1 e 7 anos.

O cluster de número 0 representam 26% dos indivíduos com características :

- Pessoas do sexo feminino, com idade entre aproximadamente 68 a 83 anos, da raça branca, viúvas e com escolaridade entre 4 e 7 anos ou 12 ou mais.

Atributo\Cluster	0	1
Idade	De 67,8 até 83,4	De 36,6 até 75,6
Sexo	Feminino	Masculino
Raça	Branca	Branca
Estado Civil	Viúvo	Casado
Escolaridade	De 4 a 7 anos e 12 ou mais	De 1 a 7 anos

Clustering em vendas de artigos esportivos

Exemplo 1 : acompanhamento da sazonalidade na venda de itens esportivos

Atributos analisados:

- **Grupo:** 102 (Calçados Femininos)
- **Tamanho:** 1, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44
- **Sub-grupo:** 101 (TENIS PASSEIO MASCULINO), 109 (TENIS FUTEBOL SALAO), 110 (TENIS INDOOR), 201 (TENIS PASSEIO FEMININOS), 202 (TENIS DE TENNIS FEM.), 203 (TENIS DE BASQUETE FEM.), 204 (TENIS DE VOLEI FEM.), 205 (TENIS DE HANDEBOL FEM.), 206 (TENIS DE CORRIDA FEM.), 207 (BOTAS FEM.), 208 (CHINELOS FEM.), 209 (SANDALIAS FEM.), 301 (AGASALHOS FEM.) , 535 (GOLF)
- **Marca:** 1 (NIKE), 2 (ADIDAS), 3 (FILA), 4 (RAINHA), 5 (REEBOK), 6 (PENALTY), 9 (MIZUNO), 10 (PUMA), 12 (UMBRO), 13 (DIADORA), 15 (ASICS), 18 (WILSON), 19 (NEW BALANCE), 22 (KAPPA), 30 (HEAD), 111 (OLYMPIKUS), 113 (DALPONTE), 148 (TIMBERLAND), 166 (BABOLAT), 167 (BULL TERRIER), 179 (QIX)
- **Ano:** 2005, 2006, 2007
- **Mês:** 1 (Janeiro), 2 (Fevereiro), 3 (Março), 4 (Abril), 5 (Maio), 6 (Junho), 7 (Julho), 8 (Agosto), 9 (Setembro), 10 (Outubro), 11 (Novembro), 12 (Dezembro)

Clustering em vendas de artigos esportivos

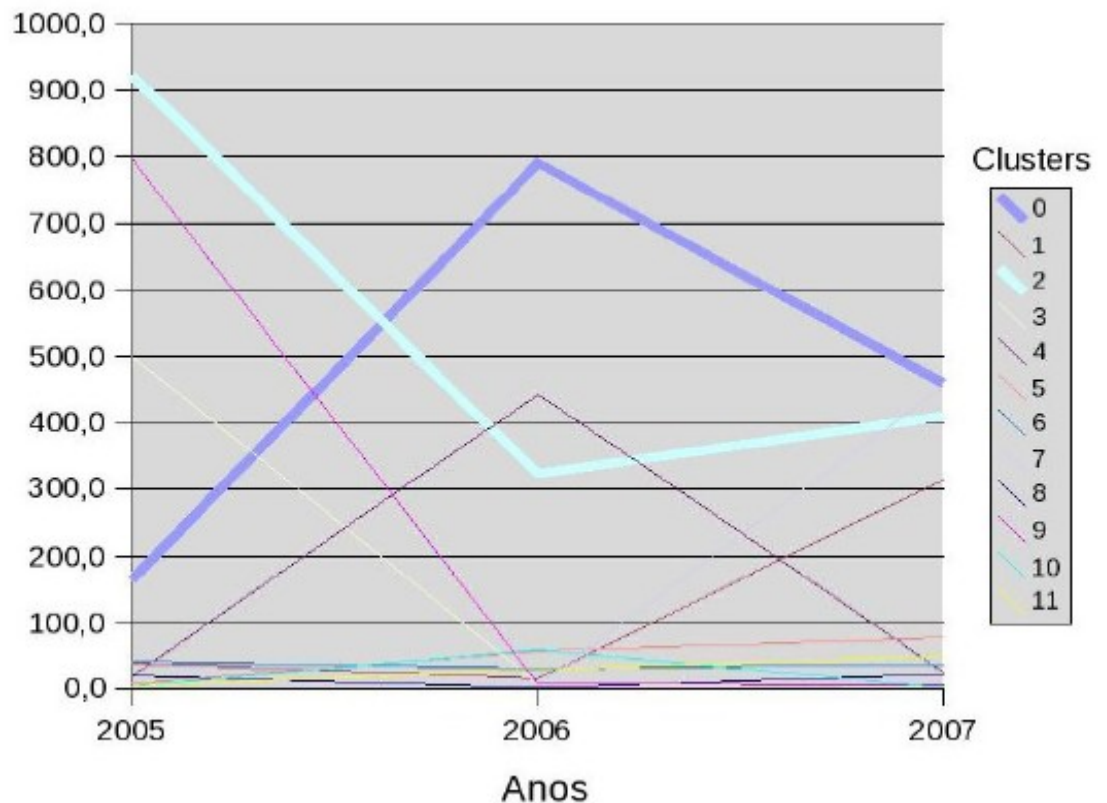
Clusters formados:

Cluster	Quantidade de Instâncias	Probabilidade
0	1377	22,00%
1	319	5,00%
2	1719	28,00%
3	493	8,00%
4	366	6,00%
5	145	2,00%
6	112	2,00%
7	475	8,00%
8	42	1,00%
9	1011	16,00%
10	50	1,00%
11	48	1,00%
Log likelihood: -7.77429		

Clustering em vendas de artigos esportivos

Número de instâncias em cada cluster em cada ano analisado:

Cluster	Ano		
	2005	2006	2007
0	164,1	790,6	460,9
1	40,8	16,1	314,4
2	922,7	323,8	411,2
3	501,7	9,6	4,4
4	21,7	444,0	26,4
5	9,6	58,8	78,8
6	43,2	33,0	39,4
7	4,3	6,2	455,4
8	20,4	3,5	22,1
9	797,6	8,5	7,2
10	4,9	60,1	2,3
11	5,2	30,8	49,6

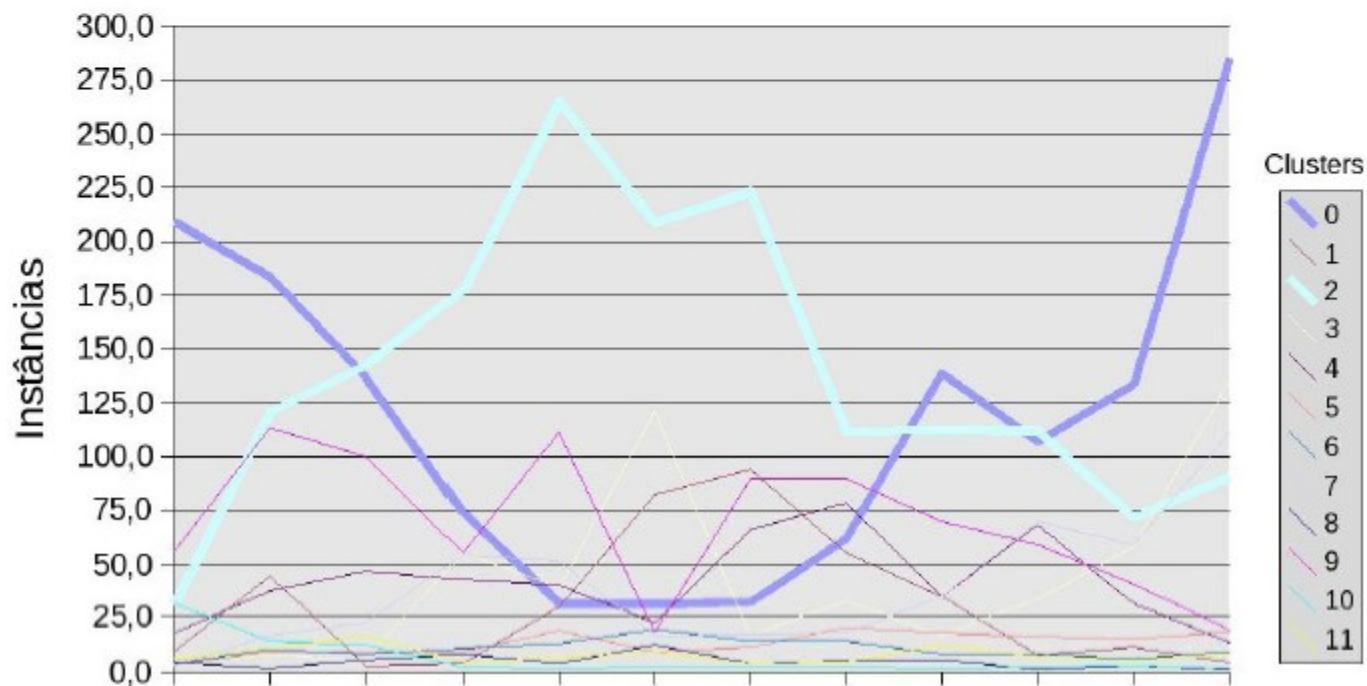


Que produtos estão agrupados em cada cluster?

Distribuição ao longo dos meses

Cluster	Meses											
	Jan 1	Fev 2	Mar 3	Abr 4	Mai 5	Jun 6	Jul 7	Ago 8	Set 9	Out 10	Nov 11	Dez 12
0	209,2	183,4	136,6	73,9	31,6	31,3	32,4	62,0	138,5	107,0	133,2	285,5
1	9,4	44,8	2,5	3,8	30,2	82,4	93,8	55,4	34,8	7,8	11,2	4,3
2	31,1	120,9	142,4	177,6	264,9	208,8	223,0	111,4	112,6	111,8	71,6	90,8
3	3,2	8,1	4,9	55,2	39,4	120,9	16,9	33,0	16,8	33,9	57,9	134,7
4	18,0	37,5	46,8	43,1	40,5	22,6	66,2	78,3	35,0	68,1	31,5	13,6
5	3,0	9,0	9,0	8,9	18,8	8,9	11,1	20,4	17,8	16,1	15,0	18,3
6	4,0	10,0	8,0	11,0	13,2	19,2	14,2	14,1	8,1	8,0	6,0	9,0
7	2,9	16,5	22,7	54,5	51,6	17,0	17,1	17,9	34,7	70,1	59,3	110,5
8	4,2	1,4	5,6	8,2	4,0	12,4	3,6	5,0	5,3	1,1	2,9	1,3
9	56,0	113,6	100,0	55,2	111,1	18,3	89,7	90,0	69,7	58,9	40,7	19,2
10	32,0	13,8	12,3	2,2	1,5	2,0	2,4	2,1	1,6	2,2	2,7	1,6
11	6,0	11,0	16,4	4,5	6,3	9,3	3,8	4,6	12,3	7,1	5,1	8,3

Figura 3.4: Clusters por Meses no Cenário 1



Clustering em vendas de artigos esportivos



Cluster 0

sub-grupo 201: Tenis de Passeio Femininos

Marcas: Olympikus, Nike, Diadora, Mizuno e Asics.

Tamanhos: de 34 a 38

Tendência maior de vendas dos produtos : Setembro e Março

Cluster 2

sub-grupos 201 e 202: Tenis de Passeio Femininos e Tenis de Tennis Femininos

Marcas: Nike, Asics, Puma, Olympikus e Diadora.

Tamanhos: de 35 a 39.

Tendência maior de vendas dos produtos: Março a Julho

Clustering em vendas de artigos esportivos

Visualização no WEKA

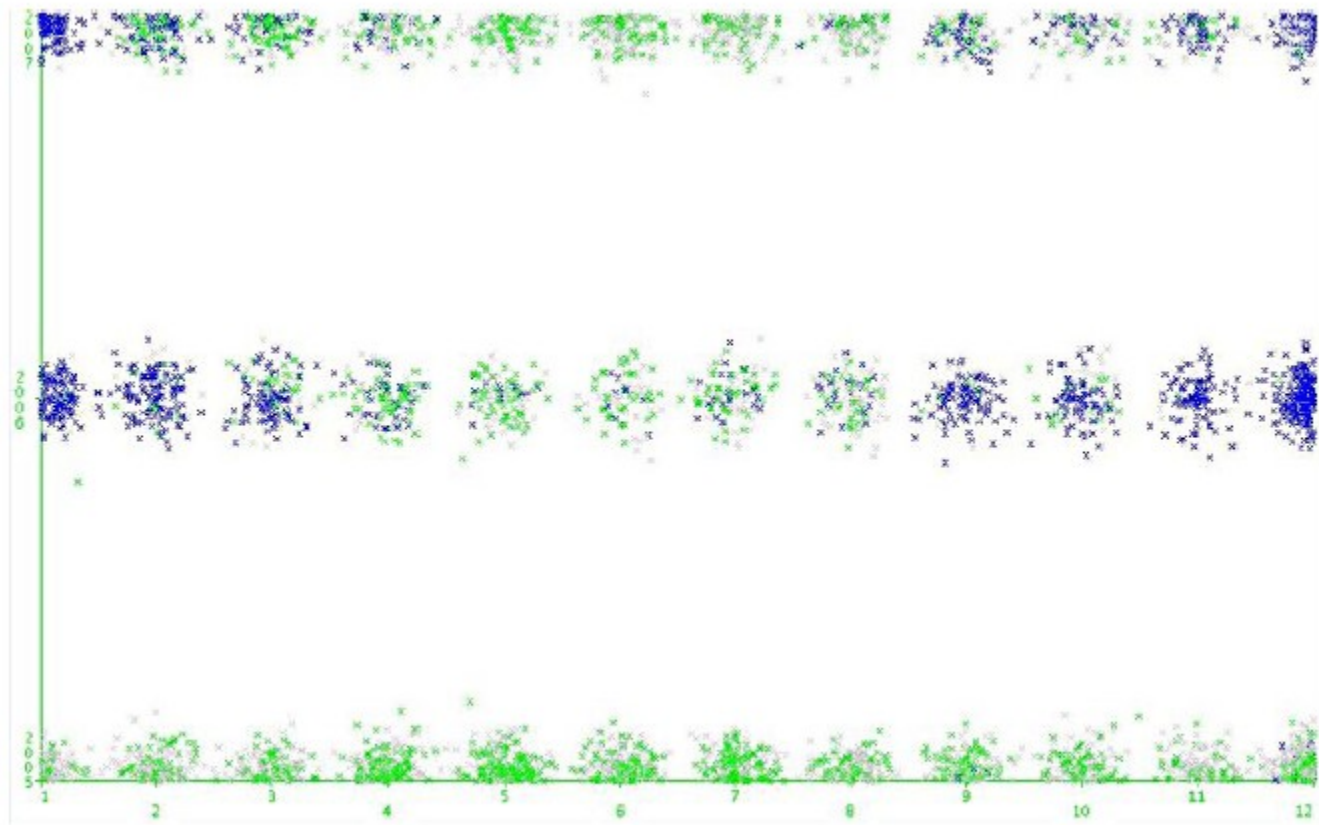
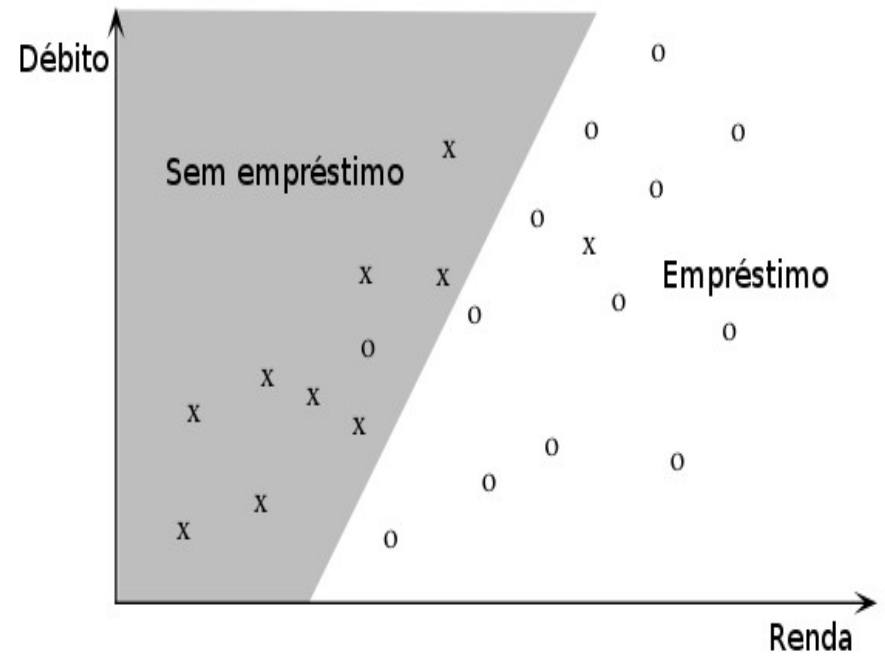
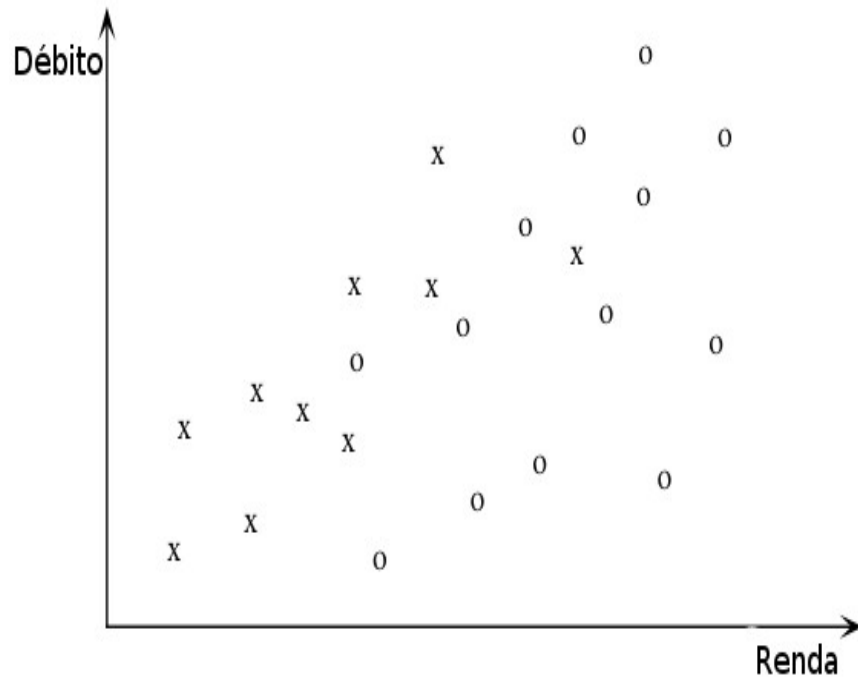


Figura 3.6: Gráfico Weka Meses por Anos no Cenário 1

Problemas de Classificação

Classificação



Etapas da Classificação

- Fase de Treinamento

- Ajuste de parâmetros
 - Utiliza conjunto de dados de treinamento

Fase de Testes

- Avalia o desempenho para novos dados em uma fase de validação
 - Utiliza conjunto de dados de teste
 - Desempenho depende da representatividade dos exemplos
 - Aprendizado é mais confiável quando exemplos de treinamento seguem uma distribuição semelhante à dos exemplos de teste

Paradigmas de Classificação

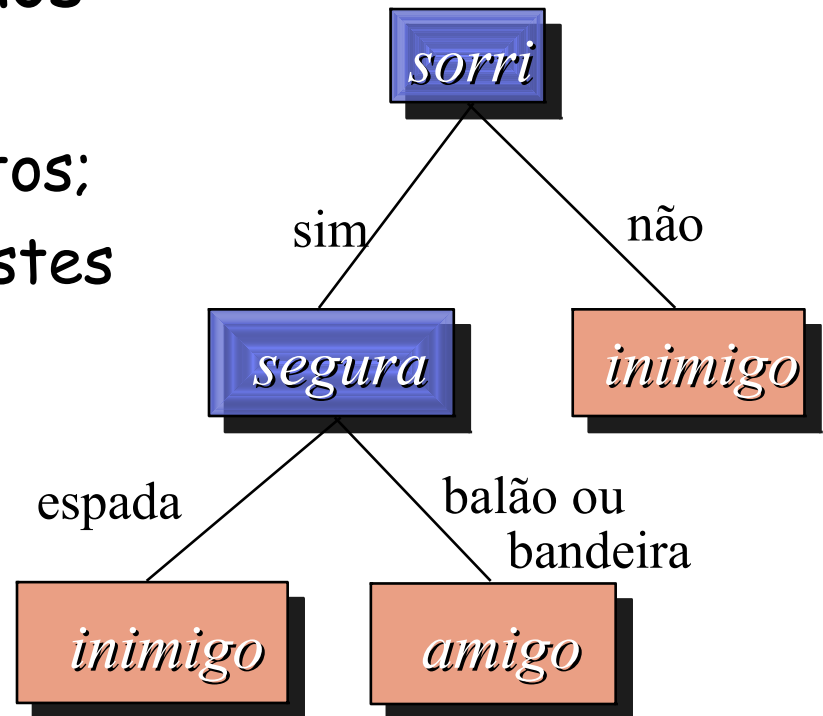
- Simbólico
- Estatístico e probabilístico
- Conexionista
- Evolucionário

Paradigma Simbólico

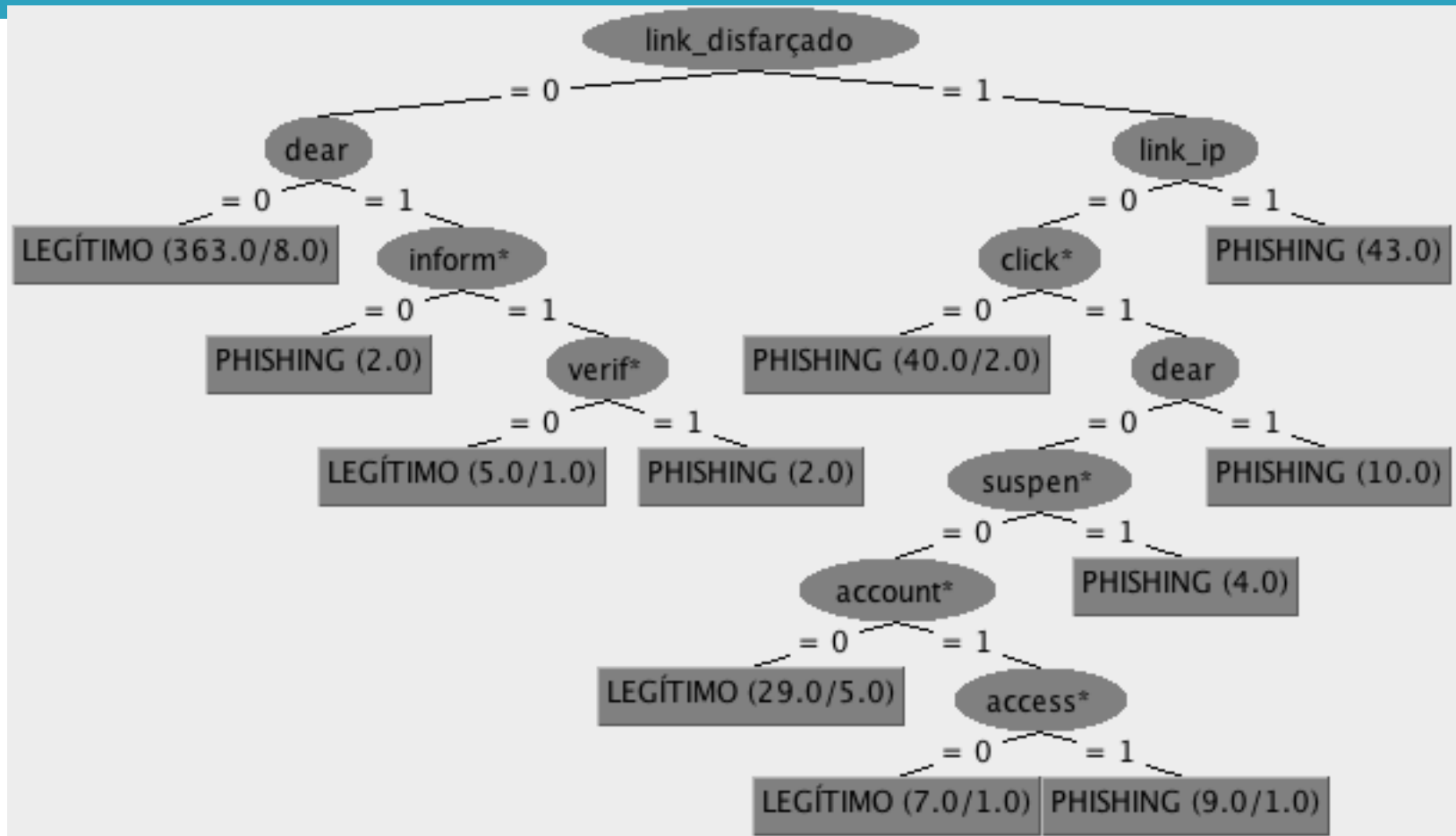
- Constrói representações simbólicas de um conceito através da análise de seus exemplos e contra-exemplos
- Representações simbólicas geralmente assumem a forma de:
 - Expressão lógica
 - Árvore de decisão
 - Regras de produção
 - Rede semântica

Árvores de Decisão

- Organizam informações em estrutura composta de nós e ramificações
 - Nós: testes sobre atributos;
 - Ramos: resultados dos testes



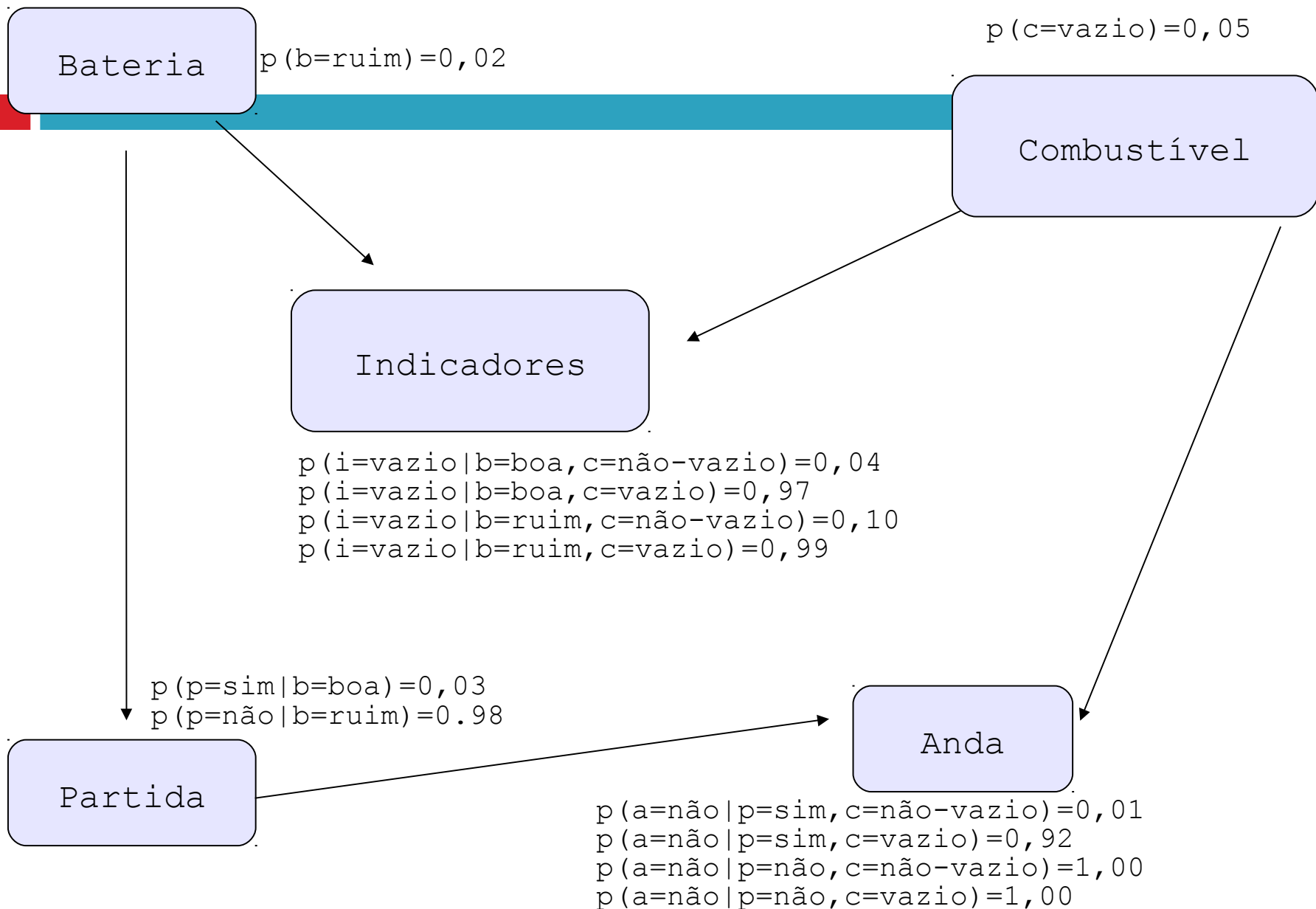
Exemplo de AD



Paradigma Estatístico

- Constrói um modelo estatístico do problema, geralmente utilizando a regra de Bayes.
- Aprendizagem de estruturas:
 - Redes Bayesianas
 - Redes Bayesianas Ingênuas

Exemplo de RB

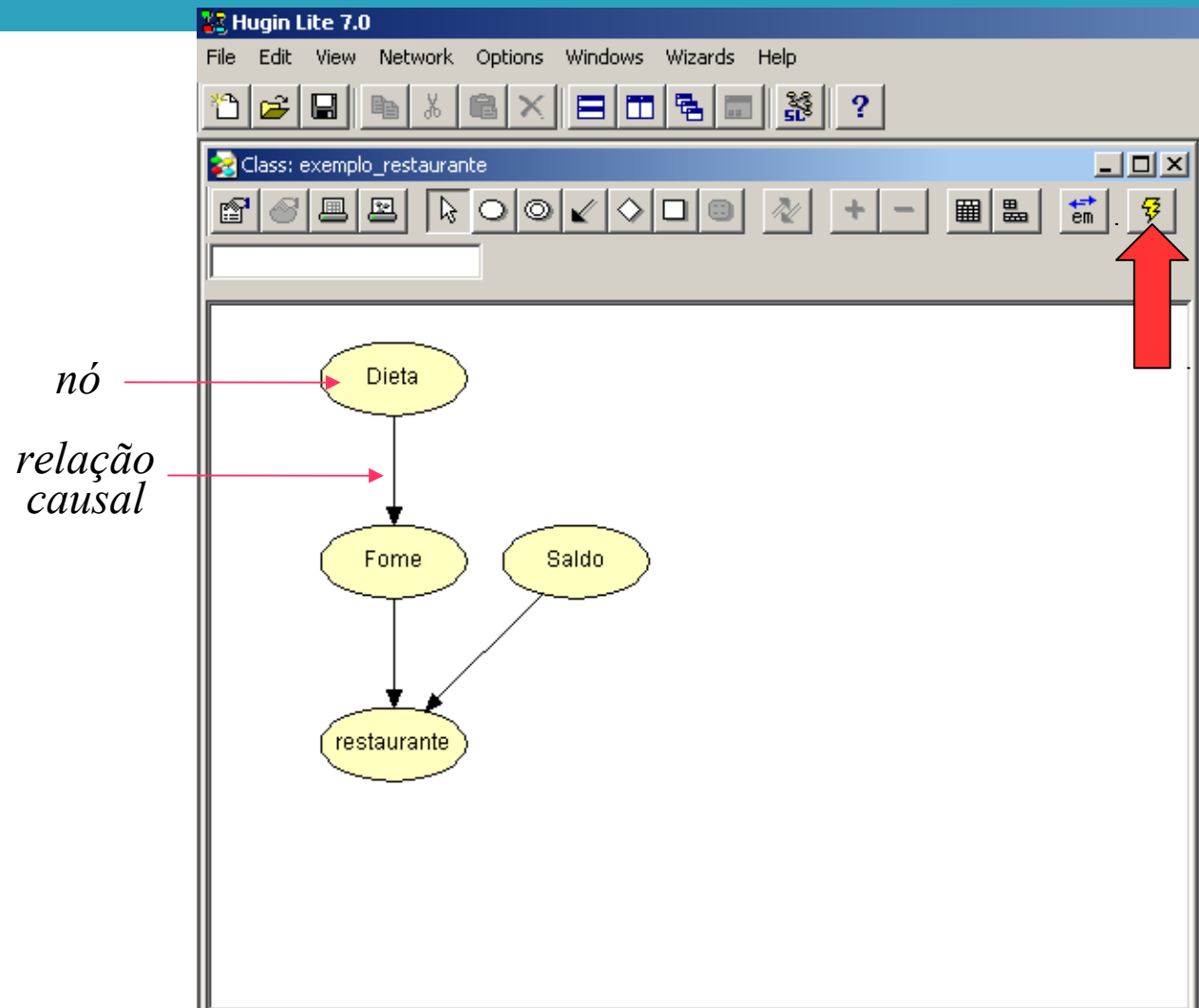


Software Hugin

- Software para construção, teste e depuração de Redes Bayesianas.
- Hugin:
 - <http://www.hugin.com/>
- Hugin Lite : versão demo free
 - http://www.hugin.com/Products_Services/Products/Demo/Lite/

Software Hugin - Exemplo

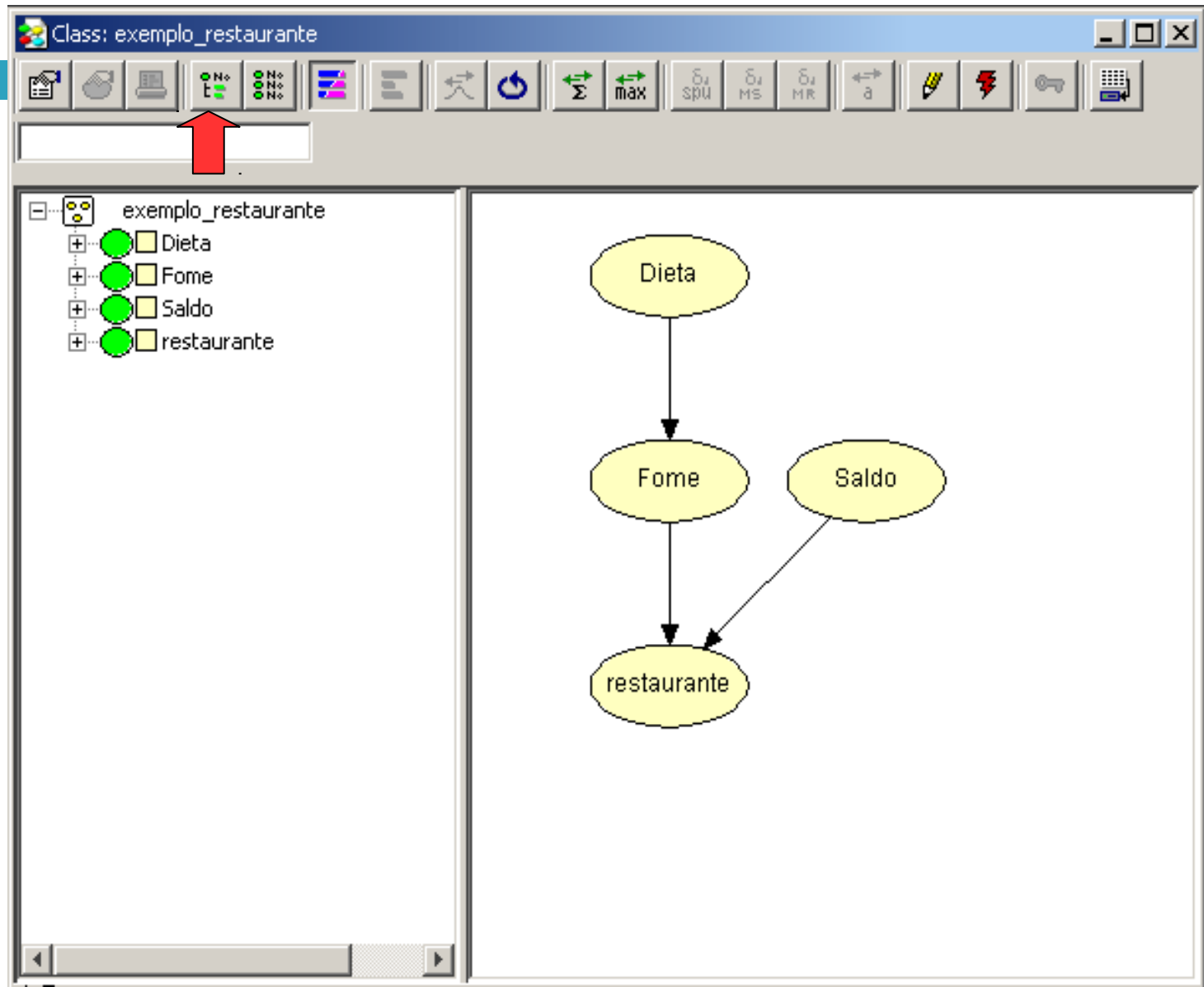
Modo de
edição



Software Hugin - Exemplo

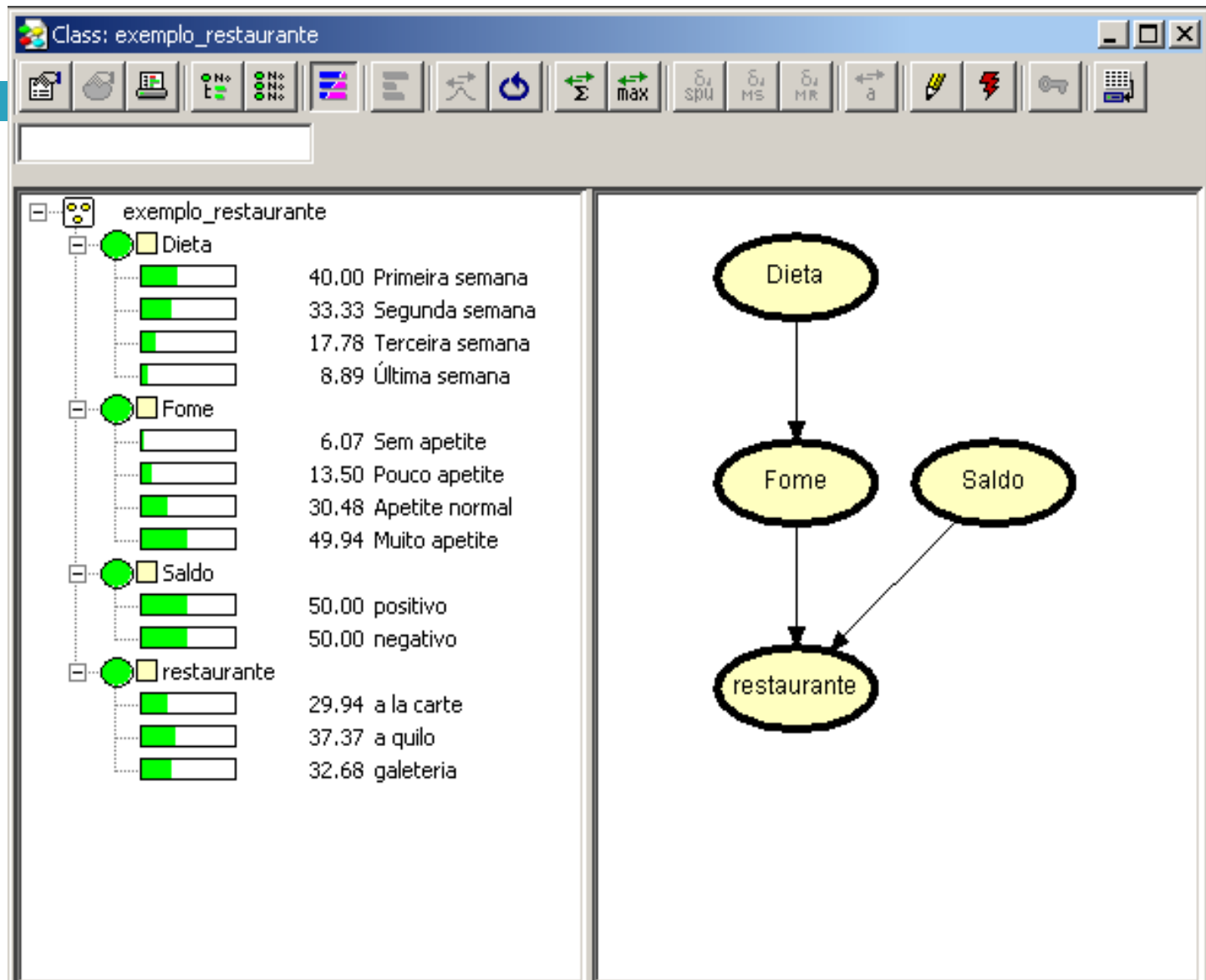
Modo de
teste e
depuração

Expansão
dos nós



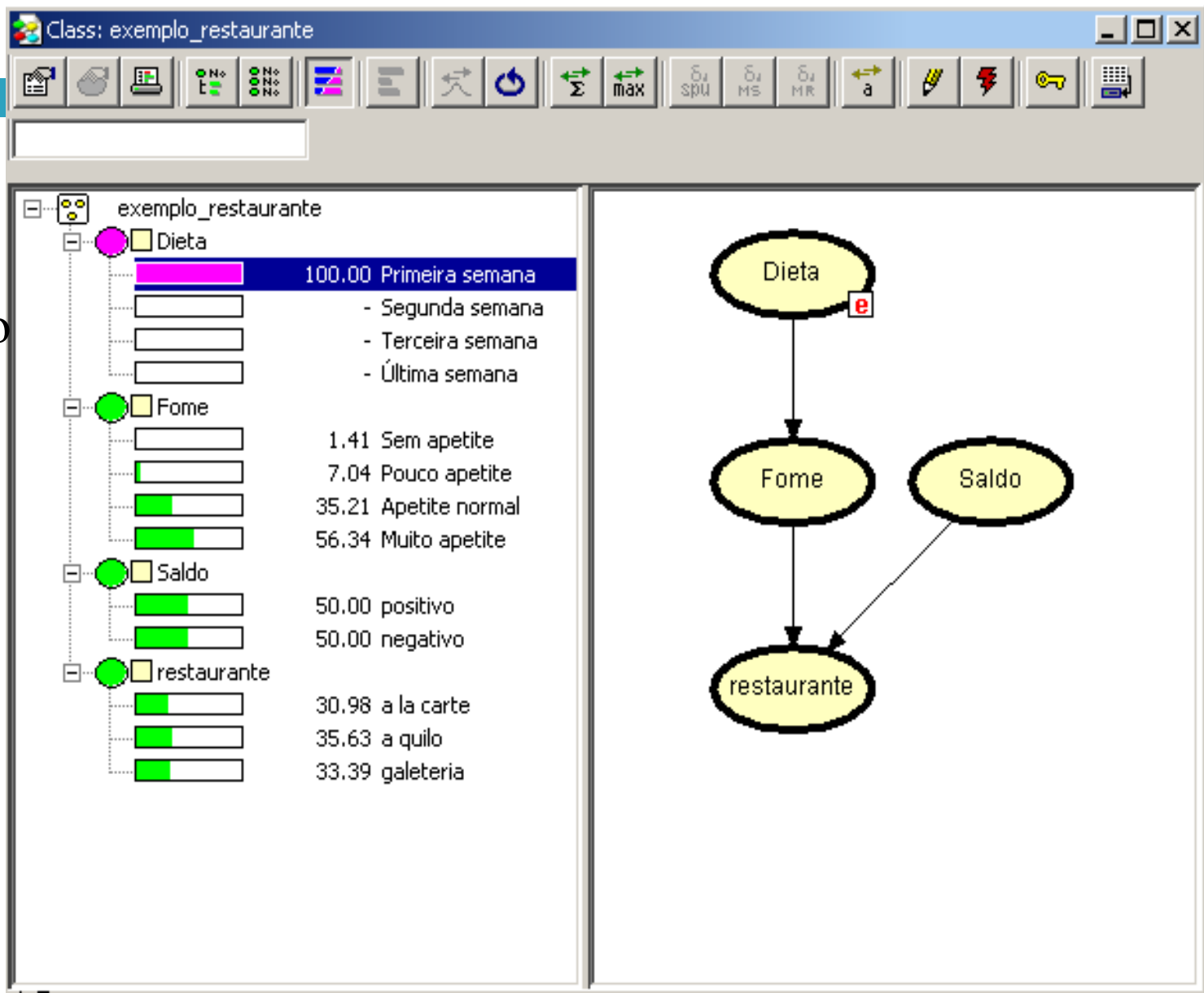
Software Hugin - Exemplo

Modo de
teste e
depuração



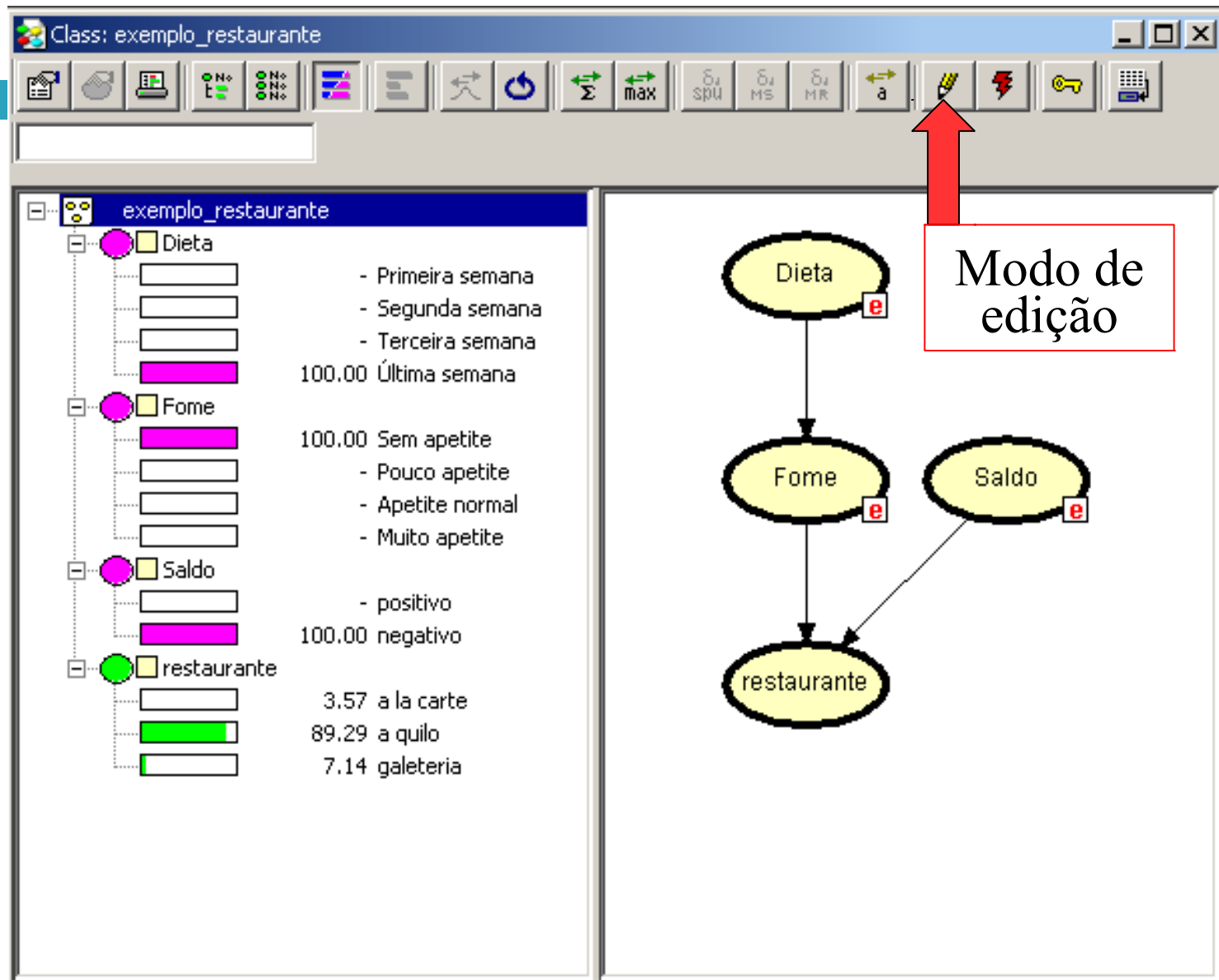
Software Hugin - Exemplo

Modo de
teste e
depuração



Software Hugin - Exemplo

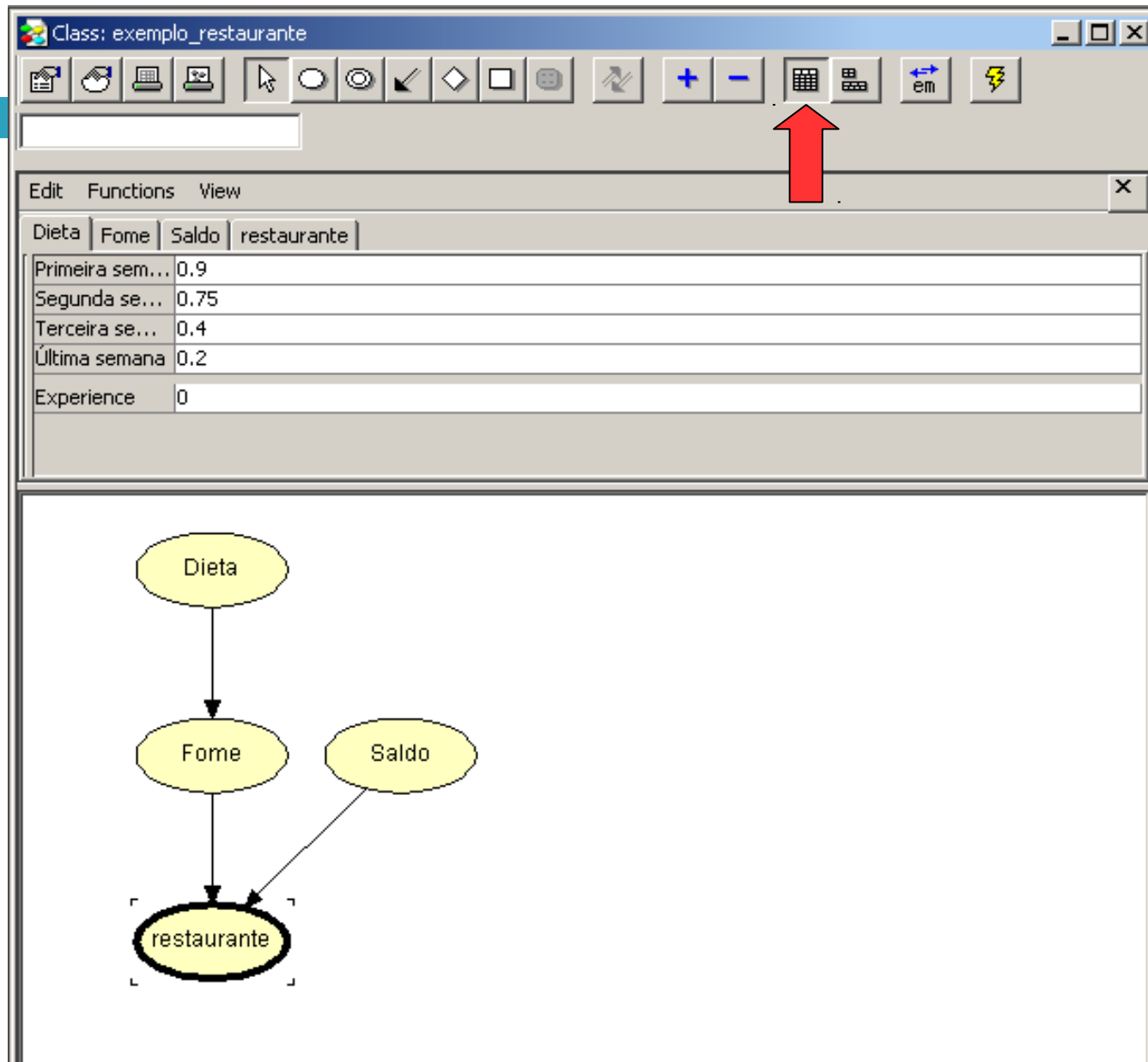
Modo de
teste e
depuração



Software Hugin - Exemplo

Modo de
edição

Probabilidades
informadas



Software Hugin - Exemplo

Modo de
edição

Probabilidades
informadas

Class: exemplo_restaurante

Edit Functions View

Dieta Fome Saldo restaurante

Dieta	Primeira semana	Segunda semana	Terceira semana	Última semana
Sem appetite	0.02	0.1	0.2	0.21
Pouco appetite	0.1	0.25	0.3	0.35
Apetite normal	0.5	0.4	0.5	0.4
Muito appetite	0.8	0.75	0.7	0.6
Experience	0	0	0	0

Diagrama de rede bayesiana:

```
graph TD; Dieta([Dieta]) --> Fome([Fome]); Saldo([Saldo]) --> restaurante([restaurante]); Fome --> restaurante;
```

Software Hugin - Exemplo

Modo de
edição

Probabilidades
informadas

Class: exemplo_restaurante

Edit Functions View

Dieta Fome Saldo restaurante

Fome	Sem appetite		Pouco appetite		Apetite normal		Muito appetite	
Saldo	positivo	negativo	positivo	negativo	positivo	negativo	positivo	negativo
a la carte	0.090909	0.035714	0.333333	0.333333	0.333333	0.333333	0.409091	0.1875
a quilo	0.606061	0.892857	0.333333	0.333333	0.333333	0.333333	0.227273	0.5
galeria	0.30303	0.071429	0.333333	0.333333	0.333333	0.333333	0.363636	0.3125

Diagrama de rede bayesiana:

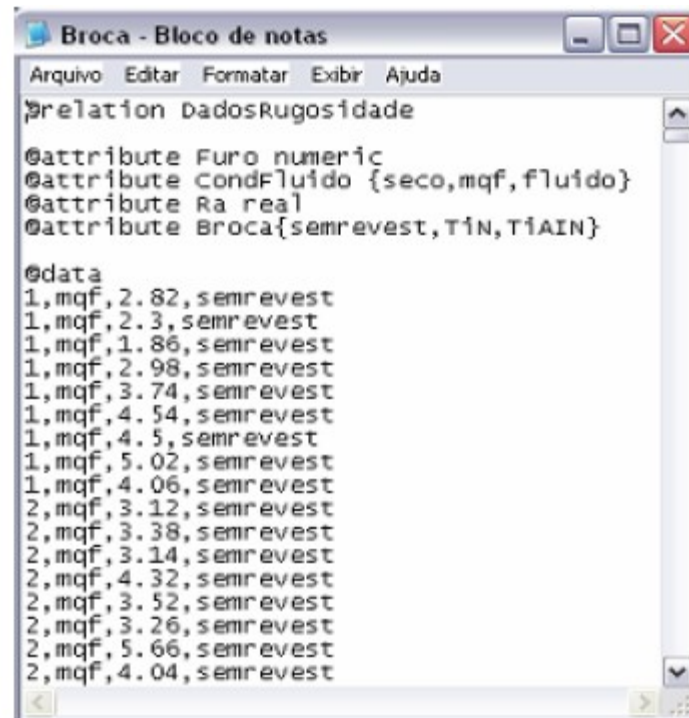
```
graph TD; Dieta([Dieta]) --> Fome([Fome]); Saldo([Saldo]) --> restaurante([restaurante]); Fome --> restaurante;
```

Ferramenta Weka

- WEKA: Waikato Environment for Knowledge Analysis
- Algoritmos para:
 - preparação de dados.
 - aprendizagem de máquina através da mineração.
 - para validação de resultados.

Ferramenta Weka

- Preparação dos Dados
 - Abertura de Arquivos ARFF



```
Arquivo  Editar  Formatar  Exibir  Ajuda
relation DadosRugosidade

@attribute Furo numeric
@attribute CondFluido {seco,mqf,fluido}
@attribute Ra real
@attribute Broca{semrevest,TiN,TiAIN}

@data
1,mqf,2.82,semrevest
1,mqf,2.3,semrevest
1,mqf,1.86,semrevest
1,mqf,2.98,semrevest
1,mqf,3.74,semrevest
1,mqf,4.54,semrevest
1,mqf,4.5,semrevest
1,mqf,5.02,semrevest
1,mqf,4.06,semrevest
2,mqf,3.12,semrevest
2,mqf,3.38,semrevest
2,mqf,3.14,semrevest
2,mqf,4.32,semrevest
2,mqf,3.52,semrevest
2,mqf,3.26,semrevest
2,mqf,5.66,semrevest
2,mqf,4.04,semrevest
```

Ferramenta Weka

■ Preparação dos Dados

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. The 'Filter' section shows 'Discretize -R first-last' applied. The 'Current relation' is 'QueryResult-weka.filters.supervised.attribute.Discretize-Rf...' with 1221 instances and 11 attributes. The 'Attributes' list on the left includes 'Broca', 'Etapas', 'Furo', 'Medida', 'Profundi', 'Ra', 'Rq', 'Rz', 'RMax', 'Sm', and 'CondFluido'. The 'Selected attribute' section shows 'CondFluido' with a nominal type and 3 distinct values. A table below shows the distribution: 'mqf' (162), 'Fluido' (681), and 'seco' (378). A bar chart at the bottom visualizes this distribution with blue bars for each category. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Edit... Save...

Filter

Choose Discretize -R first-last Apply

Current relation

Relation: QueryResult-weka.filters.supervised.attribute.Discretize-Rf...
Instances: 1221 Attributes: 11

Attributes

All None Invert

No.	Name
1	Broca
2	Etapas
3	Furo
4	Medida
5	Profundi
6	Ra
7	Rq
8	Rz
9	RMax
10	Sm
11	CondFluido

Remove

Selected attribute

Name: CondFluido
Missing: 0 (0%) Distinct: 3 Type: Nominal
Unique: 0 (0%)

Label	Count
mqf	162
Fluido	681
seco	378

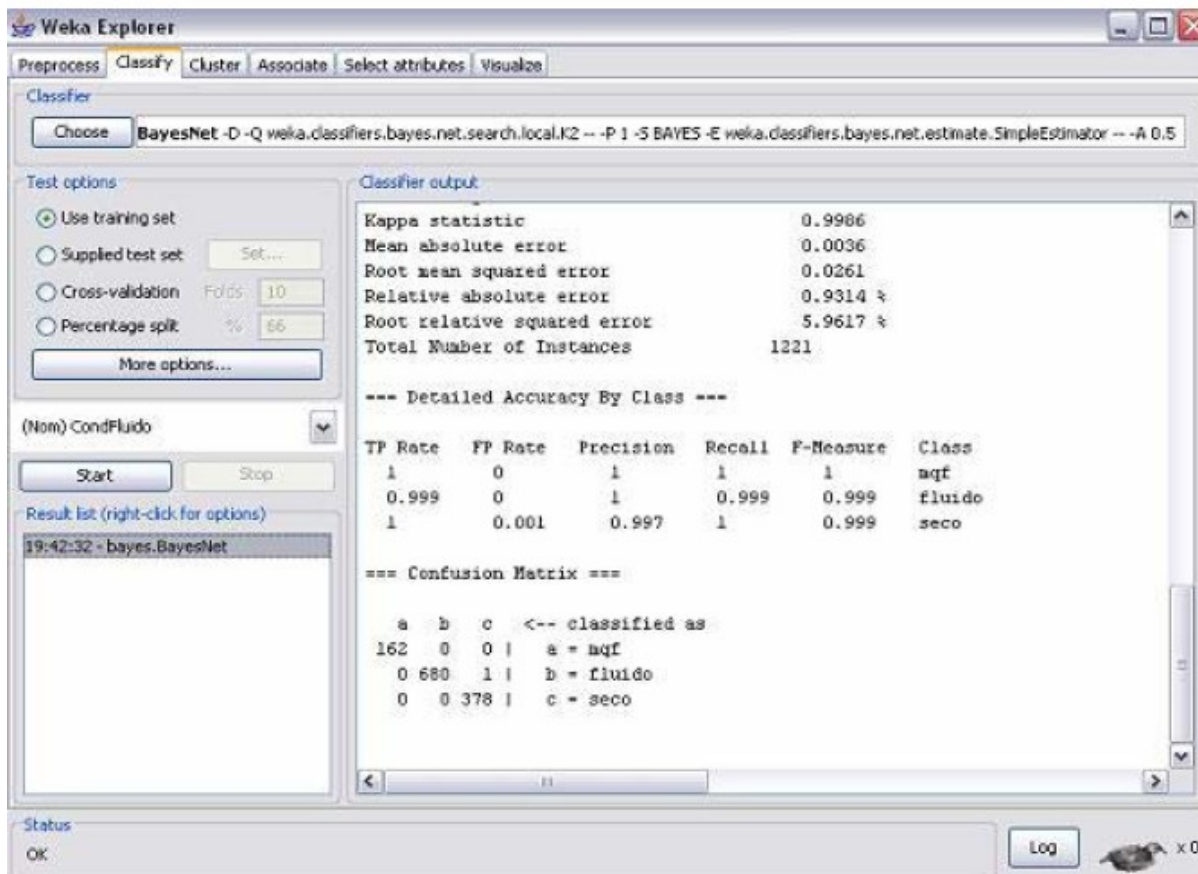
Class: Broca (Nom) Visualize All

Status

OK Log x 0

Ferramenta Weka

Classificação dos Dados



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **BayesNet** -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) CondFluido

Start Stop

Result list (right-click for options)

19:42:32 - bayes.BayesNet

Classifier output

Kappa statistic 0.9986
Mean absolute error 0.0036
Root mean squared error 0.0261
Relative absolute error 0.9314 %
Root relative squared error 5.9617 %
Total Number of Instances 1221

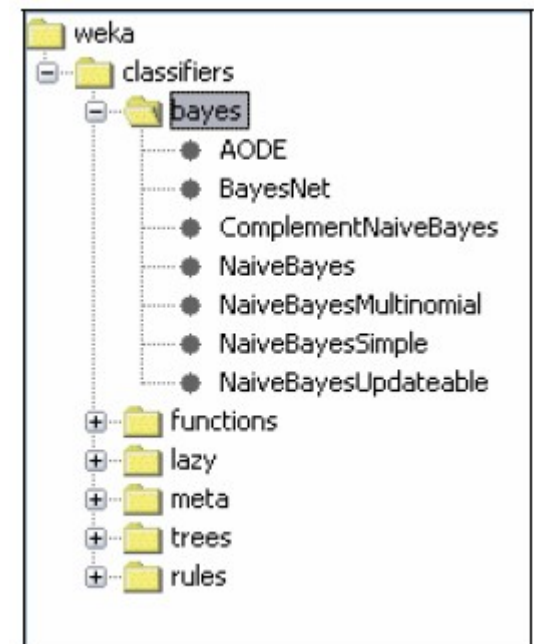
--- Detailed Accuracy By Class ---

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	mqf
0.999	0	1	0.999	0.999	fluido
1	0.001	0.997	1	0.999	seco

=== Confusion Matrix ===

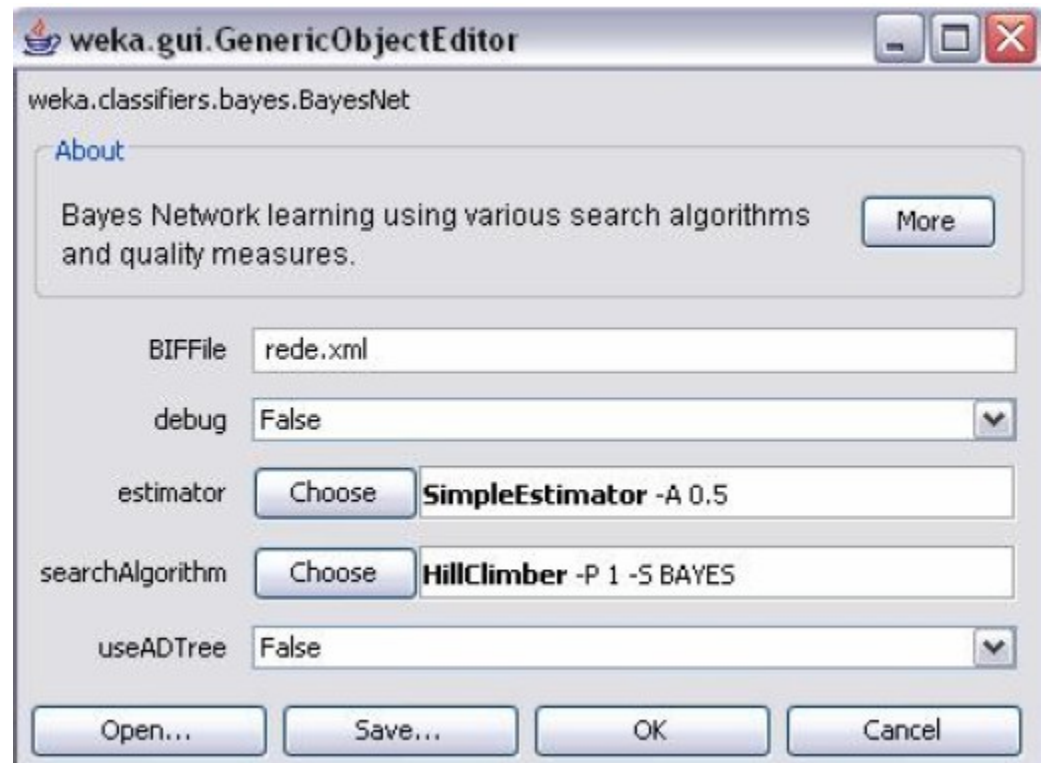
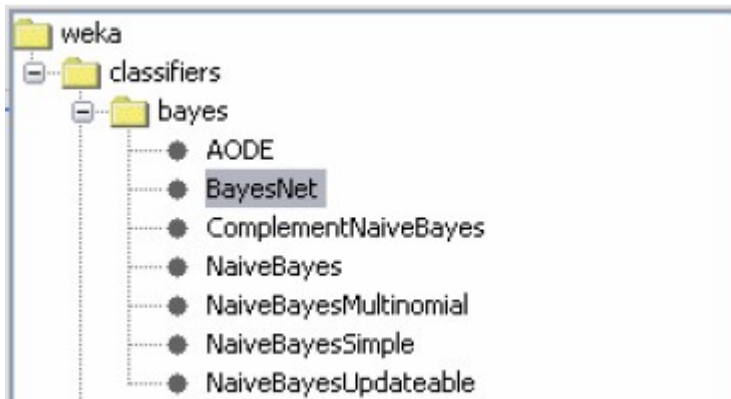
a	b	c	<-- classified as
162	0	0	a = mqf
0	680	1	b = fluido
0	0	378	c = seco

Status OK Log x 0



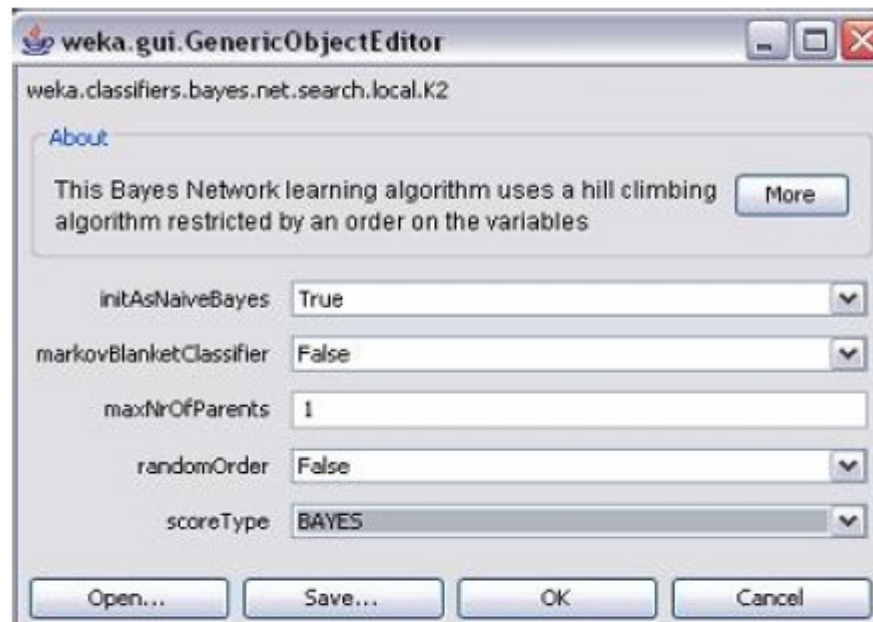
Redes Bayesianas no Weka

■ Algoritmo BayesNet



Redes Bayesianas no Weka

- Algoritmos de Busca e Pontuação
 - Hill climbing, K2, Repeated Hill Climbing, TAN, Simulated annealing, Tabu Search, Genetic Search.



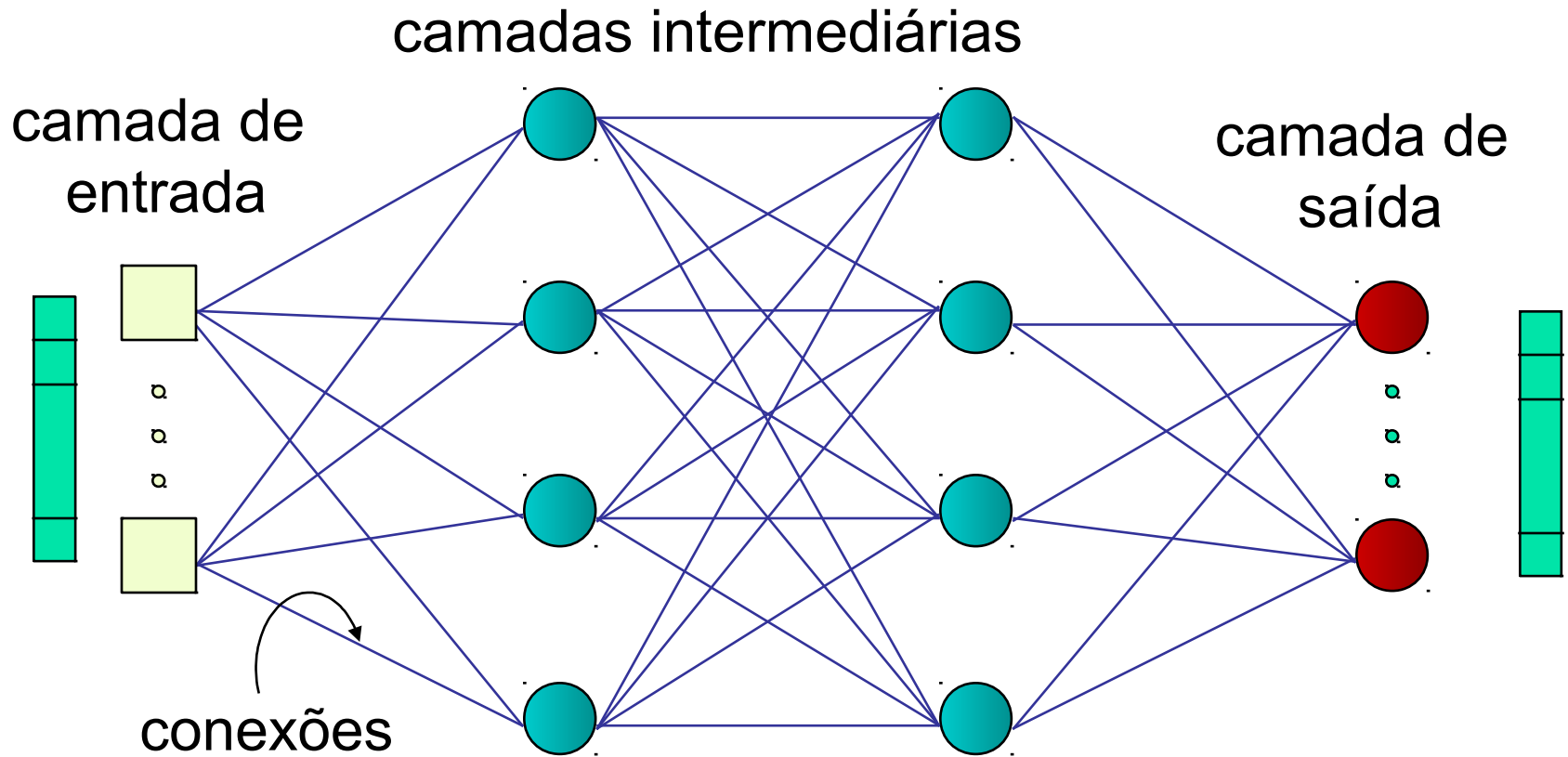
Paradigma Conexionista

- Redes Neurais
 - Estruturas distribuídas
 - Formadas por um grande número de unidades de processamento conectadas entre si
 - São pesquisadas em várias disciplinas:
 - Biologia, **Ciência da Computação**, Engenharias, Estatística, Filosofia, Física, Linguística, Matemática,

Redes Neurais Artificiais

- Sistemas computacionais distribuídos baseados na estrutura e funcionamento do sistema nervoso
 - Nós simulam neurônios
 - Conexões ponderadas simulam sinapses
- Definidas por
 - Arquitetura
 - Aprendizado

Redes Neurais Artificiais



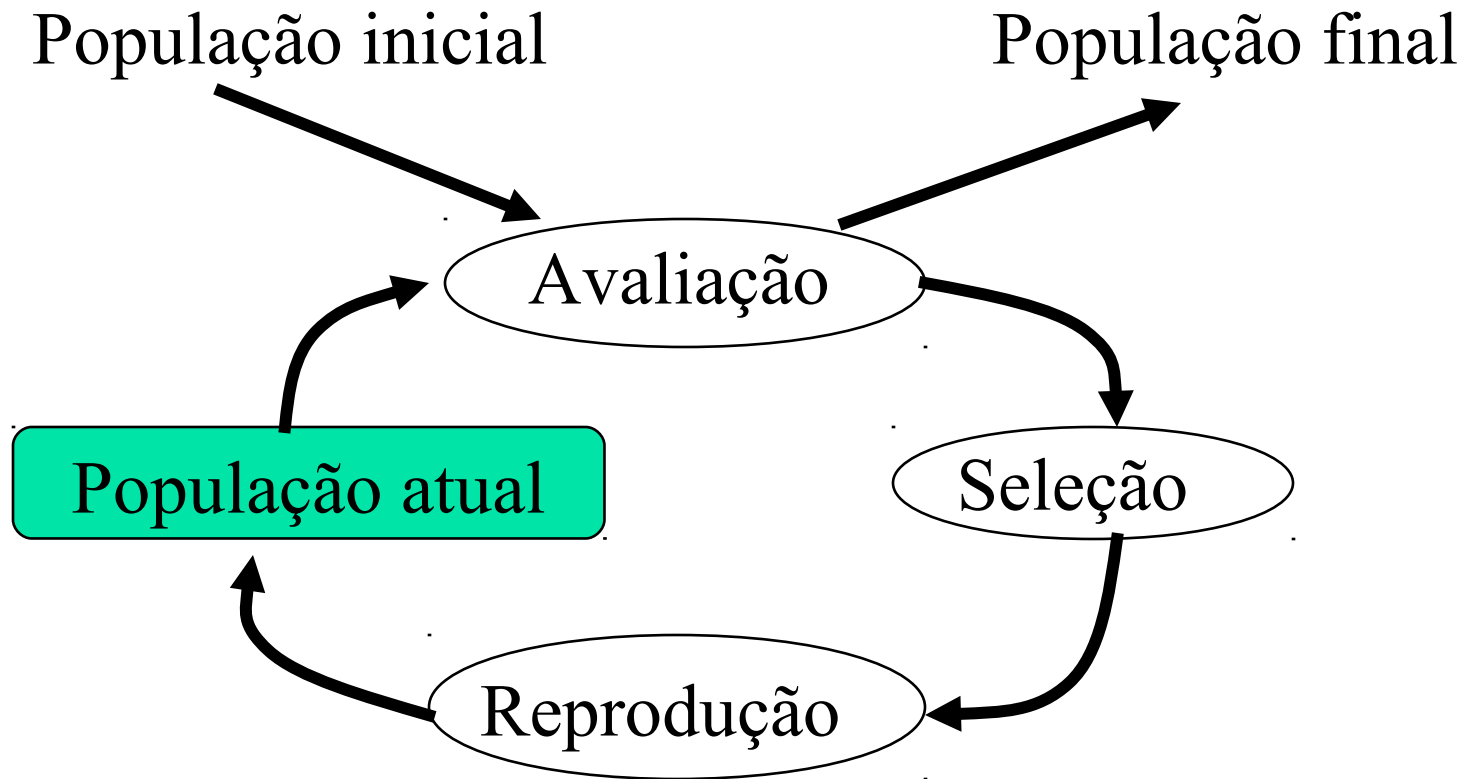
Paradigma Evolutivo

- Computação Evolutiva
 - Sistemas para a resolução de problemas que utilizam modelos computacionais baseados na teoria da evolução natural
 - Também chamados de algoritmos evolutivos
 - Inclui os Algoritmos Genéticos

Algoritmos Genéticos

- Técnica de busca e otimização
 - Baseados na genética e teoria da seleção natural
 - Utiliza uma população de soluções candidatas (indivíduos)
 - A cada indivíduo é associado um escore de aptidão, que mede o quão boa é a solução que ele representa
 - Otimização ocorre em várias gerações
 - A cada geração
 - Mecanismos de seleção selecionam os indivíduos mais aptos
 - Operadores de reprodução geram novos indivíduos

Algoritmos Genéticos





Aplicações

Aplicações

Através de registros médicos, definir que tratamentos são mais eficientes para determinadas doenças

- E para certos perfis de pacientes

Através de padrões de uso dos ocupantes de uma casa, definir como reduzir o consumo de energia

- E melhorar o conforto

Através da ordem de leitura de um jornal, destacar um conjunto de notícias na melhor ordem para o perfil do leitor.

Biologia Molecular e AM

- Problemas da Biologia Molecular que podem ser tratados por AM
 - Reconhecimento de genes
 - Construção de árvores filogenéticas
 - Análise de expressão gênica
 - Previsão de estruturas de proteínas
 - Análise de interação entre genes
 - Montagem de fragmentos
 - Alinhamento de seqüências

Reconhecimento de genes



- Um dos principais problemas em biologia molecular é a identificação de genes em seqüências de DNA não caracterizadas
- Algoritmos convencionais não têm sido eficientes
 - Variação natural dos genes
 - Complexidade dos genes
 - Natureza pouco compreendida dos genes

Aplicações

- Voz

<http://lmbarrros.tripod.com/computacao/vox.html>

- Reconhecimento de escrita

- <http://members.aol.com/Trane64/java/JRec.html>

Aplicações

- Compressão de imagens
- Reconhecimento de voz
- O problema do caixeiro viajante
- Biomedicina

- <http://www-cse.stanford.edu/classes/sophomore-college/projects-00/neural-networks/Applications/>

Bibliografia



A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, November 1977.

Mitchel, J. *Machine Learning*. NY: McGraw-Hill, 1990.