CIMPA-UCR

Continuous Optimization in Clustering

Javier Trejos – Eduardo Piza – Adilson Xavier – Alex Murillo
Universidad de Costa Rica
Universidade Federal de Río de Janeiro

CIMPA-UCR

Contents

- Hyperbolic smoothing approach
  - Problem specification
  - Hyperbolic Smoothing Clustering Method
  - Computational results
  - Partition into Boundary and Gravitational Regions
- Fuzzy clustering
  - Fuzzy clustering
  - HSCM approach to fuzzy clustering

CIMPA-UCR

Problem Specification

- Definition

Let be $\Omega = \{x_1, \ldots, x_n\}$ a set of n points (observations) in $\Re^p$ and K a given number of clusters.

In this presentation, the clustering problem considered is the partition of the set Ω in K disjoint subsets

$$\Omega = \bigcup_{k=1}^{K} C_k \ ,$$

$$C_k \cap C_{k'} = \varnothing, \ \forall k, k' = 1, \ldots, K, \ k \neq k',$$

$$C_k \neq \varnothing, \ \forall k = 1, \ldots, K.$$

Problem Specification

- The Minimum Sum-of-Squares Clustering Problem

$$\textit{minimize} \quad \sum_{k=1}^{K} \sum_{x \in C_k} \left\| \mathbf{x} - \mathbf{g}_k \right\|^2$$

$$\textit{subject to}: \; P \in \Pi_K \quad \mathbf{x} = (x_1, \ldots, x_p) \in \Re^p \;,$$

where $P = \{C_1, \ldots, C_K\}$ is the set of clusters, Π_K is the set of all possible K-partitions of the set Ω, $\mathbf{g}_k$ is the centroid of cluster C_k, $k = 1, \ldots, K$,

CIMPA-UCR

Hyperbolic Smoothing Clustering Method
General Problem Formulation

$\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$: set of n observations

$\mathbf{g}_k$, $k = 1, \ldots, K$: cluster centroids

$z_i = \min_{k=1,\ldots,K} \|\mathbf{x}_i - \mathbf{g}_k\|$

z_i is the distance, according to a given metric, between observation i and the nearest centroid $\mathbf{g}_k$

We seek for:

$$\text{minimize} \sum_{i=1}^{n} f(z_i)$$

where $f(z_i)$ is an arbitrary monotonic incresing function of the distance z_i

General Problem Formulation

Minimize $\sum_{i=1}^{n} f(z_i)$

$$z_i = \min_{k=1,\ldots,K} \|\mathbf{x}_i - \mathbf{g}_k\|$$

Trivial examples of monotonic increasing functions:

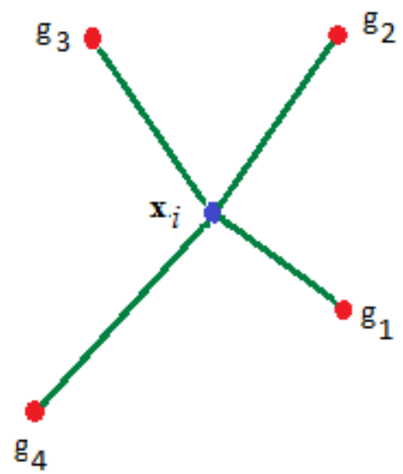
$$\sum_{i=1}^{n} f(z_i) = \sum_{i=1}^{n} z_i^2 \qquad \sum_{i=1}^{n} f(z_i) = \sum_{i=1}^{n} z_i$$

Possible distance metrics:

- L_2 (Euclidean)
- L_1 (Manhattan)
- L_p (Minkowski)
- L_∞ (Chebychev)

HSCM: Resolution Methodology

- The General Clustering Problem

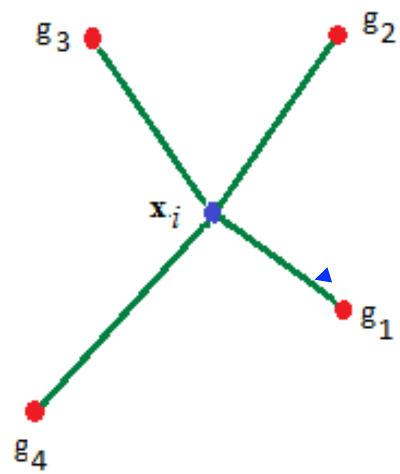


Consider a general observation point x_i

Let $\mathbf{g}_k, k = 1,\ldots,K,$ be the centroids of clusters, where each set of these centroid coordinates will be represented by $X \in \Re^{pK}$.

HSCM: Resolution Methodology

- The General Clustering Problem



$$z_i = \min_{\mathbf{g}_k \in X} \|\mathbf{x}_i - \mathbf{g}_k\|$$

Consider a general observation point x_i

Let $\mathbf{g}_k, k = 1, \ldots, K,$ be the centroids of clusters, where each set of these centroid coordinates will be represented by $X \in \Re^{pK}$.

HSCM: Resolution Methodology

- Problem transformation

The general clustering problem is equivalent to the following problem:

$$\text{Minimize} \sum_{i=1}^{n} f(z_i) \qquad (P1)$$

$$\text{subject to: } z_i = \min_{k=1,\dots,K} \|\mathbf{x}_i - \mathbf{g}_k\|, i = 1, 2, \dots, n.$$

HSCM: Resolution Methodology

- Problem transformation

From problem $(P1)$ it is obtained the **relaxed** problem:

$$\text{Minimize} \sum_{i=1}^{n} f(z_i) \qquad (P2)$$

subject to: $z_i - \|\mathbf{x}_i - \mathbf{g}_k\| \leq 0, i = 1, 2, \ldots, n;\ k = 1, \ldots, K.$

CIMPA-UCR

HSCM: Resolution Methodology

- Problem transformation

Let us use the auxiliary function

$$\varphi(y) = \max\{0, y\}.$$

From the inequalities

$$z_i - \|\mathbf{x}_i - \mathbf{g}_k\| \leq 0, \quad i = 1, 2, \ldots, n; \;\; k = 1, \ldots, K$$

it follows that

$$\sum_{k=1}^{K} \varphi\left(z_i - \|\mathbf{x}_i - \mathbf{g}_k\|\right) = 0, \quad i = 1, 2, \ldots, n.$$

HSCM: Resolution Methodology

- Problem transformation

Using this set of equality constraints in place of the constraints in $(P2)$ we obtain the following problem:

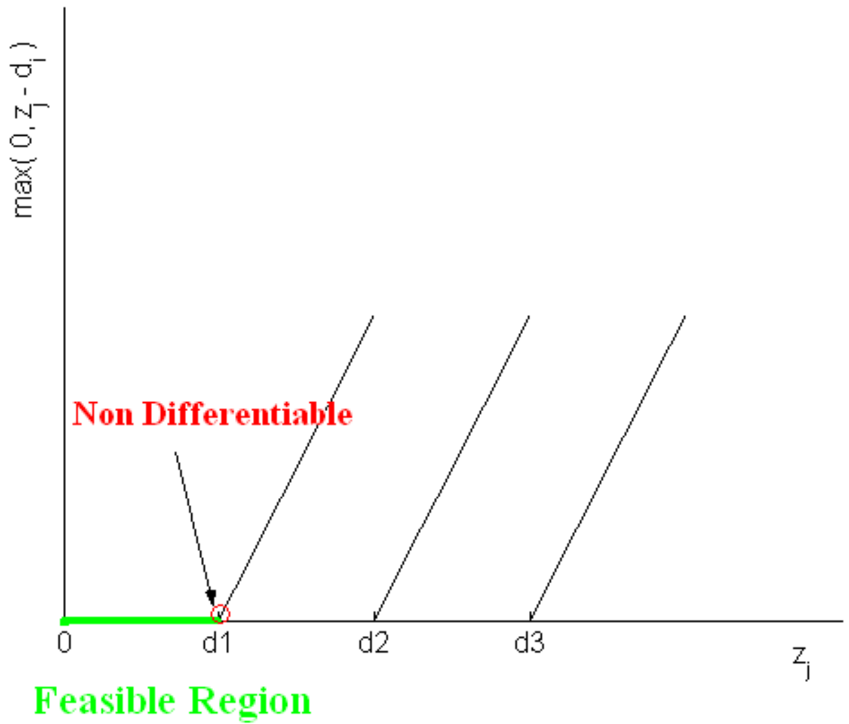
$$\text{Minimize} \sum_{i=1}^{n} f(z_i) \qquad (P3)$$

$$\text{subject to:} \quad \sum_{k=1}^{K} \varphi\left(z_i - \|\mathbf{x}_i - \mathbf{g}_k\|\right) = 0, \quad i = 1, 2, \ldots, n..$$

CIMPA-UCR

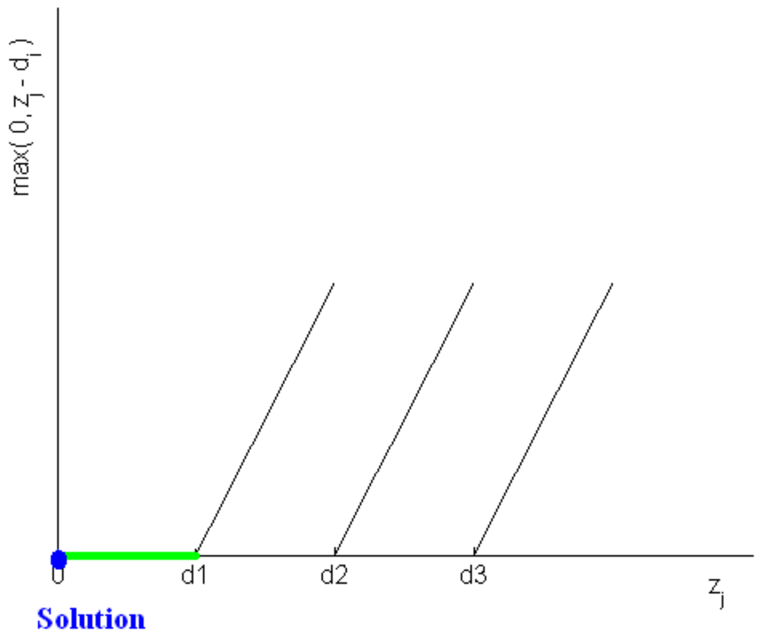
HSCM: Resolution Methodology

- Problem transformation



Objective Function Action

- Decrease $f(z_j)$



CIMPA-UCR

HSCM: Resolution Methodology

- Problem transformation

By performing a ε perturbation, it is obtained the problem:

$$\text{(P5) } \mathit{minimize} \quad \sum_{j=1}^{m} f(z_j)$$

$$\mathit{subject\ to}: \quad \sum_{i=1}^{q} \varphi(z_j - \| s_j - x_i \|) \geq \varepsilon, \quad j = 1, \ldots, m$$

for $\varepsilon > 0$.

HSCM: Resolution Methodology

- Problem transformation

Canonic

CIMPA-UCR

HSCM: Resolution Methodology

- Smoothing the problem

Let us define the auxiliary function

$$\phi(y,\tau) = (y + \sqrt{y^2 + \tau^2})/2$$

for $y \in \Re$ and $\tau > 0$.

HSCM: Resolution Methodology

- Smoothing the problem

Function ϕ has the following properties:

(a) $\phi(y,\tau) > \varphi(y), \forall \tau > 0;$

(b) $\lim_{\tau \to 0} \phi(y,\tau) = \varphi(y);$

(c) $\phi(.,\tau)$ is an increasing convex C^∞ function.

HSCM: Resolution Methodology

- Smoothing the problem

By using the Hyperbolic Smoothing approach for the problem (P5), it is obtained

$$\text{(P6) } \textit{minimize} \quad \sum_{j=1}^{m} f(z_j)$$

$$\textit{subject to:} \quad \sum_{i=1}^{q} \phi(z_j - \| s_j - x_i \|, \tau) \geq \varepsilon, \quad j = 1, \ldots, m$$

HSCM: Resolution Methodology

- Smoothing the problem

Differentiable

HSCM: Resolution Methodology

- Smoothing the distance calculation

For the Euclidian metric, let us define the auxiliary function

$$\theta(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} (s_{jl} - x_{il})^2 + \gamma^2}$$

for $\gamma > 0$.

HSCM: Resolution Methodology

- Smoothing the Euclidian distance:

Function θ has the following properties:

(a) $\lim_{\gamma \to 0} \theta(x_i, g_k, \gamma) = \|x_i - g_k\|_2$;

(b) θ is a C^∞ function.

HSCM: Resolution Methodology

- Smoothing the problem

Therefore, it is now obtained the completely smooth problem

$$(P7)\ \textit{minimize} \quad \sum_{j=1}^{m} f(z_j)$$

$$\textit{subject to:} \quad \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \geq \varepsilon, \quad j = 1, \ldots, m$$

HSCM: Resolution Methodology

- Problem resolution

By considering the KKT conditions for the problem (P7), it is possible to conclude that all inequalities will certainly be active.

$$\text{(P8)}\quad \textit{minimize} \quad \sum_{j=1}^{m} f(z_j)$$

$$\textit{subject to:}\quad h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0\,, \quad j = 1, \ldots, m.$$

HSCM: Resolution Methodology

- Problem resolution

Some remarks about the problem (P8):

The variable domain space has $(nq + m)$ dimensions.

All functions of this problem are C^∞ functions in relation to the variables (x,z).

Each variable z_j appears only in one equality constraint $h(x,z_j)$ for all $j=1,\dots,m$.

HSCM: Resolution Methodology

- Problem resolution

Some remarks about the problem (P8):

The parcial derivative of $h(x,z_j)$ in relation to z_j is not equal to zero for all $j = 1,\ldots,m$.

In the face of the previous remarks, it is possible to use the Implicit Function Theorem in order to calculate each component z_j, $j = 1,\ldots,m$, as a function of the centroids variables x_i, $i = 1,\ldots,q$.

CIMPA-UCR

HSCM: Resolution Methodology

- Problem resolution

Therefore, by using the Implicit Function Theorem, it is obtained the unconstrained problem

$$\text{(P9)}\ \mathit{minimize}\ \ F(x) = \sum_{j=1}^{m} f(z_j)$$

where each z_j is obtained by the calculation of a zero of each equation

$$h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0\,,\ \ j = 1, \dots, m.$$

HSCM: Resolution Methodology

- Problem resolution

Again, due the Implicit Function Theorem, the functions $z_j(x)$ have all derivatives in relation to the variables $x_i, i = 1,\ldots,q$. So, it is possible to calculate the gradient of the objective function of the problem (P9)

$$\nabla F(x) = \sum_{j=1}^{m} \frac{df(z_j(x))}{dz_j} \nabla z_j(x)$$

where

$$\nabla z_j(x) = -\nabla h(x, z_j) / \frac{\partial h(x, z_j)}{\partial z_j}.$$

HSCM: Resolution Methodology

- Problem resolution

Some remarks about the problem (P9):

It can be solved by making use of any method based in the first and second order derivatives information (conjugate gradient, quasi-Newton, Newton methods, etc).

It is defined in a *nq*-dimensional space, therefore a smaller dimension problem in relation to the original problem (P8).

CIMPA-UCR

Simplified HSC Algorithm

Initialization Step: Choose values $0 < \rho_1 < 1,\ 0 < \rho_2 < 1,\ 0 < \rho_3 < 1;$

Choose initial values: $x^0,\ \gamma^1,\ \tau^1,\ \varepsilon^1$; Let $k = 1$.

Main Step: Repeat indefinitely

Solve the problem (P9)

$$minimize\ \ F(x) = \sum_{j=1}^{m} z_j(x)^2$$

with $\gamma = \gamma^k,\ \tau = \tau^k,\ \varepsilon = \varepsilon^k$, starting at the initial point x^{k-1},

and let x^k be the solution obtained.

Let $\gamma^{k+1} = \rho_1 \gamma^k,\ \tau^{k+1} = \rho_2 \tau^k,\ \varepsilon^{k+1} = \rho_3 \varepsilon^k,\ k = k + 1$.

Remarks about the HSCM Algorithm

The algorithm causes τ and γ approach zero, so the constraints of the subproblems it solves, given as in (P8), tend to those of (P5).

The algorithm causes ε approaches to zero, so, in a simultaneous movement, the solved problem (P5) gradually approaches to problem (P4).

HSCM - Computational Results: TSPLIB-3038

Minimum Sum-of-Squares Clustering Formulation

q	Putative Global Minimum	Best Solution HSC Algorithm	E_{Best}	Best Solution Frequence	Start Points HSC Algorithm	E_{Mean}	CPU Time
2	0.31688E10	0.31705E10	0.05	10	10	0.05	0.60
3	0.21763E10	0.21776E10	0.06	7	10	1.08	1.18
4	0.14790E10	0.14793E10	0.02	10	10	0.02	1.81
5	0.11982E10	0.11986E10	0.03	9	10	0.06	3.09
6	0.96918E9	0.96936E9	0.02	10	10	0.02	9.56
7	0.83966E9	0.83967E9	0.001	7	10	4.66	10.65
8	0.73475E9	0.73491E9	0.02	1	10	1.35	16.24
9	0.64477E9	0.64471E9	**-0.01**	2	10	0.38	10.05
10	0.56025E9	0.56030E9	+0.01	10	10	0.01	16.16
20	0.26681E9	0.26675E9	**-0.02**	1	10	2.51	62.90
30	0.17557E9	0.17543E9	**-0.08**	3	10	0.36	169.17
40	0.12548E9	0.12541E9	**-0.06**	1	20	1.34	333.92
50	0.98400E8	0.98893E8	+0.50	1	40	1.46	662.33
60	0.82006E8	0.81115E8	**-1.09**	1	10	0.19	1049.97
80	0.61217E8	0.61025E8	**-0.31**	1	10	0.63	2038.31
100	0.48912E8	0.48470E8	**-0.90**	1	10	0.19	3499.76

HSCM: Results for the Pla85900

Minimum Sum-of-Squares Clustering Formulation

CIMPA-UCR

The Computational Task of the HSC Algorithm

- The most relevant computational task associated to HSC Algorithm is the determination of the zeros of m equations:

$$h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x, \gamma), \tau) - \varepsilon = 0, \qquad j = 1, \ldots, m$$

The Partition into Boundary and Gravitational Regions

- The basic idea of the approach is the partition of the set of observations in two non overlapping parts:

- - the first set corresponds to the observation points that are
- relatively close to two or more centroids;
-
- - the second set corresponds to the observation points that are
- significantly close to a unique centroid in comparasion with the other ones.

Partition of the Set of Observations

$\bar{x}$: the referencial point

δ: the band zone width

The Boundary Band Zone

Gravitational Region => $\overline{x}_2$

Gravitational Regions

The Partition of the Set of Observations

- Let us define:
- J_B is the set of boundary observations;
- J_G is the set of gravitational observations.

J_B

- Considering this partition, the objective function can be expressed in the following way:

$$minimize\ F(x) = \sum_{j=1}^{m} f(z_j) = \sum_{j \in J_B} f(z_j) + \sum_{j \in J_G} f(z_j)$$

So, the objective function can be expressed in the following way:

$$minimize\ F(x) = f_B(x) + f_G(x)$$

CIMPA-UCR

The First Component of the Objective Function

The component associated with the set of boundary observations, can be calculated by using the previous presented smoothing approach:

$$minimize \;\; F_B(x) = \sum_{j \in J_B} f(z_j)$$

where each z_j results from the calculation of a zero of each equation:

$$h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \qquad j \in J_B.$$

The Second Component fro the MSSC

For The Minimum Sum-of-Squares Clustering formulation,
the second component, associated with the gravitational regions,
can be calculated in a direct form:

$$minimize \;\; F_G(x) = \sum_{i=1}^{q} \sum_{j \in J_i} \left\| s_j - v_i \right\|_2^2 + \sum_{i=1}^{q} |J_i| \left\| x_i - v_i \right\|_2^2$$

where J_i is the set of observations associated to centroid i

v_i is the gravity center of cluster i

The Gradient of the Second Component for the MSSC

The gradient of the second component, associated with the gravitational regions, can be calculated in an easy way:

$$\nabla F_G(x) = \sum_{\iota=1}^{q} 2|J_i| \left(x_i - v_i\right)$$

where J_i is number the of observations associated to centroid i

Simplified Accelerated Algorithm: AHSCM

Initialization Step: Choose values $0 < \rho_1 < 1,\ 0 < \rho_2 < 1,\ 0 < \rho_3 < 1;$

Choose initial values: $x^0,\ \gamma^1,\ \tau^1,\ \varepsilon^1$; Let $k = 1$.

Specify the initial boundary band width: δ^1.

Main Step: Repeat indefinitely

By using $\bar{x} = x^{k-1}$ and $\delta = \delta^k$ determine partitions J_B and J_G.

Solve the problem

$$minimize\ \ f(x) = \sum_{j \in J_B} z_j(x)^2 + \sum_{j \in J_G} z_j(x)^2$$

with $\gamma = \gamma^k,\ \tau = \tau^k,\ \varepsilon = \varepsilon^k$, starting at the initial point x^{k-1},

and let x^k be the solution obtained.

Redefine the boundary band width value δ^{k+1}.

Let $\gamma^{k+1} = \rho_1 \gamma^k,\ \tau^{k+1} = \rho_2 \tau^k,\ \varepsilon^{k+1} = \rho_3 \varepsilon^k,\ k = k+1.$

CIMPA-UCR

Comparison of Results: TSPLIB-3038 Minimum Sum-of-Squares Clustering Formulation

q	f_{opt}	Algorithm HSC			Algorithm XHSC		
		f_{Best}	E	$Time$	f_{Best}	E	$Time$
2	0.31688E10	0.31705E10	0.05	0.60	0.31705E10	0.05	0.07
3	0.21763E10	0.21776E10	0.06	1.08	0.21776E10	0.06	0.12
4	0.14790E10	0.14793E10	0.02	1.81	0.14793E10	0.02	0.13
5	0.11982E10	0.11986E10	0.03	3.09	0.11984E10	0.02	0.17
6	0.96918E09	0.96936E09	0.02	9.56	0.96936E09	0.02	0.23
7	0.83966E09	0.83967E09	0.001	10.65	0.83967E09	0.001	0.19
8	0.73475E09	0.73491E09	0.02	16.24	0.73491E09	0.02	0.27
9	0.64477E09	0.64471E09	**-0.01**	10.05	0.64551E09	0.11	0.27
10	0.56025E09	0.56030E09	0.01	16.16	0.56030E09	0.01	0.28
20	0.26681E09	0.26675E09	**-0.02**	62.90	0.26694E09	0.05	0.59
30	0.17557E09	0.17543E09	**-0.08**	169.17	0.17612E09	0.31	0.86
40	0.12548E09	0.12541E09	**-0.06**	333.92	0.12533E09	**-0.11**	1.09
50	0.98400E08	0.98893E08	0.50	662.33	0.98834E08	0.44	1.36
60	0.82006E08	0.81115E08	**-1.09**	1049.97	0.81349E08	**-0.80**	1.91
80	0.61217E08	0.61025E08	**-0.31**	2038.31	0.60773E08	**-0.73**	6.72
100	0.48912E08	0.48470E08	**-0.90**	3499.76	0.48615E08	**-0.60**	9.79

Comparison of Results: Pla85900

Minimum Sum-of-Squares Clustering Formulation

q	$f_{Calculated}$	Algorithm HSC			Algorithm XHSC		
		Occur.	E_{Mean}	$Time_{Mean}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.37491E16	4	0.86	23.07	5	0.58	3.65
3	0.22806E16	10	0.00	47.41	7	0.04	4.92
4	0.15931E16	10	0.00	76.34	10	0.00	5.76
5	0.13397E16	1	0.80	124.32	1	1.35	7.78
6	0.11366E16	8	0.12	173.44	2	1.25	7.87
7	0.97110E15	4	0.42	254.37	1	0.87	9.33
8	0.83774E15	8	0.55	353.61	4	0.37	12.96
9	0.74660E15	3	0.68	438.71	1	0.25	13.00
10	0.68294E15	4	0.29	551.98	3	0.46	14.75

CIMPA-UCR

XHSCM: Results for the FL3795 Instance

Minimum Sum-of-Squares Clustering Formulation

XHSCM: Results for the FNL4461 Instance

CIMPA-UCR

XHSCM: Results for the RL5915 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.100036E+12	10	0.00	0.14
3	0.642032E+11	10	0.00	0.23
4	0.485154E+11	4	1.32	0.31
5	0.379585E+11	8	1.01	0.45
6	0.318584E+11	3	0.93	0.51
7	0.267754E+11	7	0.67	0.60
8	0.234633E+11	8	0.48	0.54
9	0.208867E+11	4	0.70	0.70
10	0.187794E+11	1	0.41	0.74

XHSCM: Results for the RL5934 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.920845E+11	1	0.00	0.14
3	0.681904E+11	5	1.77	0.23
4	0.488114E+11	10	0.00	0.31
5	0.393650E+11	1	1.69	0.39
6	0.317269E+11	9	0.00	0.52
7	0.279379E+11	5	0.58	0.48
8	0.244060E+11	4	1.23	0.57
9	0.215563E+11	2	1.56	0.69
10	0.191761E+11	2	2.35	0.76

CIMPA-UCR

XHSCM: Results for the Pla7397 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.178155E+15	10	0.00	0.09
3	0.111206E+15	10	0.00	0.15
4	0.629983E+14	10	0.00	0.23
5	0.506247E+14	3	1.94	0.34
6	0.396774E+14	4	12.15	0.41
7	0.352141E+14	3	0.58	0.48
8	0.308094E+14	2	4.26	0.66
9	0.272683E+14	1	2.37	0.77

CIMPA-UCR

XHSCM: Results for the RL11849 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.210287E+12	9	4.01	0.26
3	0.152046E+12	3	0.89	0.49
4	0.104973E+12	9	0.00	0.68
5	0.809552E+11	1	1.11	0.83
6	0.637439E+11	10	0.00	0.99
7	0.552323E+11	4	0.61	1.41
8	0.472981E+11	2	0.53	1.29
9	0.416333E+11	2	0.84	1.36
10	0.369192E+11	8	0.53	1.55

XHSCM: Results for the USA13509 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.109756D+15	7	0.29	0.34
3	0.573853D+14	10	0.00	0.43
4	0.434554D+14	8	0.47	0.82
5	0.329511D+14	2	0.01	1.01
6	0.265986D+14	10	0.00	0.82
7	0.222716D+14	5	0.67	1.05
8	0.194845D+14	5	0.81	1.27
9	0.167980D+14	1	4.65	1.41
10	0.149816D+14	3	1.39	1.69

CIMPA-UCR

XHSCM: Results for the BRD14051 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.371152E+11	10	0.00	0.32
3	0.197682E+11	10	0.00	0.35
4	0.152334E+11	9	0.33	0.54
5	0.122288E+11	7	1.20	0.82
6	0.101914E+11	5	0.18	0.94
7	0.832838E+10	4	0.00	1.64
8	0.736798E+10	1	1.76	1.19
9	0.656428E+10	1	3.29	1.66
10	0.593928E+10	1	1.17	2.00

XHSCM: Results for the D15112 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.368403E+12	10	0.00	0.36
3	0.253240E+12	10	0.00	0.62
4	0.173603E+12	10	0.00	0.77
5	0.132707E+12	10	0.00	0.88
6	0.111553E+12	10	0.00	1.05
7	0.994046E+11	4	0.03	1.41
8	0.816951E+11	6	1.75	1.77
9	0.713070E+11	7	0.49	1.82
10	0.644901E+11	5	0.71	2.27

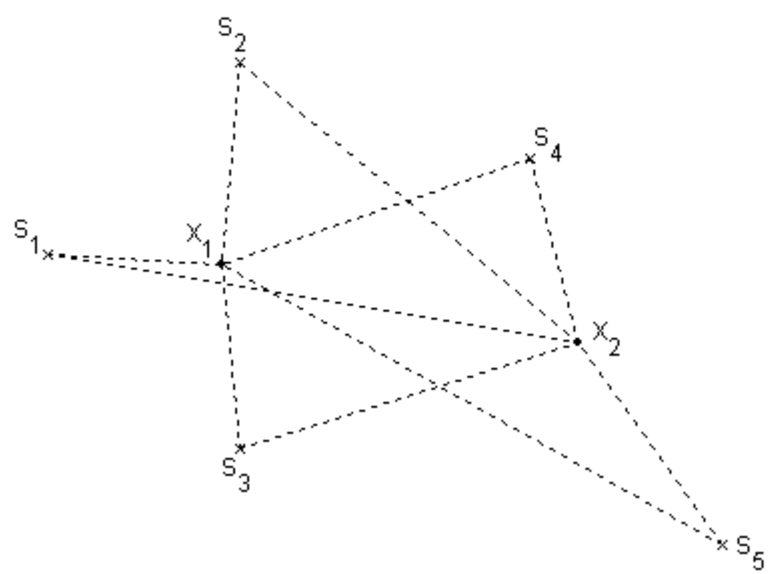
XHSCM: Results for the BRD18512 Instance

XHSCM: Results for the Pla33810 Instance

q	$f_{HSC_{Best}}$	Occur.	E_{Mean}	$Time_{Mean}$
2	0.946269E+15	6	5.23	1.11
3	0.605695E+15	7	0.47	2.68
4	0.404399E+15	10	0.00	2.35
5	0.335680E+15	5	0.22	3.54
6	0.280962E+15	3	1.24	3.69
7	0.238079E+15	3	1.48	4.26
8	0.204086E+15	6	1.34	4.22
9	0.179965E+15	3	0.01	4.29
10	0.164824E+15	2	0.68	5.14

Performance of the HSCM

- The performance of HSCM can be attributed to the complete differentiability of the approach. So, each centroid sees permanently every observation point, conversely, each observation point can permanently see every centroid and attract it .



Performance of the XHSCM Algorithm

- The robustness performance of the XHSCM Algorithm can be attributed to the complete differentiability of the approach.
- The speed performance of the XHSCM Algorithm can be attributed to the partition of the set of observations in two non overlapping parts.This last approach engenders a drastic simplification of computational tasks.

Conclusions

- In view of the preliminary results obtained, where the proposed methodology performed efficiently and robustly, these algorithms can represent a possible approach for dealing with the solving of clustering of problems of real applications.

CIMPA-UCR

Correspondencias en Minería de Datos

Javier Trejos – Mayra Ruano

Universidad de Costa Rica

Backcountry.com

Contenidos

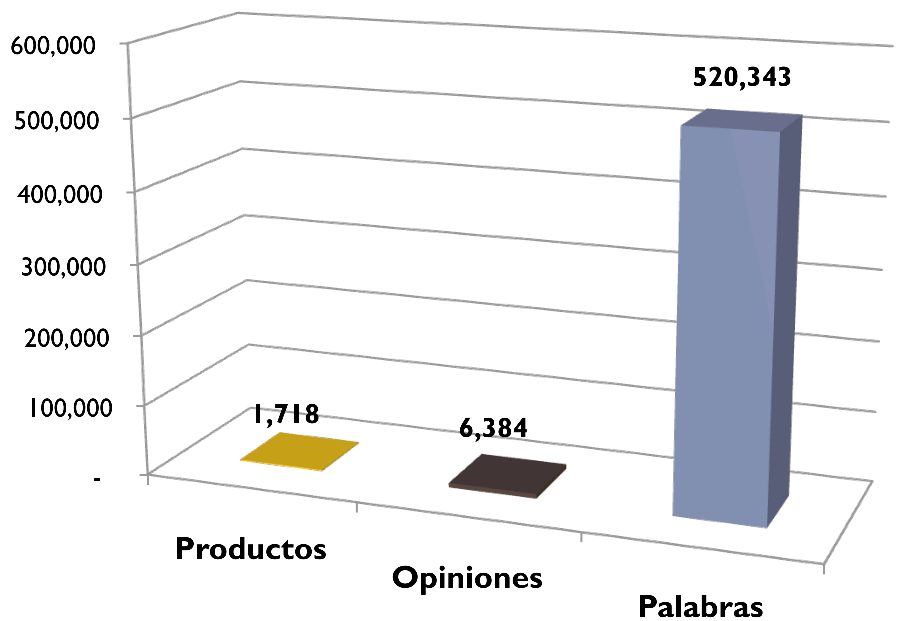
- Motivación
- ¿Cómo resolver el problema?
  - Herramientas existentes
  - Método Análisis Factorial de Correspondencias
- Ejemplo aplicado en Adjetivos
- Ejemplo aplicado en Patrones de Tamaños
- Conclusión

¿Por qué es importante para Backcountry.com?

- La comunidad de usuarios de Backcountry.com ha publicado más de **134,000** opiniones desde Setiembre de 2002
  → Mucho conocimiento esperando a ser descubierto
- Al entender las percepciones de los clientes es posible:
  - Mejorar la experiencia de los usuarios en el sitio
  - Incrementar las ventas

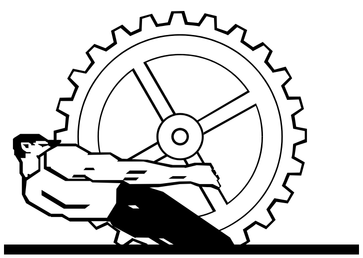
Dimensión del problema

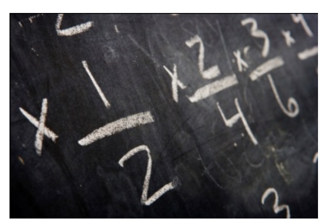
- Los datos de tipo texto son complejos para manipular
- Categoría de Zapatos, desde Enero de 2009 a Enero de 2010



Solución propuesta

- Herramientas existentes
  - Librerías de Procesamiento de Lenguaje Natural
  - **GATE**: *General Architecture for Text Engineering*
- Método de análisis de datos que trabajen bien con los datos
  - **Análisis Factorial de Correspondencias**(AFC)

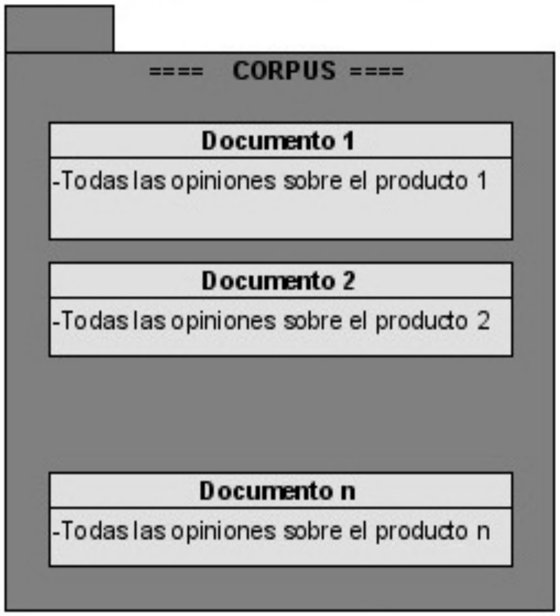


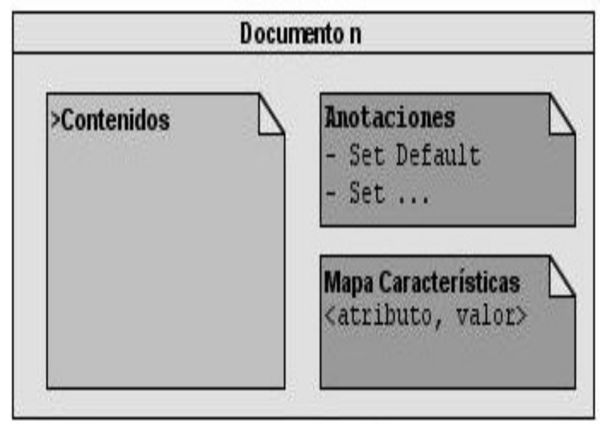


Librería GATE

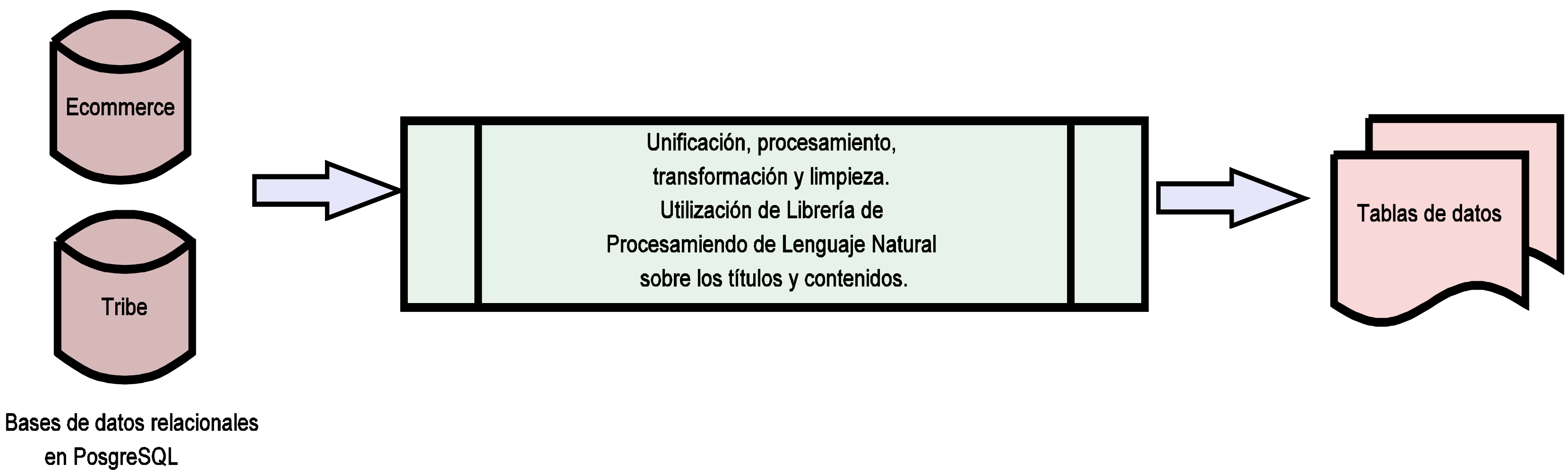
- Marco de trabajo para realizar minería de texto y procesamiento de lenguaje natural
- Manipulación de texto no estructurado, para convertirlo en datos estructurados
  - Identificación de *tokens*
  - Eliminación de *palabras de parada (Porter, 1979)*
  - Extracción de raíces de las palabras
  - Identificación de las *partes del discurso*
  - Anotación de patrones en el texto

Corpus y documentos GATE





Proceso de generación de las tablas de datos



CIMPA-UCR

Repaso teórico del método AFC

- Propuesto por Jean Paul Benzécri (1967)
- Objetivo: analizar dos variables cualitativas x y y
- Cada variable tiene modalidades: $x^1, x^2, \ldots, x^p$

  $y^1, y^2, \ldots, y^q$

  – Producto: Codigo1, Codigo2, etc..

  – Palabras: *heavy, nice, comfortable*, etc…

  – Patrones de tamaño: size up, size down, true size

Tabla de Contingencia

$$
\begin{array}{c|ccccc|c|}
 & y^1 & \cdots & y^k & \cdots & y^q & \\
\hline
x^1 & x_{11} & \cdots & x_{1k} & \cdots & x_{1q} & x_{1\bullet} \\
 & \vdots & & \vdots & & \vdots & \\
x^j & x_{j1} & \cdots & x_{jk} & \cdots & x_{jq} & x_{j\bullet} \\
 & \vdots & & \vdots & & \vdots & \\
x^p & x_{p1} & \cdots & x_{pk} & \cdots & x_{pq} & x_{p\bullet} \\
\hline
 & x_{\bullet 1} & & x_{\bullet k} & & x_{\bullet q} & x_{\bullet\bullet} \\
\hline
\end{array}
$$

$$x_{j\bullet} = \sum_{k=1}^{q} x_{jk} \qquad x_{\bullet k} = \sum_{j=1}^{p} x_{jk} \qquad x_{\bullet\bullet} = \sum_{j=1}^{p} \sum_{k=1}^{q} x_{jk}$$

Perfiles Fila

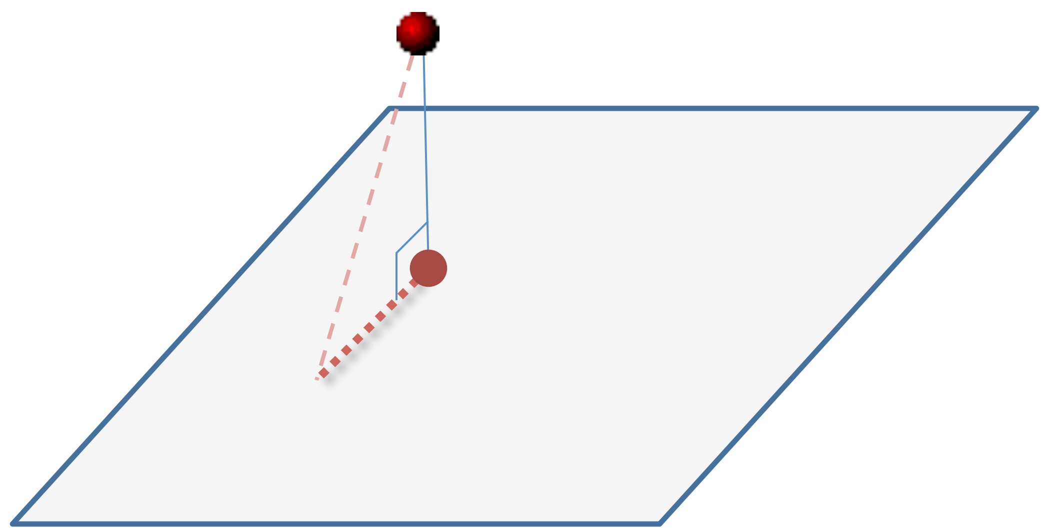
$$
\begin{array}{c|ccccc|}
 & y^1 & \cdots & y^k & \cdots & y^q \\
\hline
x^1 & \dfrac{x_{11}}{x_{1\bullet}} & \cdots & \dfrac{x_{1k}}{x_{1\bullet}} & \cdots & \dfrac{x_{1q}}{x_{1\bullet}} \\
 & \vdots & & \vdots & & \vdots \\
x^j & \dfrac{x_{j1}}{x_{j\bullet}} & \cdots & \dfrac{x_{jk}}{x_{j\bullet}} & \cdots & \dfrac{x_{jq}}{x_{j\bullet}} \\
 & \vdots & & \vdots & & \vdots \\
x^p & \dfrac{x_{p1}}{x_{p\bullet}} & \cdots & \dfrac{x_{pk}}{x_{p\bullet}} & \cdots & \dfrac{x_{pq}}{x_{p\bullet}} \\
\hline
\end{array}
\qquad
\begin{array}{ccc}
x_{1\bullet}/x_{\bullet\bullet} & & f_{1\bullet} \\
\vdots & & \\
x_{j\bullet}/x_{\bullet\bullet} & = & f_{j\bullet} \\
\vdots & & \\
x_{p\bullet}/x_{\bullet\bullet} & & f_{p\bullet}
\end{array}
$$

$$
\frac{x_{\bullet 1}}{x_{\bullet\bullet}} \cdots \frac{x_{\bullet k}}{x_{\bullet\bullet}} \cdots \frac{x_{\bullet q}}{x_{\bullet\bullet}} \quad = \quad f_{\bullet 1} \cdots f_{\bullet k} \cdots f_{\bullet q}
$$

CIMPA-UCR

AFC

- Encontrar un subespacio de menor dimensión para representar los perfiles



CIMPA-UCR

Algunas definiciones

- Distancia d:
  - $d(x,x) = 0 \ \forall x \in \mathbb{R}^q$.
  - $d(x,y) = d(y,x) \ \forall x,y \in \mathbb{R}^q$.
  - $d(x,z) < d(x,y) + d(y,z) \ \forall x,y,z \in \mathbb{R}^q$.

$$\mathbb{R}^q : \langle x,y \rangle_M = x^t M y$$

- Métrica M:

$$D = diag(p_i) \qquad \sum_{i=1}^{p} p_i = 1$$

- Métrica de pesos D:

$$f_{ij} = \frac{x_{ij}}{x_{..}}$$

- Frecuencias F $1 = \sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.j} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}$

Algunas definiciones (cont.)

- Centro de gravedad: $$g_x = \sum_{i=1}^{p} p_i(pf_i)$$

- Nube: $$N(X, M, D) \longrightarrow N_x(X_{pf}, D_y^{-1}, D_x)$$

$$D_x = diag(\frac{x_{i.}}{x_{..}}) = diag(f_{i.})$$

$$D_y = diag(\frac{x_{.j}}{x_{..}}) = diag(f_{.j})$$

$$\begin{aligned} d^2_{\chi^2}(pf_i, pf_i') &= \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \\ &= (pf_i - pf_{i'}) D_y^{-1} (pf_i - pf_{i'})^t \end{aligned}$$

AFC como caso particular del ACP

$$\max_u \left\{ \sum_{i=1}^{p} f_{i.} d^2_{\chi^2}(pf_i, g_x) \right\}$$ sujeto a $||u||_{D_y^{-1}} = 1$

$$\longrightarrow \quad \max_u \{ I_{\Delta u}(N_x) \}$$ sujeto a $u^t D_y^{-1} u = 1$

$$\longrightarrow \quad \max_u \{ u^t D_y^{-1} F^t D_x^{-1} F D_y^{-1} u \}$$

La solución es el vector propio U de la matriz $S = F^t D_x^{-1} F D_y^{-1}$
Asociado con el valor propio λ más grande diferente de 1

Cada vector propio $u_0, u_1, \ldots, u_r$ asociado a $1 > \lambda_0 > \lambda_1 > \ldots > \lambda_r > 0$
determina una un eje que pasa por el origen de la nube.

Estos vectores propios son utilizados para definir los planos factoriales.

CIMPA-UCR

Coordenadas en el plano factorial

- La proyección D_y^{-1} ortogonal sobre el vector u_h es

$$coord_{u_h}(pf_i) = pf_i D_y^{-1} u_h$$

- Así, la proyección del i-*ésimo* perfil fila sobre el plano factorial generado por u_h, u_l es:

$$(coord_{u_h}(pf_i), coord_{u_l}(pf_i))$$

Calidad de la representación Contribución absoluta

$$CntrAbs_h(i) = \frac{f_{i.}coord_{u_h}(pf_i)^2}{\lambda_h}$$

- Mide el aporte que hace un eje a la inercia total.
- Inercia proyectada sobre eje *h*: $I_h = \lambda_h$
- Muestra en qué proporción contribuye un perfil i a la inercia de la nube proyectada sobre el eje h.

Calidad de la representación
Contribución relativa

- Es un indicador de cuán bien representado está un punto sobre el subespacio factorial

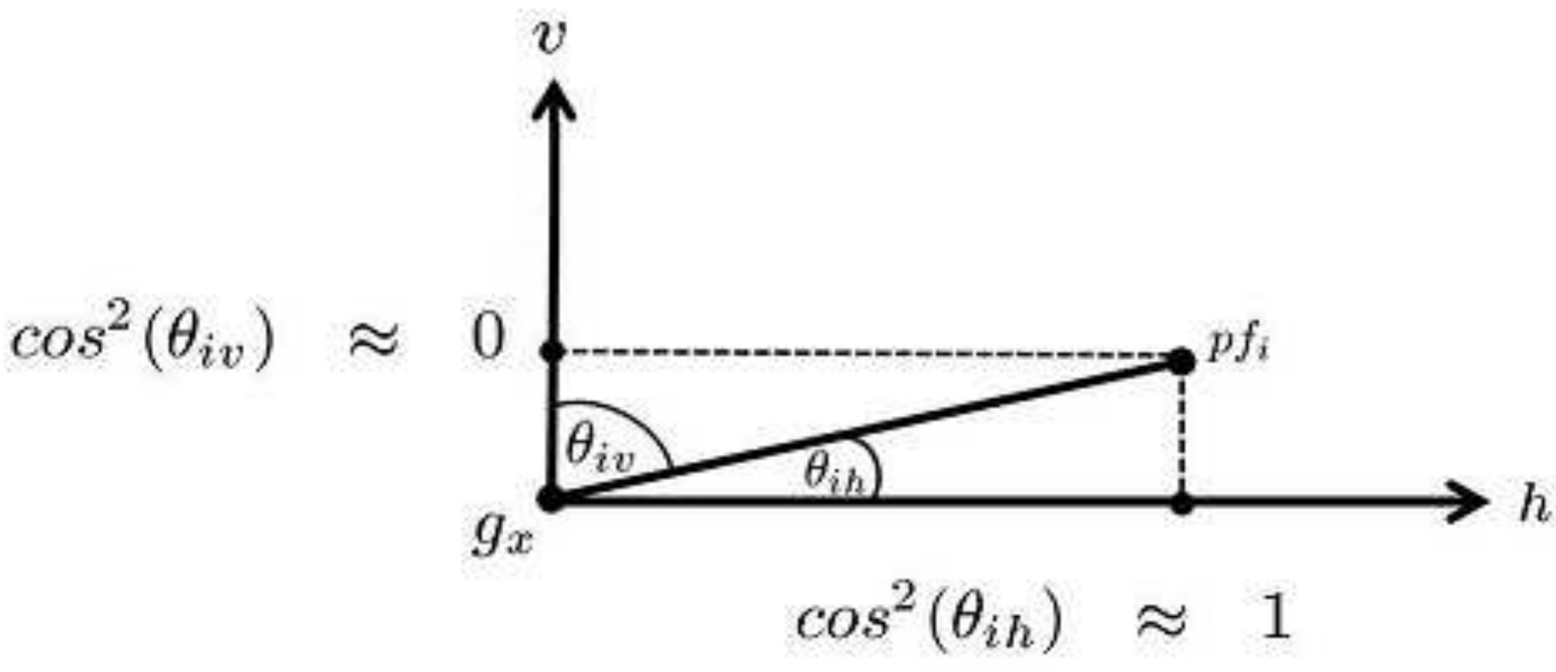


Tabla de contingencia - Adjetivos

- Marca: **Scarpa**
- Fechas: del 1° Enero de 2010 al 15 Junio de 2010
- Frecuencia mínima: 2

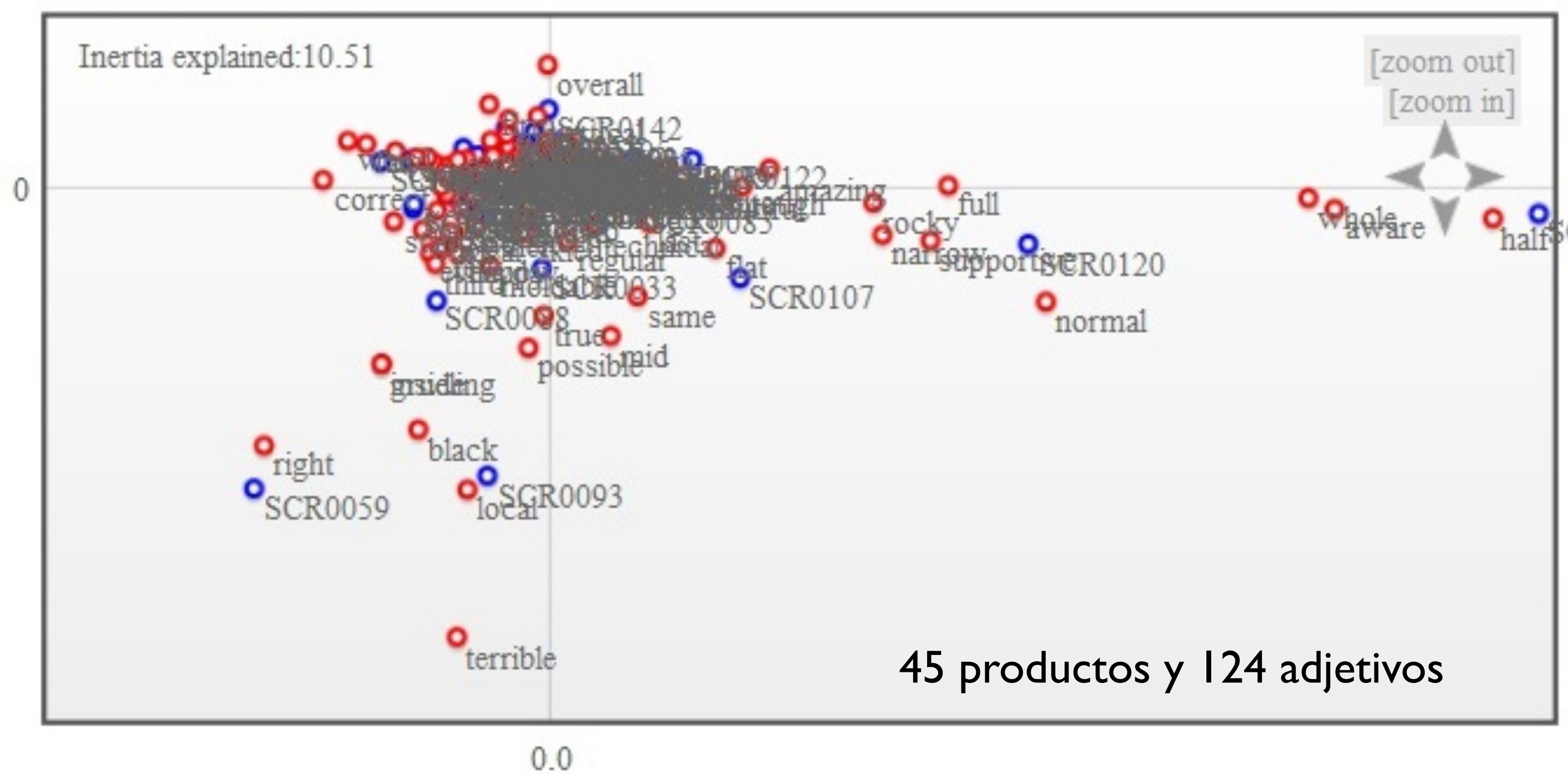
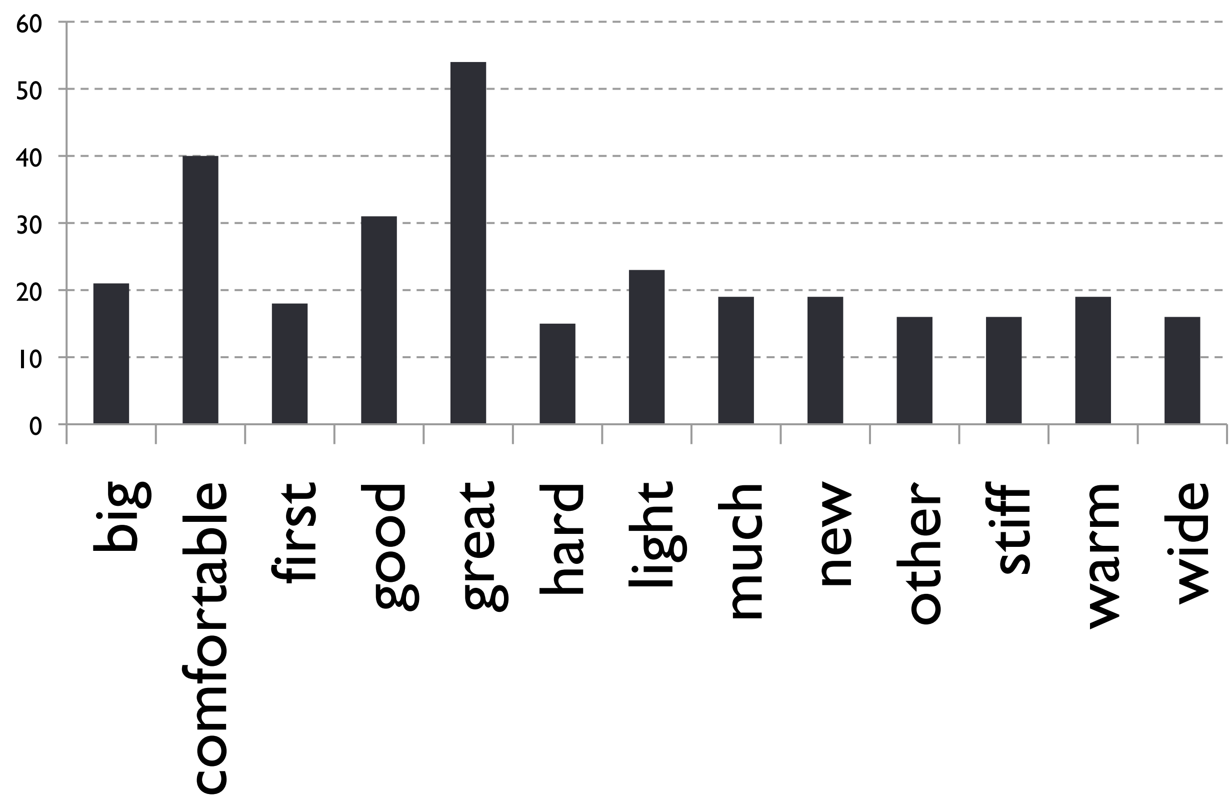


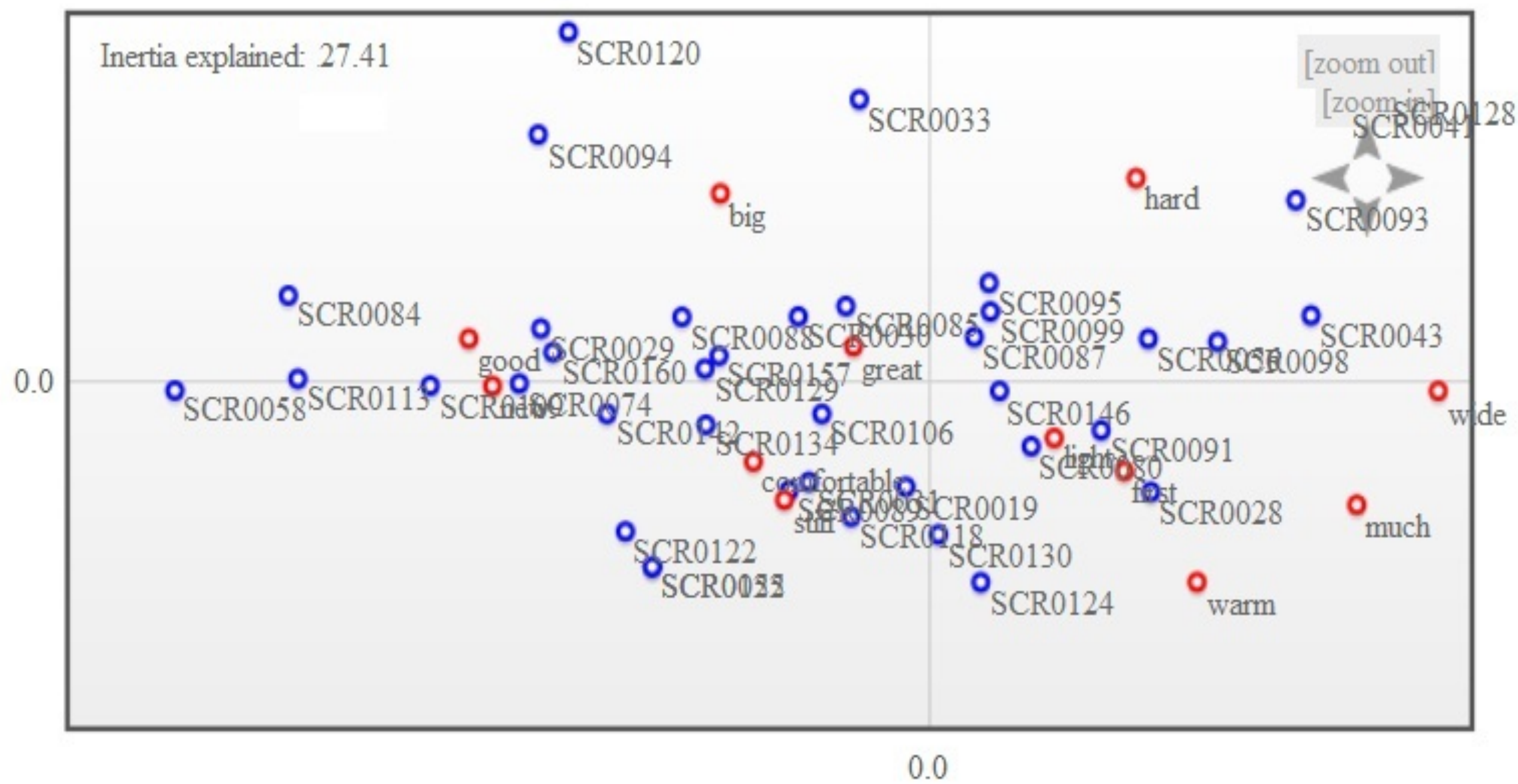
Tabla de contingencia - Adjetivos (cont.)

- Marca: **Scarpa**
- Fechas:
  - 1° Enero, 2010
  - 15 Junio, 2010
- Frecuencia mínima: 15



CIMPA-UCR

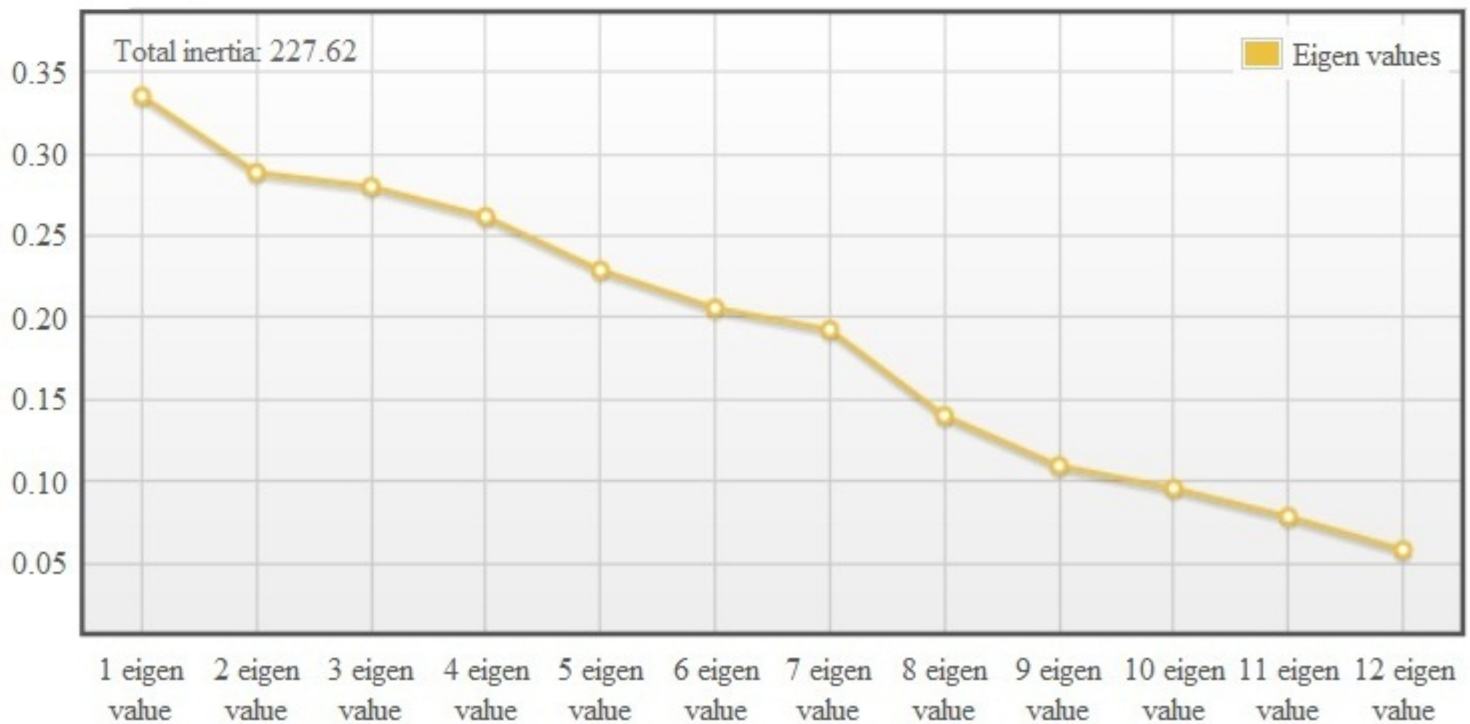
Primer plano factorial



44 productos and 13 adjetivos

CIMPA-UCR

Gráfico de los valores propios



CIMPA-UCR

Algunos casos a observar

- ***good – new***
  - SCR0084 - *Scarpa Techno Climbing Shoe*
  - SCR0058 - *Scarpa T-Race Telemark Ski Boot*
  - SCR0113 - *Scarpa Meridian GTX Hiking Shoe - Mens*
  - SCR0109 - *Scarpa Apex GTX Hiking Shoe - Mens*
  - SCR0074 - *Scarpa Zen Approach Shoe - Mens*
  - SCR0029 - *Scarpa Booster Climbing Shoe*
  - SCR0160 - *Scarpa Epic Trail Running Shoe - Mens*

CIMPA-UCR

Algunos casos a observar (cont.)

- ***Big***
  - SCR0094 - *Scarpa Terminator X Pro Telemark Ski Boot → big cuffs (parte alta de la bota)*
  - *SCR0033 - Scarpa Marathon Rock Climbing Shoe → relacionado con tamaño de la bota*
- ***Hard***
  - SCR0093 - *Scarpa T1 Lady Telemark Ski Boot - Women's*
  - *SCR0095 - Scarpa Terminator X Telemark Ski Boot*
  - *SCR0099 – Scarpa Kailash GTX Hiking Boot - Women's*

  → *Relacionadas con hard-core hiking, hard snow and hard groomers*

CIMPA-UCR

Algunos casos a observar (cont.)

- ***Warm, light***
  - SCR0028 - *Scarpa Phantom Lite Mountaineering Boot - Men's*
  - *SCR0124 - Scarpa Hurricane Alpine Touring Boot*
  - SCR0091 - *Scarpa Diva Alpine Touring Boot - Women's*
  - SCR0080 - *Scarpa Summit GTX Mountaineering Boot - Men's*

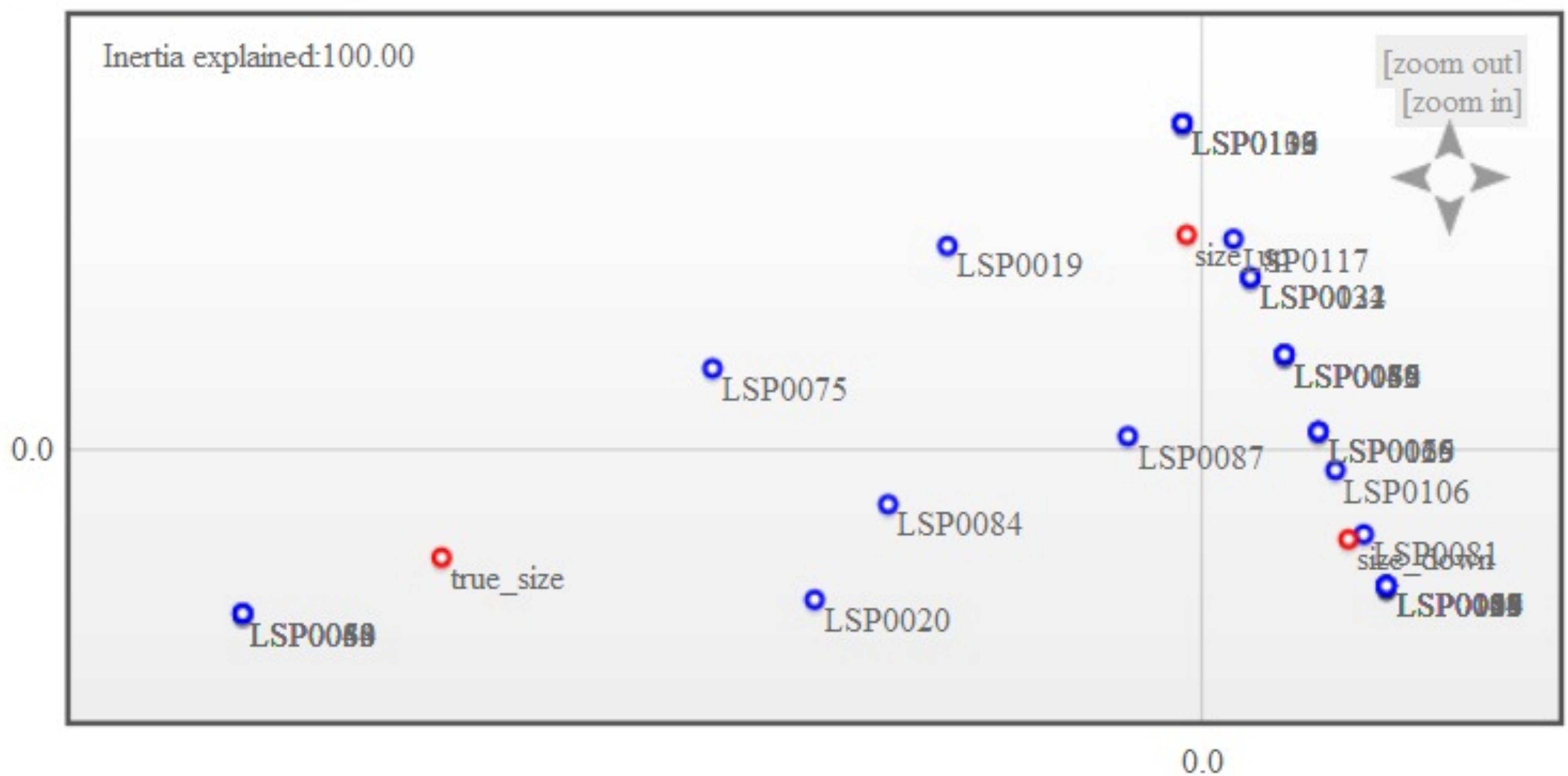
- *Estas botas son percibidas por los clientes como cálidas y livianas.*

Tabla de contingencia
Patrones de tamaño

- Marca: La Sportiva
- Fechas: Mayo, 2009 - Mayo, 2010
- Utiliza JAPE: *Java Annotation Pattern Engine*

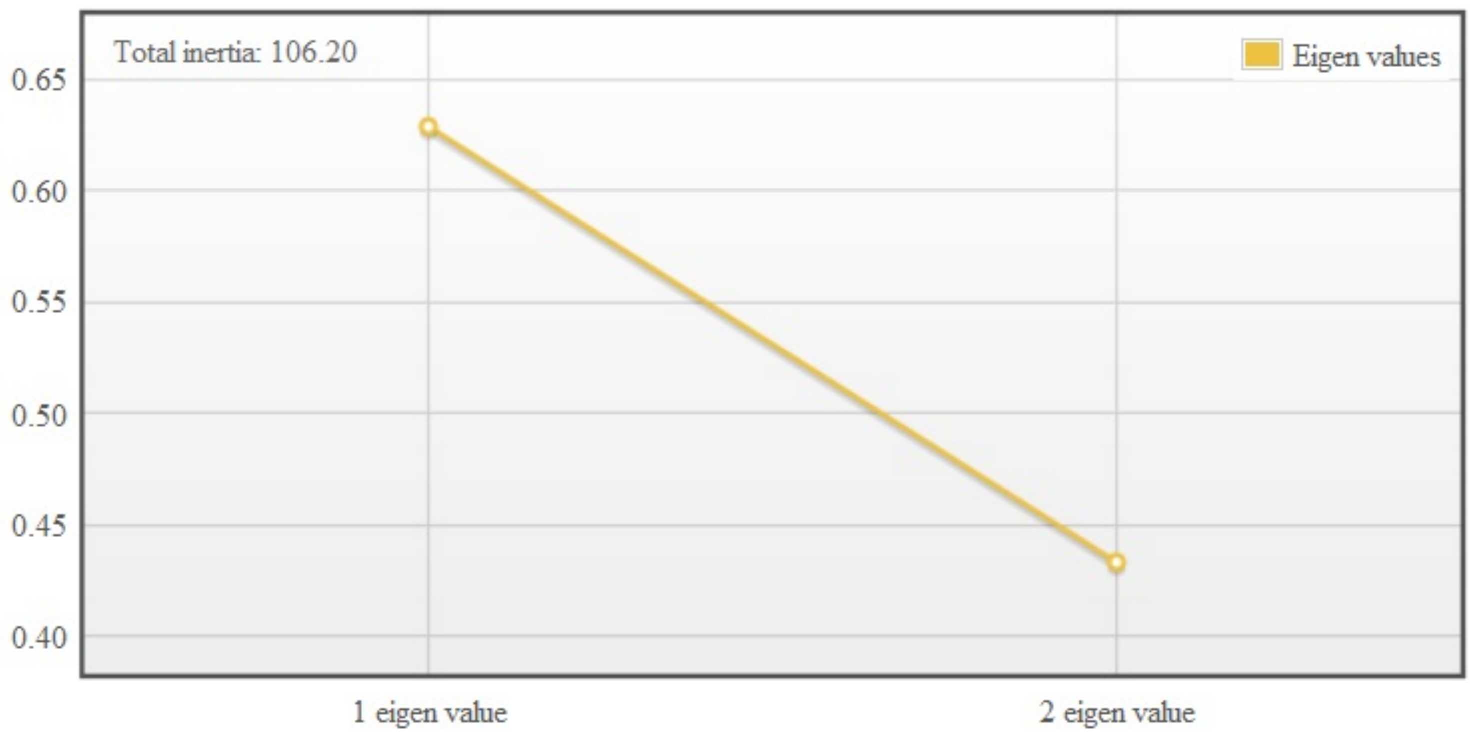
Size up	Size down	True to size
Go a size up	Go a size down	This is true-to-size
Buy a half size up,	Buy a half size down	Fits true to size
Order a full size bigger	Get 1/2 size down	Size is true to fit
Get 1/2 size up	Order down	Negations:
Order up	These run large	don´t go a size down
Order a 1/2 size larger		didn’t have to size up.
These run small		

Primer plano factorial



CIMPA-UCR

Gráfico de los valores propios



True to size (tamaño real)

– LSP0043 - *La Sportiva Glacier EVO **Mountaineering** Boot - Men's - 2007*

– LSP0058 - *La Sportiva Shark **Climbing** Shoe - Kids*

– LSP0064 - *La Sportiva Trango Trek GTX **Backpacking** Boot - Women's - 2007*

– LSP0080 - *La Sportiva Ultranord GTX-XCR Trail **Running** Shoes - Women's - 2007*

Size up (incrementar tamaño)

- LSP0031 - *La Sportiva Miura **Climbing** Shoe - Men's*
- LSP0102 - *La Sportiva Lynx Trail **Running** Shoe - Men's*
- LSP0113 - *La Sportiva FC 1.1 **Hiking** Shoe - Men's*
- LSP0117 - *La Sportiva Wildcat Trail **Running** Shoe - Men's*
- LSP0136 - *La Sportiva Baruntse **Mountaineering** Boot - Men's*
- LSP0139 - *La Sportiva Wildcat GTX Trail **Running** Shoe - Men's*

Size down (disminuir tamaño)

- LSP0015 - *La Sportiva Testarossa **Climbing** Shoe*
- LSP0016 - *La Sportiva Katana Rock **Climbing** Shoe - Women's*
- LSP0029 - *La Sportiva Nepal EVO GTX Mountaineering Boot - Men's*
- LSP0048 - *La Sportiva Mythos **Climbing** Shoe - Mens - 2007*
- LSP0054 - *La Sportiva Solution **Climbing** Shoe*
- LSP0055 - *La Sportiva Venom **Climbing** Shoe - 2007*
- LSP0069 - *La Sportiva Sandstone GTX-XCR Hiking Shoe - Men's*
- LSP0081 - *La Sportiva Miura VS **Climbing** Shoe*
- LSP0083 - *La Sportiva Mythos **Climbing** Shoe - Women's*
- LSP0085 - *La Sportiva Nago **Climbing** Shoe - Women's - 2008*
- LSP0104 - *La Sportiva Imogene Trail Running Shoe - Women's*
- LSP0107 - *La Sportiva Trango Extreme Evo Light GTX - Men's*
- LSP0157 - *La Sportiva Mythos **Climbing** Shoe - Men's*

CIMPA-UCR

Demostración

- Se implementó una aplicación web utilizando
  - Grails 1.3.2
  - Java 6
  - GATE
  - JAPE
  - PosgreSQL

Conclusiones

- Es posible aplicar AFC a las opiniones de Backcountry.com desde diversos ángulos
  - Adjetivos
  - Patrones de tamaño

- Trabajo futuro
  - Utilizar otros atributos para identificar patrones. Por ejemplo, la calidad, el peso, qué tan caliente es, etc.
  - Analizar por marca (perfiles filas serían marcas en lugar de productos)
  - Incorporar listas de palabras (inclusivas o exclusivas)
  - Analizar adjetivos utilizados en las descripciones actuales de los productos
    - Mejorar uso de palabras en el mercadeo

Conclusiones (cont.)

- Conocer las percepciones de los clientes a partir de sus propias palabras provee conocimiento nuevo y de mucho valor:
  - Reducción de devoluciones por problemas de tamaño
  - Creación de nuevos productos o aplicaciones en el sitio web para mejorar la experiencia de navegación y compras del usuario
  - Facilitar las decisiones de compra en la categoría de zapatos
  - Mejorar la satisfacción de los clientes

CIMPA-UCR

Clustering by moving the centers

(Clasificación moviendo los centros de clases)

Mario Villalobos Arias

CIMPA, Universidad de Costa Rica

Instituto Tecnológico de Costa Rica

Pasi 2011

Xalapa, México

Contents

- The partitioning problem
- The idea of moving the centers of the clusters
- Metaheuristics in partitioning
- Simulated annealing
- Particle swarm optimization (PSO)
- Some results
- Concluding remarks

Introduction

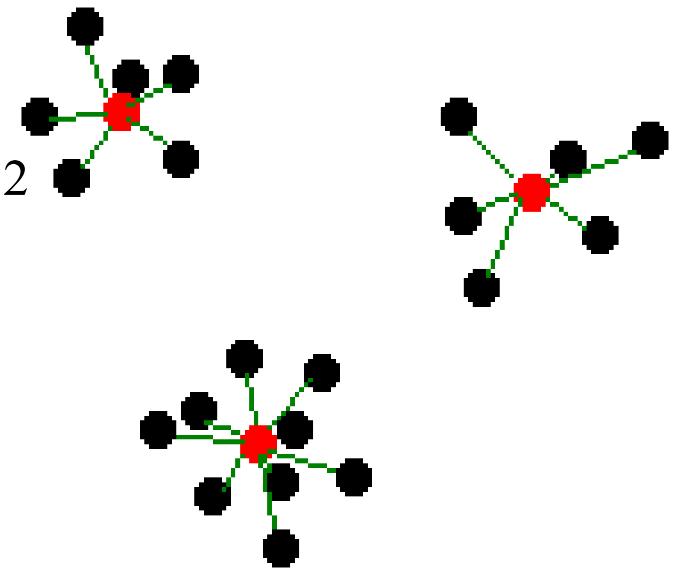
- Partitioning in Cluster Analysis
- We seek for K clusters, well separated and homogenous internally
- Objects: $\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbf{R}^p$
- Clusters: $C_1, C_2, \ldots, C_K$
- Partition: $P = (C_1, C_2, \ldots, C_K)$
- K given a priori

Homogeneity

- Minimize the within clusters variance:

$$W(P) = \frac{1}{n} \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - g_k \|^2$$

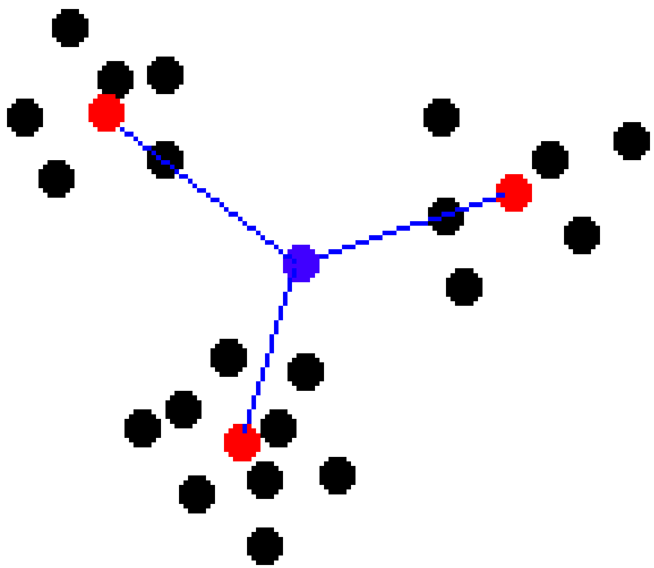
- where g_k is the barycenter of C_k



Separation

- Maximize the between-clusters variance:

$$B(P) = \sum_{k=1}^{K} \frac{|C_k|}{n} \| g_k - g \|^2$$



Remark:

$$Total = W(P) + B(P)$$

min max

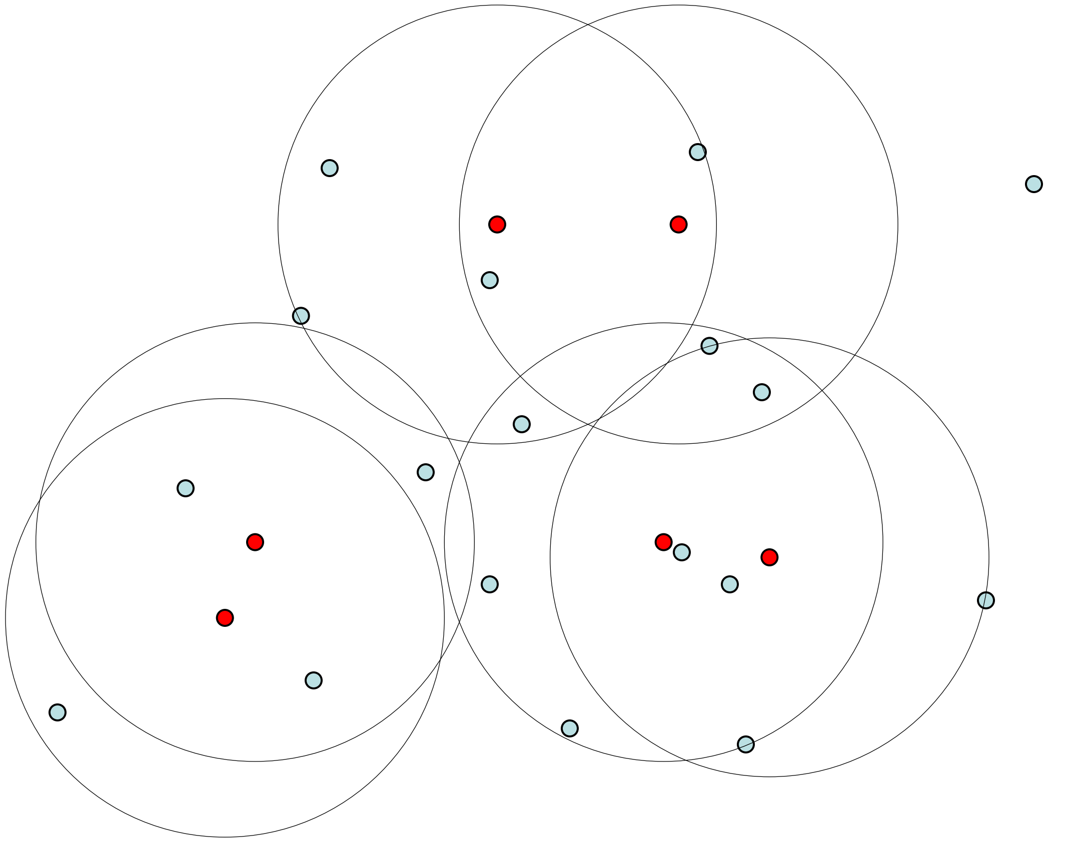
Combinatorial Problem

- Number of partitions in non empty clusters:

$$\begin{aligned} S(n,K) &= S(n-1,K-1) + K \cdot S(n-1,K) \\ &= \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\binom{K}{i} i^n \end{aligned}$$

- Optimal partition problem: NP-hard
- Approximative algorithms are needed

The idea of moving the centers of the clusters



Advantage:

there are fewer cluster than individuals

CIMPA-UCR

Metaheuristics in partitioning

- Classical methods (e.g. K-means) find local optima of *W* criterion

- We have applied several heuristics with good characteristics:
    a. Simulated annealing
    b. Genetic algorithms
    c. Tabu search
    d. Ant colonies

CIMPA-UCR

Simulated Annealing

- Introduced by Kirkpatrick, Gelatt y Vecchi (1982,1983) and independently by Cerny (1985) and Pincus (1968)
- It is based on “Annealing process” that produce low energy states of a solid when heated to high temperatures.
- It has two phases:
  - Increase the temperature until the solid melts.
  - Decreasing the temperature **slowly** until the particles are balanced to form a very pure crystal.

The Metropolis Algorithm

- Annealing has been modeled, Binder (1978)
- Monte Carlo technique : succession of states

$$(I,\ E_I) \rightarrow (J,\ E_J)$$

If $E_J - E_I < 0$ then J is acepted

Else J is acepted with probability

$$\exp\left(\frac{E_J - E_I}{\kappa_B T}\right)$$

T is temperature, κ_B Boltzman's constant

"Metropolis rule"

Simulated Annealing Algorithm

The solutions of the problem

$\Leftrightarrow$ states the physical system.

The value of the solution $\Leftrightarrow$ State energy

Optimization problem $\rightarrow$ (V, f). $I \in V$

J is a **neighbor** state I, $J \in V_I$

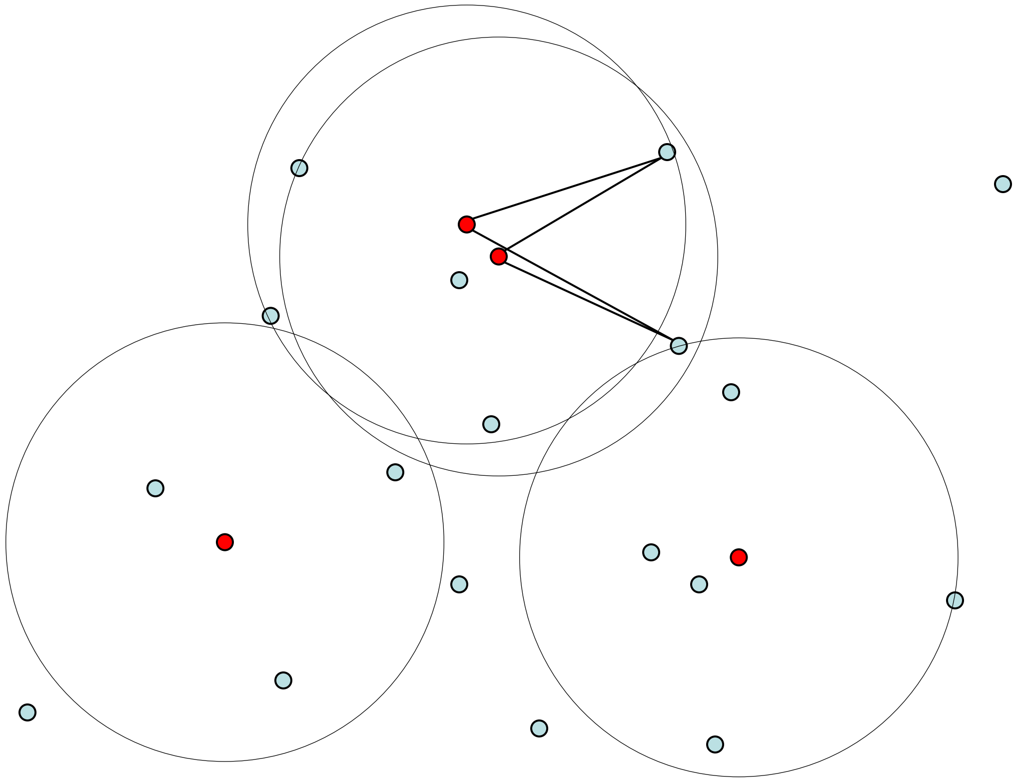
$$P_c\{\text{aceptar } J\} = \begin{cases} 1 & \text{si } f(J) \leq f(I) \\ \exp\left(\dfrac{f(I) - f(J)}{c}\right) & \text{si } f(J) > f(I) \end{cases}$$

Simulated Annealing Algorithm

```
PROCEDURE S_Simulado;

BEGIN
   Inicializar (I_o, c_o, L_o);
   k := 0;
   I := I_o;
   REPEAT
      FOR L := 1 TO L_k DO BEGIN
         Generar(J de V_I);
         IF E_J ≤ E_I THEN
            I := J;
         ELSE
            IF exp((E_I − E_J)/c_k) >random [0,1) THEN
               I := J;
      END;
      k := k + 1;
      Cacular_Largo(L_k);
      Cacular_Control(c_k);
   UNTIL Criterio_de_Parar;
END;
```

The idea of moving the centers of the clusters in SA



CIMPA-UCR

The idea of moving the centers of the clusters in SA

- compute the new state (moving a center).
- It is checked if this change altered the classes.
- if the individual is of the class of center moved, it have to change the minimum distance
- It should be checked if the center walked away from an individual and it changes to another class
- verify if individuals from other classes were moved to the class, of the center moved.

Advantage:

there are fewer cluster that individuals

Particle Swarm Optimization

- *Particle Swarm Optimization* (PSO):

  Kennedy & Eberhardt (1995)

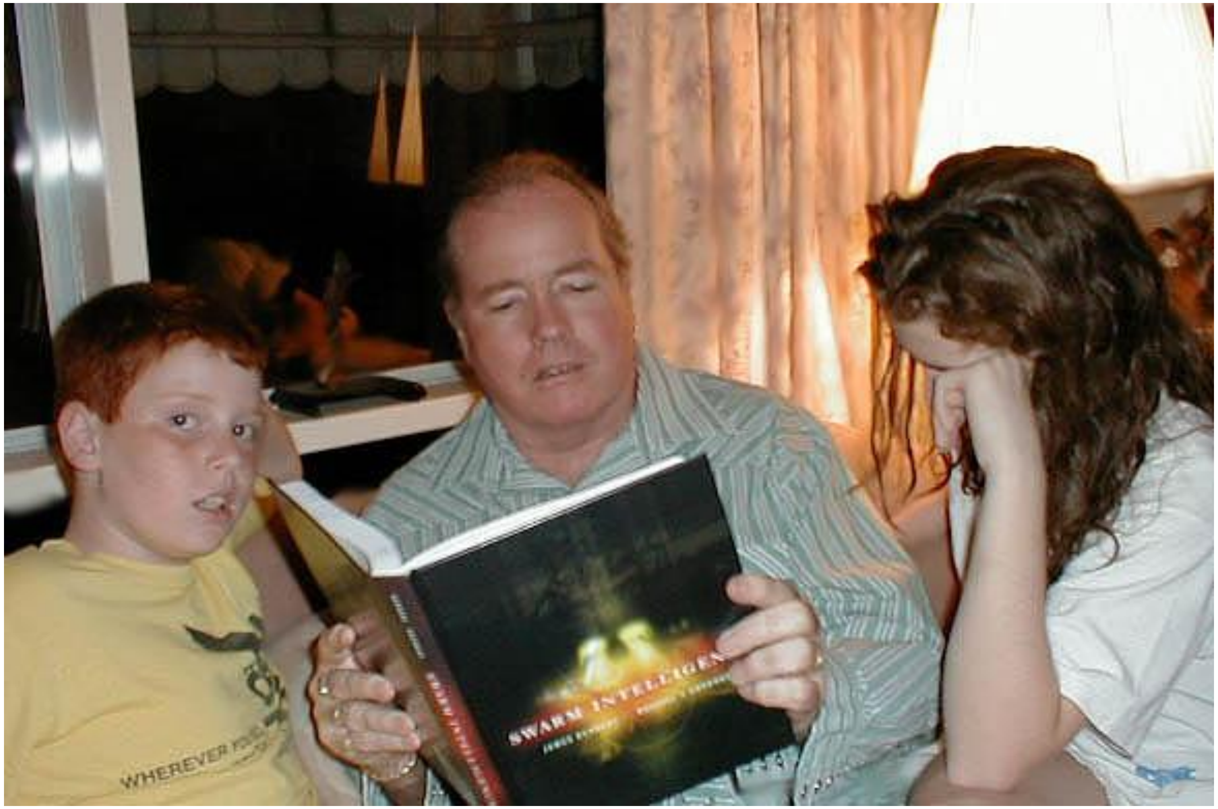
- http://www.particleswarm.net

- Models social behavior: each individual tries to perform better according to its own experience and looking at his neighbors' experience

- It handles a set of *M* particles or agents in a multidimensional space

CIMPA-UCR

The "inventores"



James Kennedy

Kennedy_Jim@bls.gov

CIMPA-UCR

The “inventores”

Russell
Eberhart

eberhart@engr.iupui.edu

Principles of PSO

For each particle:

- Remember my best position in the past
- Ask the neighbors for their best position

Tendencies:

1. Follow my path (*inertia*)
2. Go back to my best position (*conservative*)
3. Imitate the leader (*imitation*)

CIMPA-UCR

PSO: modeling

$$
\begin{aligned}
v(t+1) &= \alpha\, v(t) + \lambda_1\Big(z_m^*(t) - z_m(t)\Big) + \\
&\qquad + \lambda_2\Big(z^*(t) - z_m(t)\Big), \\
z_m(t+1) &= z_m(t) + v_m(t+1),
\end{aligned}
$$

where:

$$\lambda_1 = \mathrm{rand}\,(0, \varphi_1) \quad , \quad \lambda_2 = \mathrm{rand}\,(0, \varphi_2)$$

$$\varphi = \varphi_1 + \varphi_2$$

PSO Model

$$
\begin{aligned}
v_m(t+1) \;=\;& \alpha\, v_m(t) + \\
& +\mathrm{rand}\,(0,\varphi_1)\cdot\left(z_m^*(t) - z_m(t)\right) + \\
& +\mathrm{rand}\,(0,\varphi_2)\cdot\left(z^*(t) - z_m(t)\right)
\end{aligned}
$$

Inertia

Go back

Imitation

$$\boxed{z_m(t+1) = z_m(t) + v_m(t+1)}$$

CIMPA-UCR

Illustration of PSO

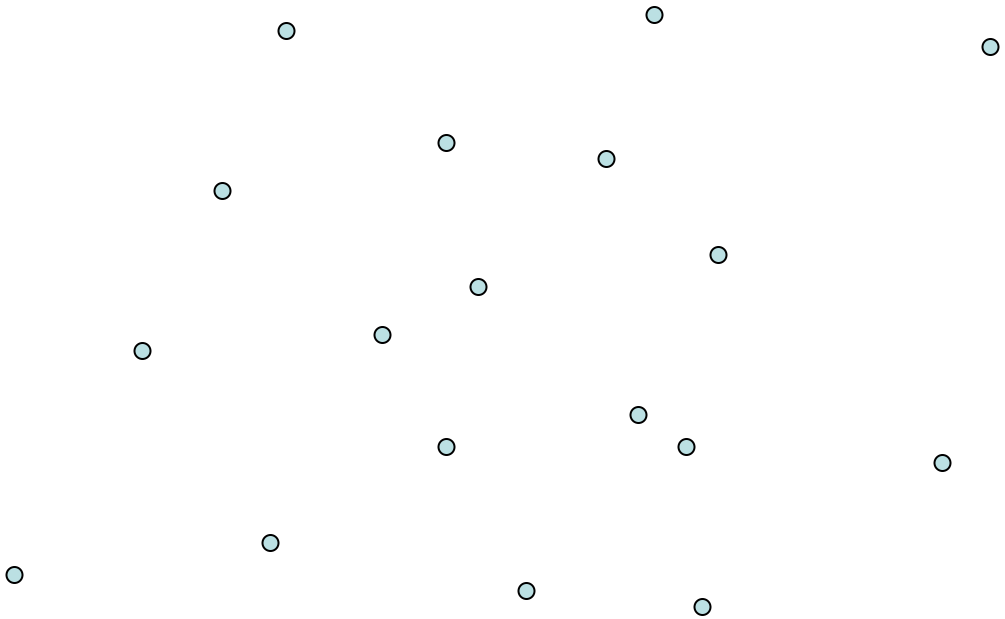
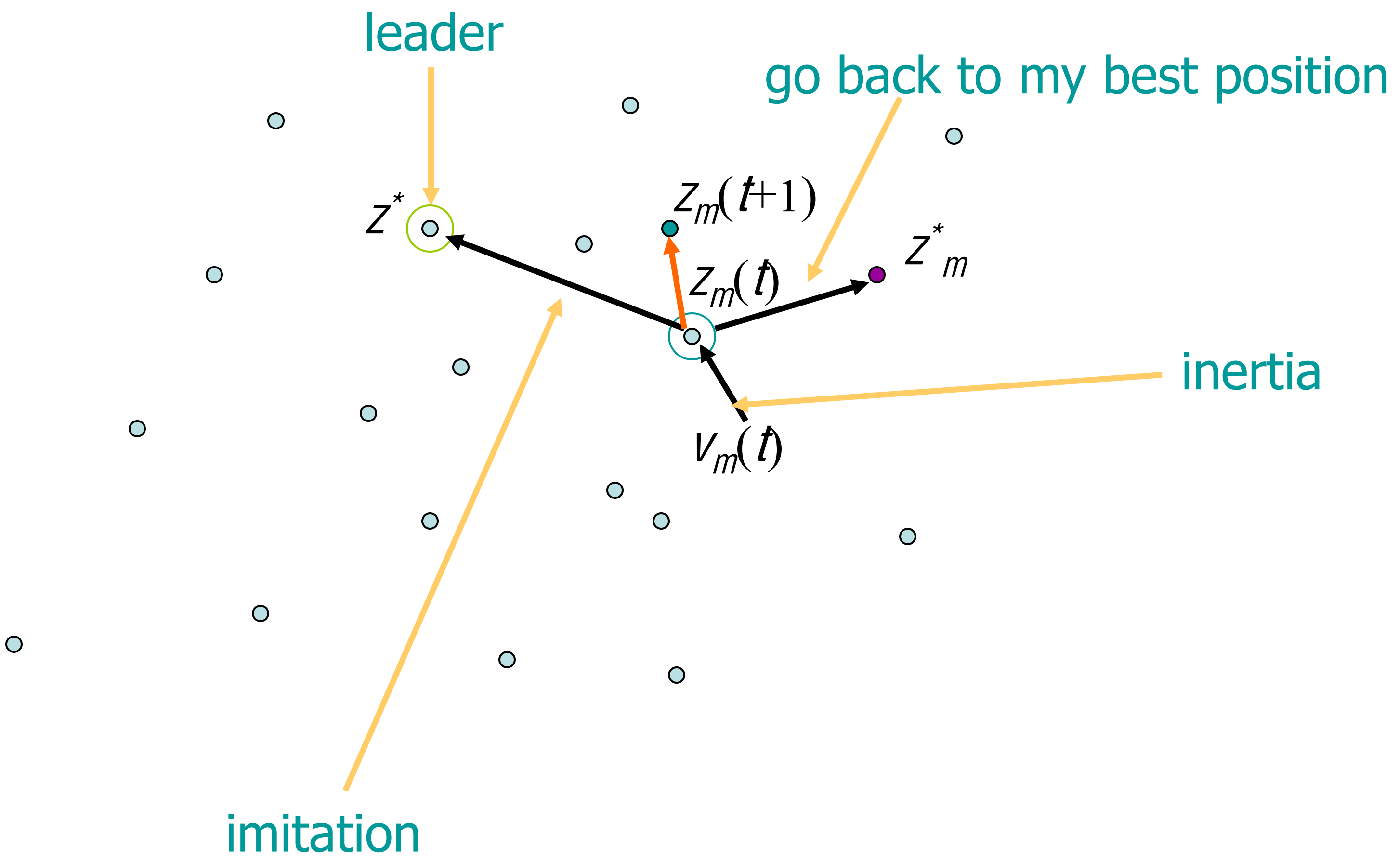
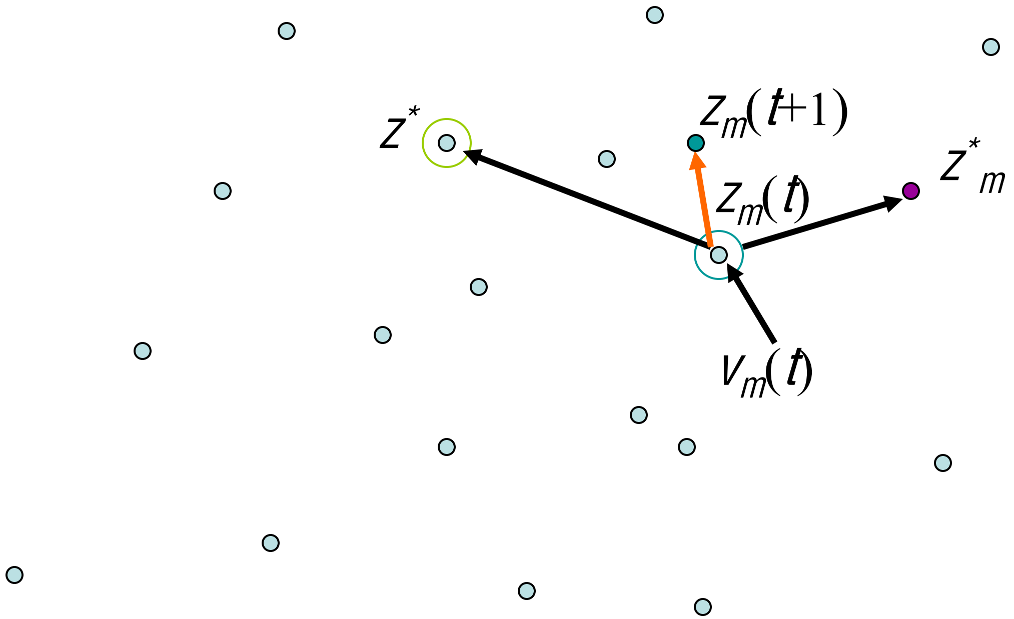


Illustration of PSO



CIMPA-UCR

Illustration of PSO



$$v_m(t+1) = \alpha\, v_m(t) + \lambda_1 \left[z^*_m(t) - z_m(t) \right] + \lambda_2 \left[z^*(t) - z_m(t) \right]$$

$$z_m(t+1) = z_m(t) + v_m(t+1)$$

CIMPA-UCR

Conditions

Define:

$$
\begin{aligned}
\kappa &= \left[(e-2)(\varphi-1)-1\right]\left[1+\frac{\sqrt{\varphi^2-4\varphi}}{\varphi-2}\right] \\
\chi &= \frac{2\kappa}{\varphi-2+\sqrt{\varphi^2-4\varphi}}
\end{aligned}
$$

Heuristics, partial results: $\alpha = e - 2 \approx 0.718$

Conditions for non divergence: $\left\{\begin{aligned} \kappa &\in (0,1) \\ \alpha &= \frac{1+\chi\cdot(\varphi-2)}{\varphi-1} \end{aligned}\right.$

Empirical result:

$$\boxed{\kappa \geq 0.577 \quad , \quad 4 < \varphi \leq 4.276}$$

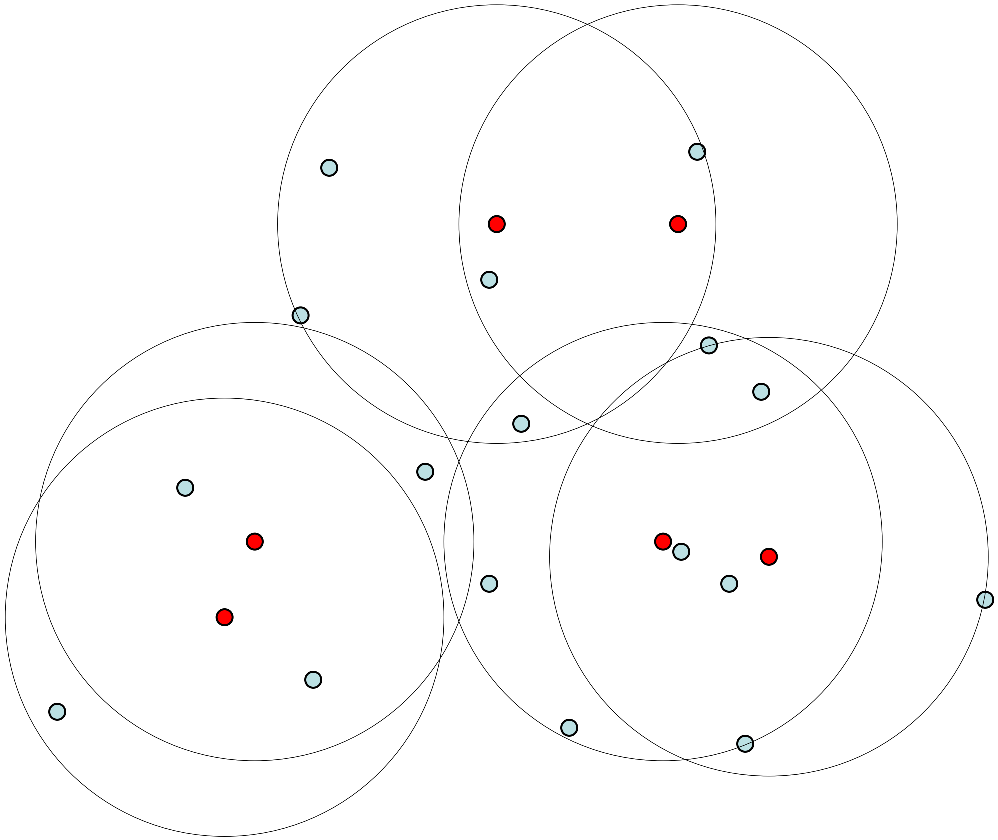
CIMPA-UCR

PSO in partitioning

- Each particle is a set of K centroids
$$g_1,\dots g_K \in \mathrm{R}^p$$
- Each centroid has associated a cluster C_k by allocating the objects to the cluster of the nearest centroid
- The centroids move in R^p according to the principles of PSO and the partition is redefined by allocation
- Particles move in R^{Kp}

CIMPA-UCR

The idea of moving the centers of the clusters in PSO



Algorithm

Repeat for t=1,2,... max_iter **or** conv

For m=1: M (*particles*)

For k=1: K (*clusters*)

use PSO equation for each numerical variable

Allocate objects to the nearest centroid

Update best position of particle m

Update best overall particle

CIMPA-UCR

Parameters

- α = 0.5: coefficient for $V_i(t)$
- $\varphi = \varphi_1 + \varphi_2 \geq 4$: sum of random coefficients
- ε (tolerance): for convergence
- Maximum # of iterations (= 200)

CIMPA-UCR

Preliminary results

Four data tables from the literature:

- French scholar notes (9 x 5)
- Amiard's fishes (23 x 16)
- Thomas' sociomatrix (24 x 24)
- Fisher's iris (150 x 4)

CIMPA-UCR

Methods compared

- PSO: Particle swarm optimization
- MC:Moving centers + SA
- SA: Simulated annealing
- TS: Tabu search
- GA: Genetic algorithm
- ACO: Ant colony optimization
- KM: *k*-means (local search)
- Ward: Hierarchical agglomerative (cut the tree)

CIMPA-UCR

Results: Scholar Notes

Table 9 x 5

K	*W*	MC #10+	PSO #=100	SA #=150	TS #=1000	GA #=100	ACO #=25	KM #=10000	Ward
2	28.2	61	92	100	100	100	100	12	No
3	16.8	95	57	100	100	95	100	12	No
4	10.5	48	51	100	100	97	100	5	Yes
5	4.9	100	29	100	100	100	100	8	Yes

Results: Amiard's fishes

Table 23 x 16

K	*W*	MC # =10	PSO #=100	SA #=150	TS #=200	GA #=100	ACO #=25	KM #=10000	Ward
3	32213	90	51	100	100	87	100	8	No
4	18281	40	23	100	100	0	100	9	No
5	14497	10	6	100	97	0	68	1	Yes

Results: Thomas' sociomatrix

Table 24 x 24

K	*W*	MC #=10	PSO #=100	SA #=150	TS #=200	GA #=100	ACO #=25	KM #=10000	Ward
3	271	90	7	100	100	85	100	2	No
4	235	40	7	100	100	24	96	0.15	No
5	202	40	7	100	98	0	84	0.02	No

Results: Fisher's iris

Table 150 x 4

K	*W*	MC #=10	PSO #=100	SA #=150	TS #=1000	GA #=100	ACO #=25	KM #=10000	Ward
2	0.99	100	76	100	100	100	100	100	No
3	0.52	60	79	100	76	100	100	4	No
4	0.38	40	55	55	60	82	100	1	No
5	0.32 0.312	? 10	28	0	32	6	100	0.24	No

CIMPA-UCR

Concluding remarks

- Generally, results are better or at least the same as those of classical methods
- Simulated annealing and ant colonies are generally better
- Some work on tuning the parameters must be done:

  α, φ, max_iter, population size

Concluding remarks

Monte-Carlo simulation:

Gaussian random variables $N(\mu,\sigma^2)$

- Number of objects (50,100,200,500)
- Number of clusters (3,5,7)
- Variance σ^2 (similar, different)
- Cardinalities of clusters (similar, different)

CIMPA-UCR

Some References

- Trejos, J.; Murillo, A.; Piza, E. (1998) "Global stochastic optimization techniques applied to partitioning", in: A. Rizzi et al. (Eds.) *Advances in Data Science and Classification*, Springer, Berlin: 185-190.
- Trejos, J.; Murillo, A.; Piza, E. (2004) "Clustering by Ant Colony Optimization", In: D. Banks et (Eds.) *Classification, Clustering, and Data Mining Applications*, Springer, Berlin, 25-32.

CIMPA-UCR

Questions?

- E-mail:

  mario.villalobos@ucr.ac.cr, mario.cr@gmail.com

Thank you!

Optimización multi-objetivo en Clasificación algunas ideas

Mario Alberto Villalobos Arias

Escuela de Matemática y CIMPA

Universidad de Costa Rica

mario.villalobos@ucr.ac.cr, mario.cr@gmail.com

Pasi 2011

Xalapa, México

Organización

- Introducción – Problema Multiobjetivo.
- MO en clasificación
- Trabajo futuro.

Problema de optimización multiobjetivo

Ejemplo: Selección óptima de portafolios de inversión.

El inversionista quiere

- maximizar su rendimiento y
- minimizar el riesgo.

Riesgo $\longrightarrow$ Markowitz, Target Shortfall Probability (TSP)

Problemas con los datos en clasificación

- las variables no son del mismo tipo
- las dimensiones de estas

Problema de optimización multiobjetivo

En general, en un PMO tenemos:

$$F(x) = (f_1(x), ..., f_n(x)).$$

Enfoque usual: $\longrightarrow$ un problema escalar o de un objetivo.

Esto puede no tener sentido si $f_1, ..., f_n$ no son del mismo tipo.

Por ejemplo: si f_1 denota distancia, f_2 denota tiempo, ...

La escalarización puede perder sentido.

Nuestro enfoque: **Tratar el PMO directamente.**

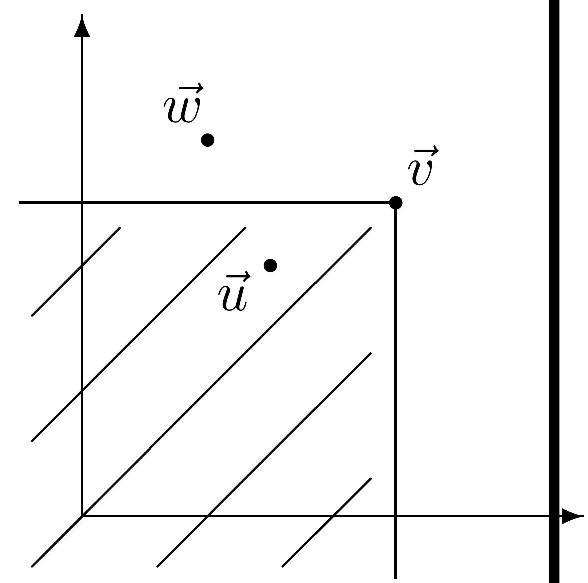
Para comparar vectores en $\mathbb{R}^d$ $\longrightarrow$ **orden de Pareto.**

$\vec{u} = (u_1, u_2, \ldots, u_d)$, $\vec{v} = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$, entonces

$$\vec{u} \preceq \vec{v} \iff u_i \leq v_i \;\forall\; i \in \{1, \ldots, d\}.$$

Orden parcial . $\vec{u} \prec \vec{v} \iff \vec{u} \preceq \vec{v}$ y $\vec{u} \neq \vec{v}$.

$\vec{v}$ domina a $\vec{u}$



$F : X \longrightarrow \mathbb{R}^d$ función vectorial,

$f_i : X \longrightarrow \mathbb{R}$ para cada $i \in \{1, \ldots, d\}$.

El problema de optimización multiobjetivo es:

Encontrar $x^* \in X$ tal que

$$F(x^*) = \min_{x \in X} F(x) = \min_{x \in X} [f_1(x), \ldots, f_n(x)], \tag{1}$$

Definición 1:

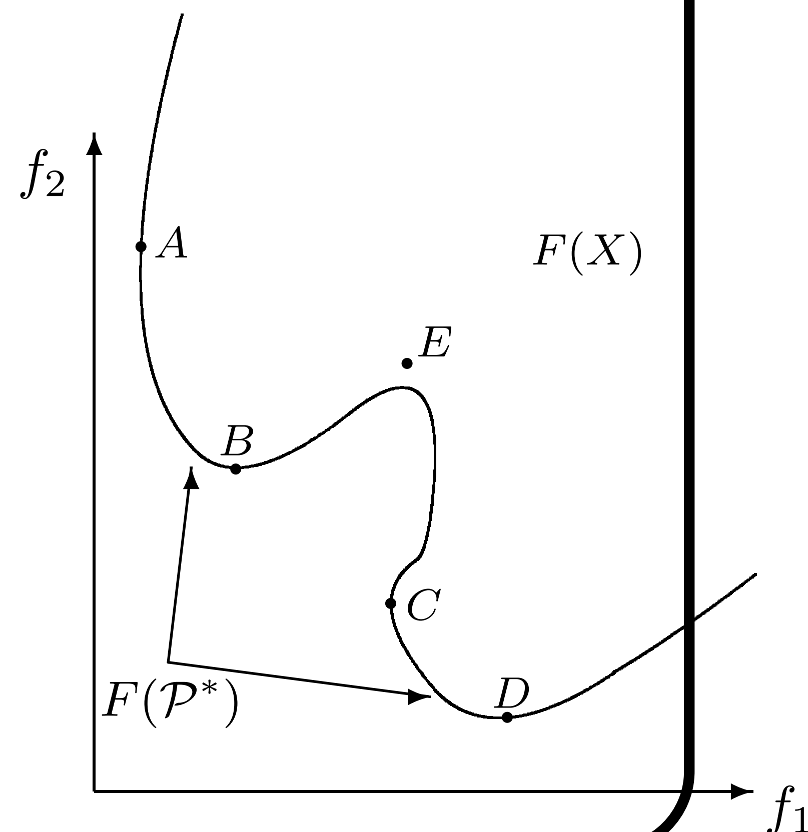
$x^* \in X$ se llama una **solución óptima de Pareto** para MOP (1) si no existe $x \in X$ tal que $F(x) \prec F(x^*)$. Sea

$$\mathcal{P}^* = \{x \in X \ : \ x \text{ es una solución óptima de Pareto}\}$$

el conjunto de óptimos de Pareto, y

$$F(\mathcal{P}^*) \ := \ \{F(x) \ : \ x \in \mathcal{P}^*\}$$

el frente de Pareto.



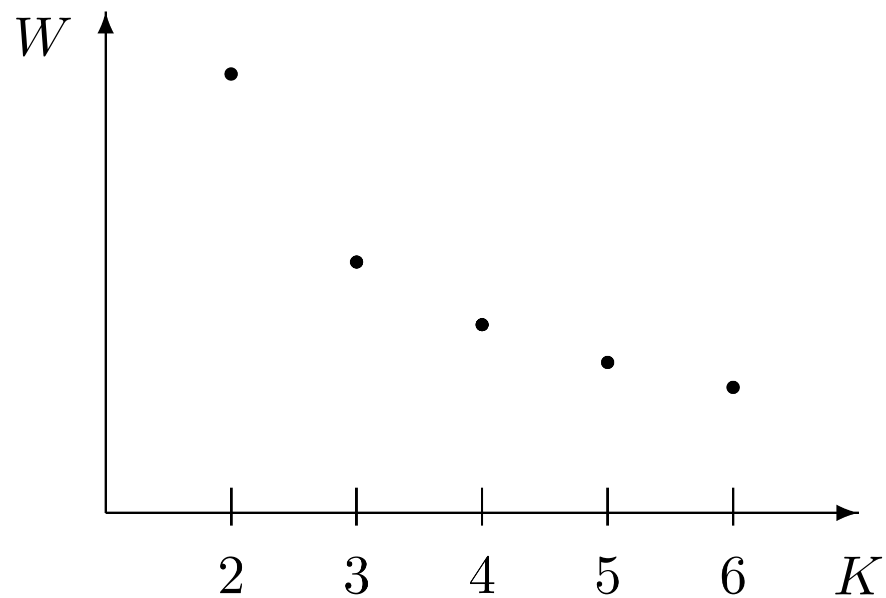
Ideas de uso de Optimización MO en clasificación

- Tomar

. $f_1 =$ Número de clases

. $f_2 = W$ (inercia)

llevar a encontrar los óptimos de los diferentes números de clases



Ideas de uso de Optimización MO en clasificación

- dividir W (inercia) en partes de acuerdo al tipo de las variables

$$W = (W_1, W_2, \ldots)$$

debería encontrar la solución clásica y elimina el efecto por las diferencias entre las variables

Trabajo Futuro

- programar los algoritmos presentados
- buscar otras posibilidades de usar optimización MO

Preguntas

gracias