Computer Science Laboratory of
Paris Nord University

Paris 13 University - Institut Galilée - LIPN, UMR 7030 du CNRS
99 Avenue J-B. Clément - 93430 Villetaneuse - France

Advances in unsupervised dimensionality reduction

through topological clustering and variable weighting

Younès BENNANI, Guénaël CABANES and Nistor GROZAVU

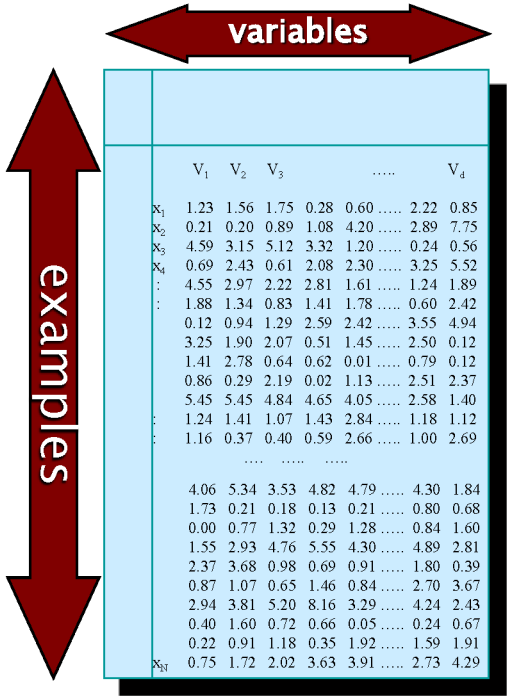
Plan

- Introduction & Problematic
- Part I
  - Reducing de number of columns
    - Variables selection through weighting distance (lwd-SOM)
    - Variables selection through weighting observations (lwo-SOM)
- Part II
  - Reducing de number of rows
    - A Simultaneous Two-Level Clustering Algorithm for Automatic Model Selection (S2L-SOM)
    - A Local Density-Based Simultaneous Two-Level Algorithm for Topographic Clustering (DS2L-SOM)
- Conclusion

Data dimensionality

The data size can be measured in two ways:
- The number of variables (d),
- The number of examples (N).

N and d can take very high values , which may present a major obstacle for machine learning.



	V_1	V_2	V_3			.....		V_d
x_1	1.23	1.56	1.75	0.28	0.60	.....	2.22	0.85
x_2	0.21	0.20	0.89	1.08	4.20	.....	2.89	7.75
x_3	4.59	3.15	5.12	3.32	1.20	.....	0.24	0.56
x_4	0.69	2.43	0.61	2.08	2.30	.....	3.25	5.52
:	4.55	2.97	2.22	2.81	1.61	.....	1.24	1.89
:	1.88	1.34	0.83	1.41	1.78	.....	0.60	2.42
	0.12	0.94	1.29	2.59	2.42	.....	3.55	4.94
	3.25	1.90	2.07	0.51	1.45	.....	2.50	0.12
	1.41	2.78	0.64	0.62	0.01	.....	0.79	0.12
	0.86	0.29	2.19	0.02	1.13	.....	2.51	2.37
	5.45	5.45	4.84	4.65	4.05	.....	2.58	1.40
:	1.24	1.41	1.07	1.43	2.84	.....	1.18	1.12
:	1.16	0.37	0.40	0.59	2.66	.....	1.00	2.69
		....	.....	.....				
	4.06	5.34	3.53	4.82	4.79	.....	4.30	1.84
	1.73	0.21	0.18	0.13	0.21	.....	0.80	0.68
	0.00	0.77	1.32	0.29	1.28	.....	0.84	1.60
	1.55	2.93	4.76	5.55	4.30	.....	4.89	2.81
	2.37	3.68	0.98	0.69	0.91	.....	1.80	0.39
	0.87	1.07	0.65	1.46	0.84	.....	2.70	3.67
	2.94	3.81	5.20	8.16	3.29	.....	4.24	2.43
	0.40	1.60	0.72	0.66	0.05	.....	0.24	0.67
	0.22	0.91	1.18	0.35	1.92	.....	1.59	1.91
x_N	0.75	1.72	2.02	3.63	3.91	.....	2.73	4.29

Curse of dimensionality

Term coined by Richard Bellman * (1961) to describe the problem of explosive growth in data volume associated with adding extra dimensions in a mathematical space.

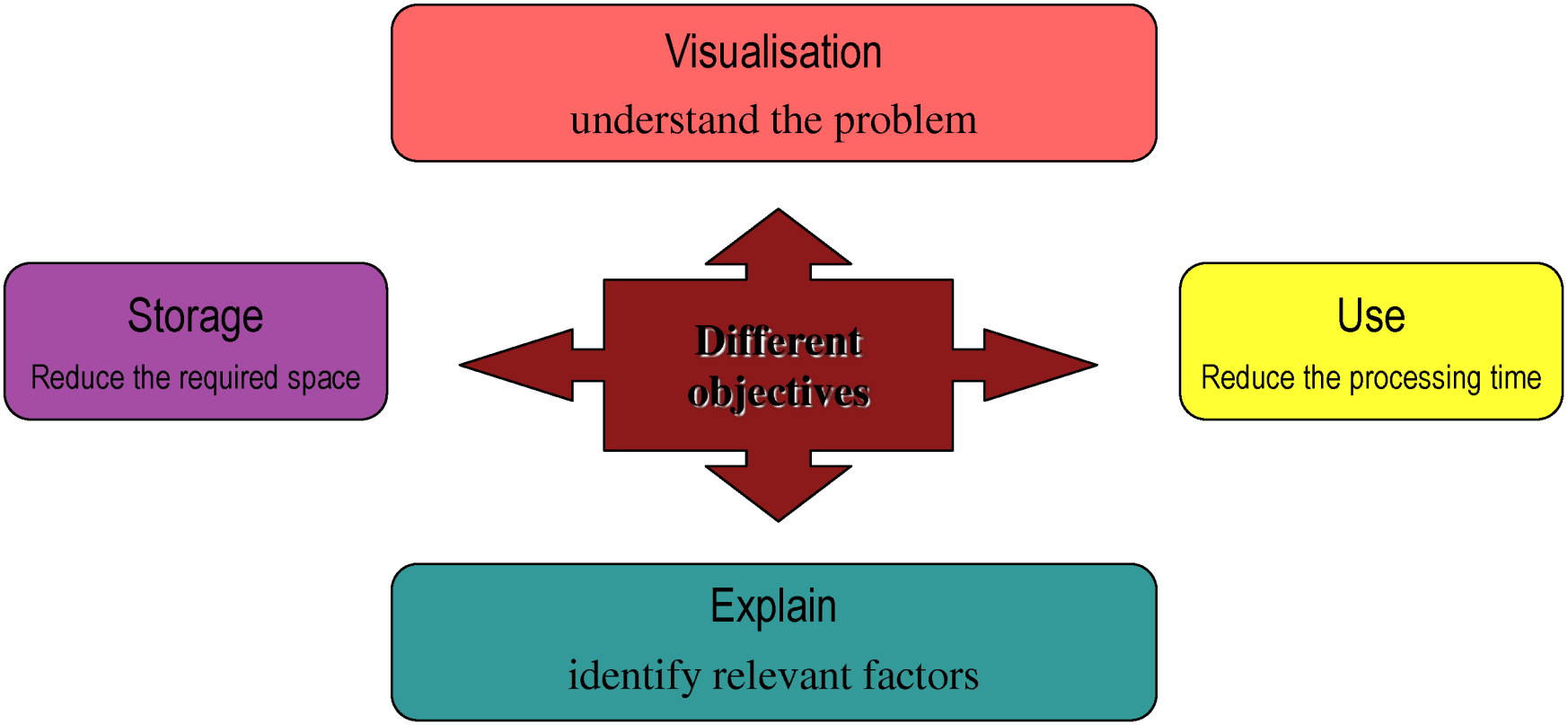
In non-parametric classification:

- Practical point of view:
it takes a lot of observations to estimate properly a function of several variables.

- Theoretical point of view:
estimation error of a density of d variables from N observations is of the order $N^{-1/(d+2)}$

* Bellman R. (1961) : « Adaptive Control Processes: A Guided Tour »
Princeton University Press.

Why to reduce the dimensionality ?



Dimensionality reduction

Dimentionality reduction:
Two dual problems

N → M / M<<N

examples

Construction

"Clustering"

Selection

Sampling

d → p / p<<d

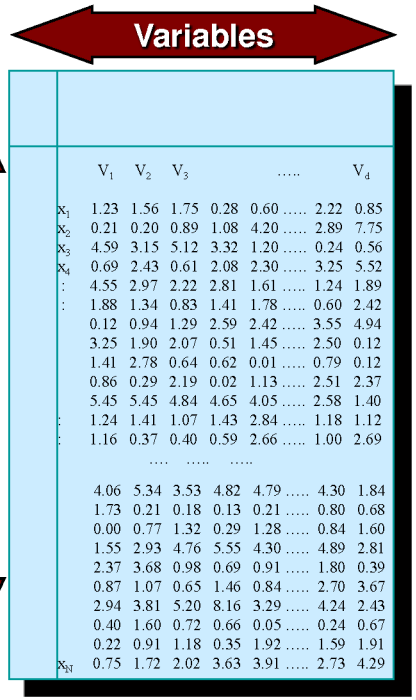
Variables

variables

Construction

Selection

Examples



	V_1	V_2	V_3			.....	V_d
x_1	1.23	1.56	1.75	0.28	0.60	.....	2.22	0.85
x_2	0.21	0.20	0.89	1.08	4.20	.....	2.89	7.75
x_3	4.59	3.15	5.12	3.32	1.20	.....	0.24	0.56
x_4	0.69	2.43	0.61	2.08	2.30	.....	3.25	5.52
:	4.55	2.97	2.22	2.81	1.61	.....	1.24	1.89
:	1.88	1.34	0.83	1.41	1.78	.....	0.60	2.42
	0.12	0.94	1.29	2.59	2.42	.....	3.55	4.94
	3.25	1.90	2.07	0.51	1.45	.....	2.50	0.12
	1.41	2.78	0.64	0.62	0.01	.....	0.79	0.12
	0.86	0.29	2.19	0.02	1.13	.....	2.51	2.37
	5.45	5.45	4.84	4.65	4.05	.....	2.58	1.40
:	1.24	1.41	1.07	1.43	2.84	.....	1.18	1.12
:	1.16	0.37	0.40	0.59	2.66	.....	1.00	2.69
		....	.....	.....				
	4.06	5.34	3.53	4.82	4.79	.....	4.30	1.84
	1.73	0.21	0.18	0.13	0.21	.....	0.80	0.68
	0.00	0.77	1.32	0.29	1.28	.....	0.84	1.60
	1.55	2.93	4.76	5.55	4.30	.....	4.89	2.81
	2.37	3.68	0.98	0.69	0.91	.....	1.80	0.39
	0.87	1.07	0.65	1.46	0.84	.....	2.70	3.67
	2.94	3.81	5.20	8.16	3.29	.....	4.24	2.43
	0.40	1.60	0.72	0.66	0.05	.....	0.24	0.67
	0.22	0.91	1.18	0.35	1.92	.....	1.59	1.91
x_N	0.75	1.72	2.02	3.63	3.91	.....	2.73	4.29

Dimentionality reduction approaches

- Clustering
  - Statistical methods, connexionism
- Selection / features weighting
  - Statistical approaches
- Features extraction
  - Linear methods

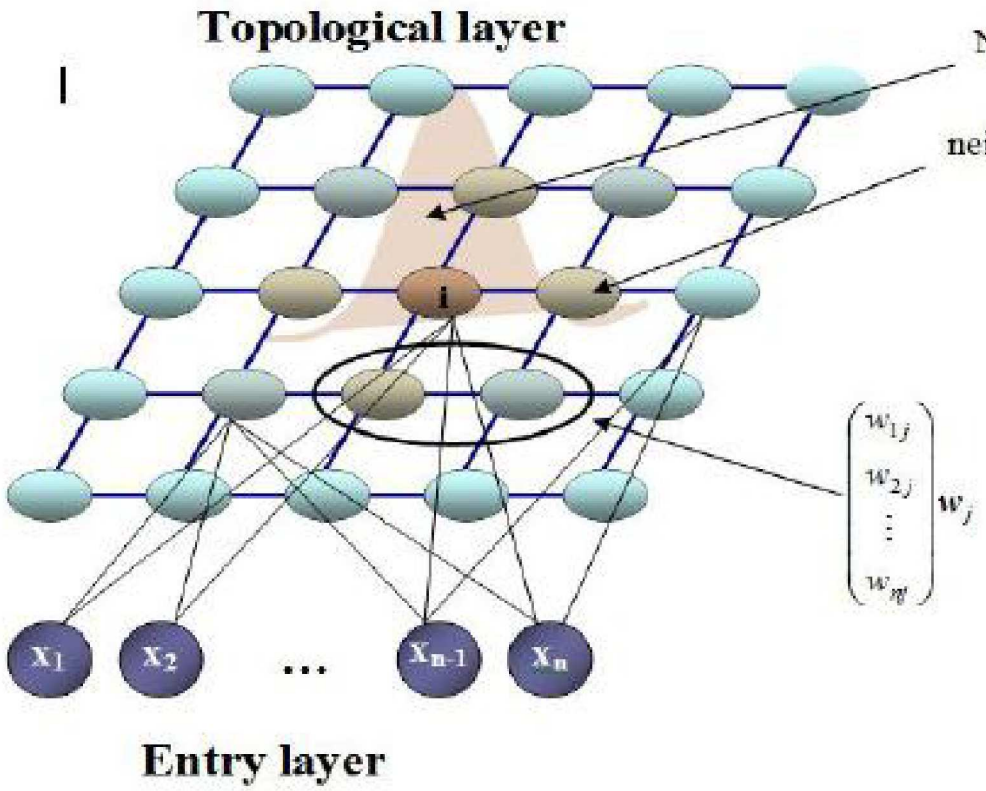
    Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS)
  - Non-linear methods

    Kernel PCA, Isometric feature mapping (Isomap), Locally Linear Embedding (LLE)

Part I
Reducing the number of columns

Unsupervised features weighting

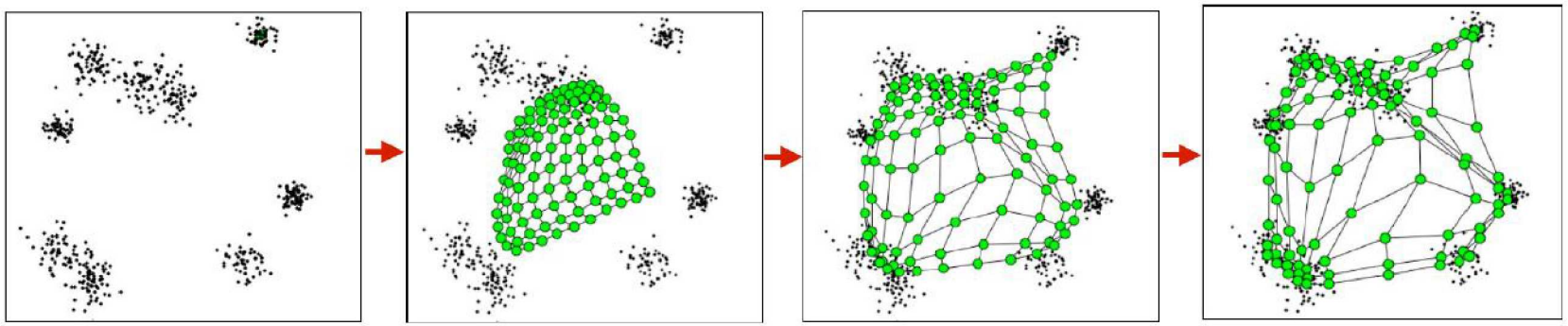
Basic model : Kohonen Self-Organizing Map's (SOM)



$$R_{SOM}(\chi, \mathcal{W}) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} \mathcal{K}_{j,\chi(\mathbf{x}^{(i)})} \|\mathbf{x}^{(i)} - \mathbf{w}^{(j)}\|^2$$

Cost function

Basic model : example



Weighting SOM

Weighting distances

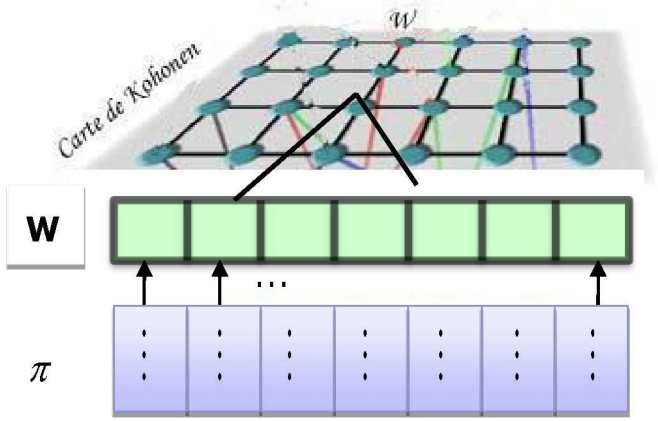
$$R_{lwd-SOM}(\chi, \mathcal{W}, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \left(\pi_j\right)^{\beta} \|\mathbf{x}_i - \mathbf{w}_j\|^2$$

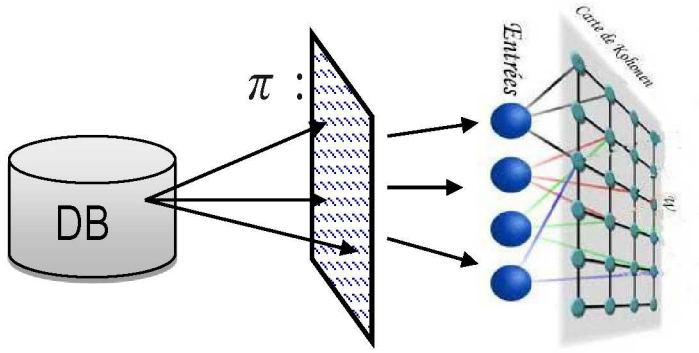
Weighting observations

$$R_{lwo-SOM}(\chi, \mathcal{W}, \Pi) = \sum_{i=1}^{|N|} \sum_{j=1}^{|\mathcal{W}|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \|\pi_j \mathbf{x}_i - \mathbf{w}_j\|^2$$

Gradient descent

Batch version





Weighted local distance - *lwd*-SOM

Objective function :

$$R_{lwd-SOM}(\chi, \mathcal{W}, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \left(\pi_j\right)^{\beta} \|\mathbf{x}_i - \mathbf{w}_j\|^2$$

Assignment phase :

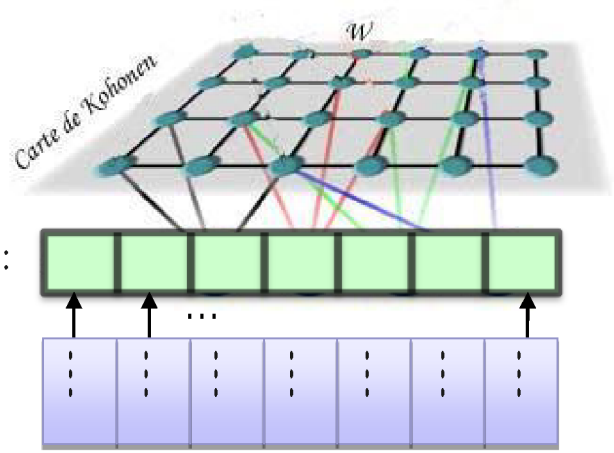
$$\chi(\mathbf{x}_i) = \arg\min_j \left(\left(\pi_j\right)^{\beta} \|\mathbf{x}_i - \mathbf{w}_j\|^2 \right)$$

Quantification phase :

$$w_j(t+1) = w_j(t) + \varepsilon(t) K_{j,\chi(x_i)} (\pi_j)^{\beta} \left(x_i - w_j(t)\right)$$

Weighting phase :

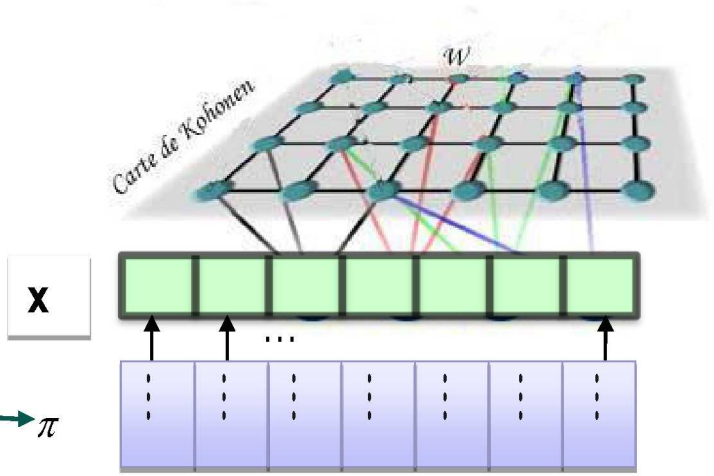
$$\pi_j(t+1) = \pi_j(t) + \varepsilon(t) K_{j,\chi(x_i)} \beta(\pi_j(t))^{\beta-1} \left(x_i - w_j(t)\right)$$



Local observations weighting - *lwo*-SOM

The objective function :

$$R_{lwo-SOM}(\chi, \mathcal{W}, \Pi) = \sum_{i=1}^{|N|} \sum_{j=1}^{|\mathcal{W}|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \|\pi_j \mathbf{x}_i - \mathbf{w}_j\|^2$$



Affectation phase :

$$\chi(\mathbf{x}_i) = \arg\min_j \left(\|\pi_j \mathbf{x}_i - \mathbf{w}_j\|^2 \right)$$

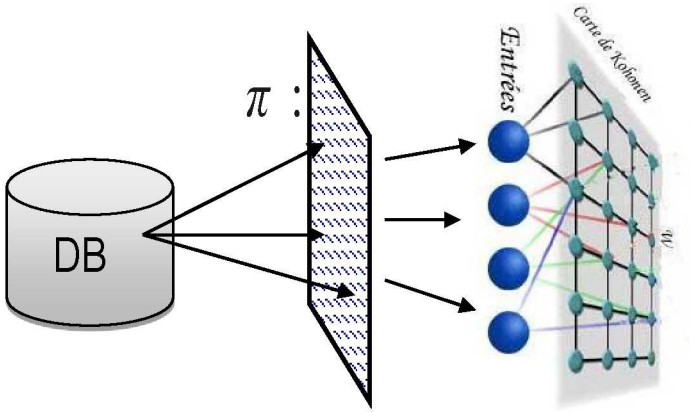
Quantification phase :

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \epsilon(t) \mathcal{K}_{j,\chi(\mathbf{x}_i)} \left(\pi_j \mathbf{x}_i - \mathbf{w}_j(t) \right)$$

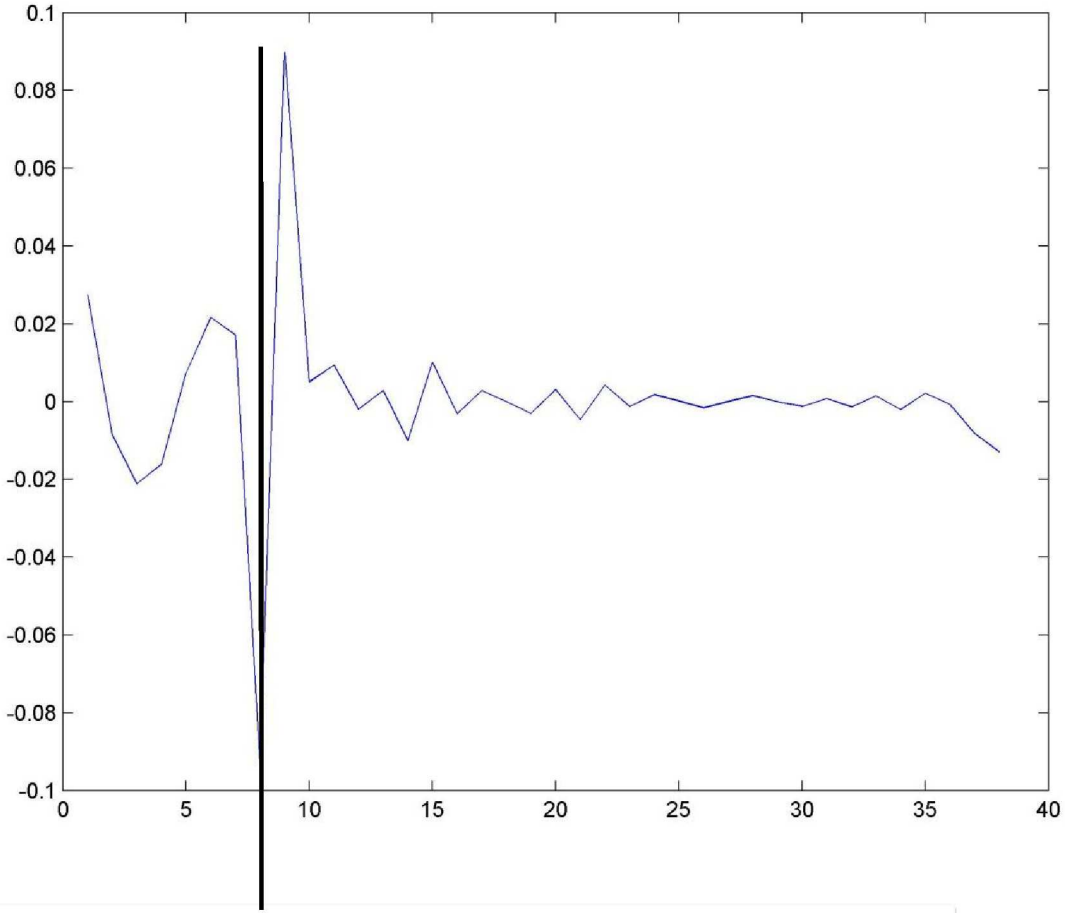
Weighting phase :

$$\pi_j(t+1) = \pi_j(t) + \epsilon(t) \mathcal{K}_{j,\chi(\mathbf{x}_i)} \mathbf{x}_i \left(\pi_j(t) \mathbf{x}_i - \mathbf{w}_j(t) \right)$$

Data filtering process



Features selection using the weights



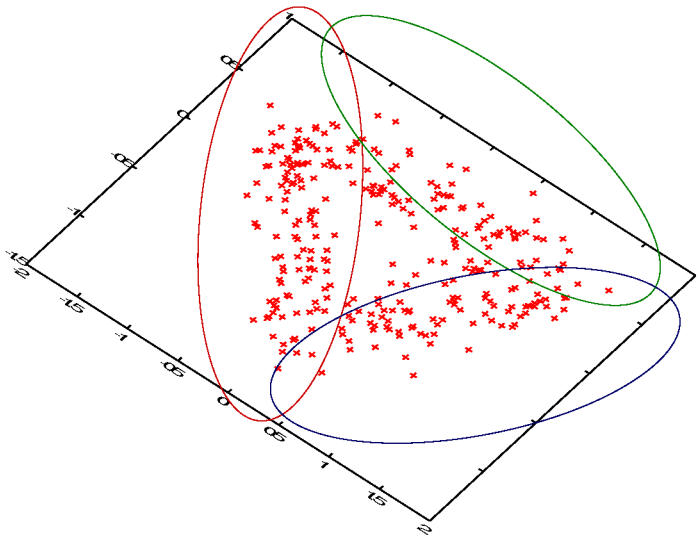
Statistical Test – Scree Test (Cattell, 1966)

Experimentations :data sets

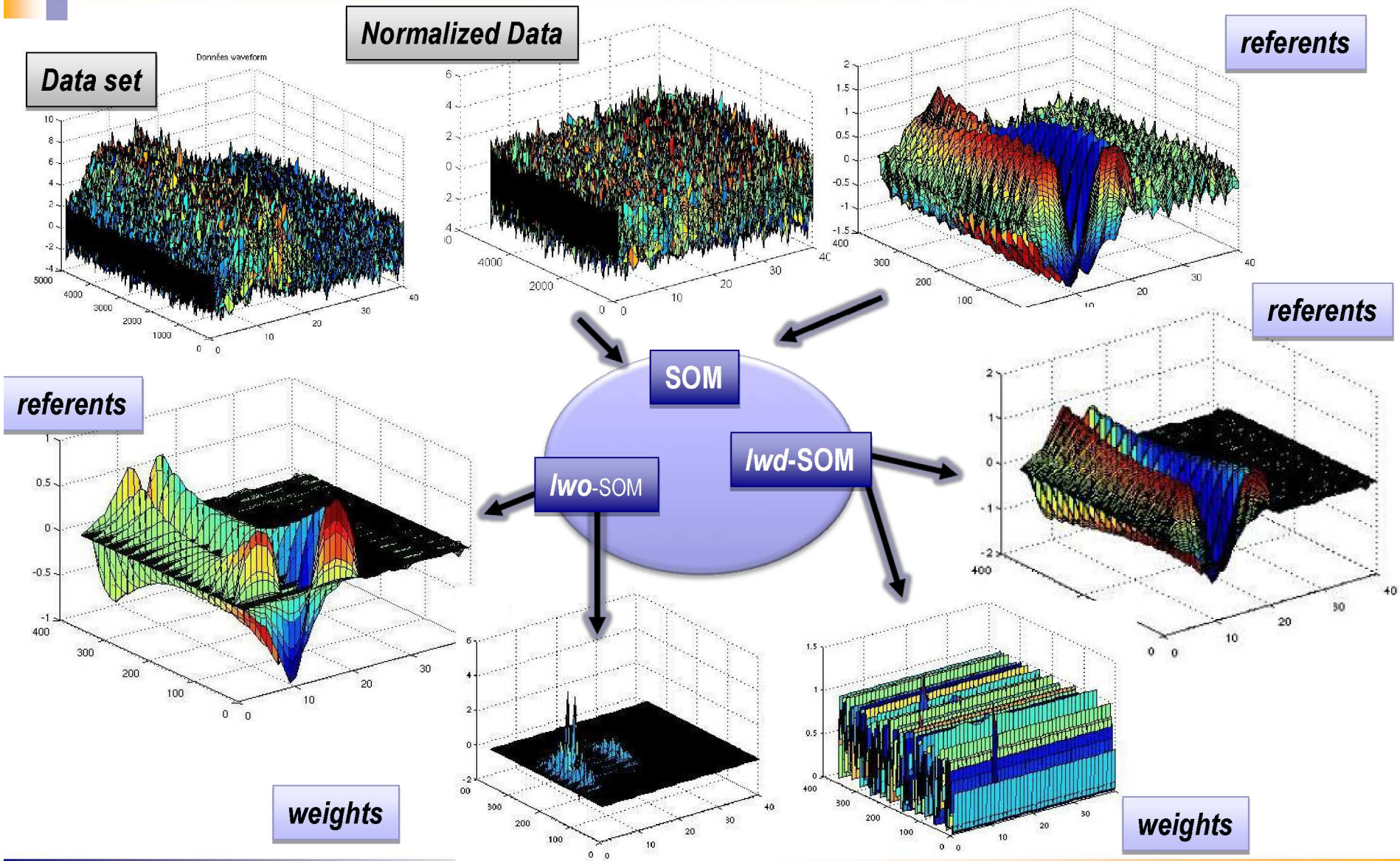
- Academical DB:
  - Iris (150 x 4)
  - Waveform (5000 x 40)
  - Wdbc (569 x 32)
  - Spambase (4601 x 57)
  - Madelon (2000 x 500)
  - Isolet (6238 x 617)

- DB infom@gic:
  - Assurance (800.000 x 32)
  - Wikipedia (19.000 x 6400) x 2

Example : waveform



Experimentations : Waveform



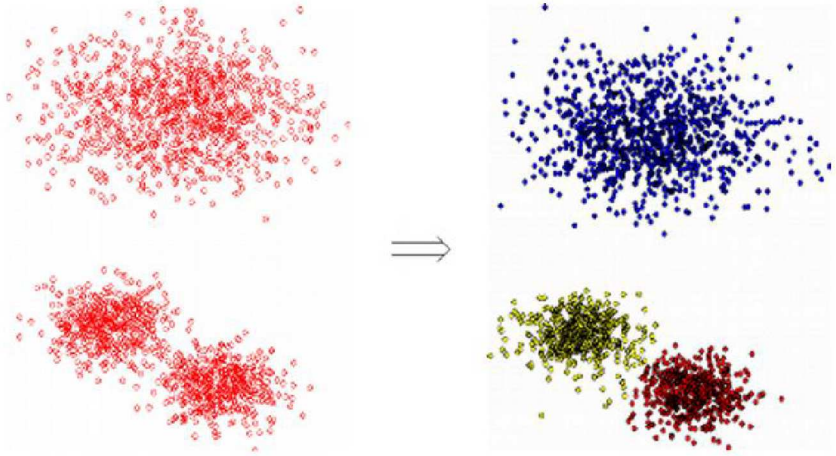
Evaluation of the quality using Purity Index

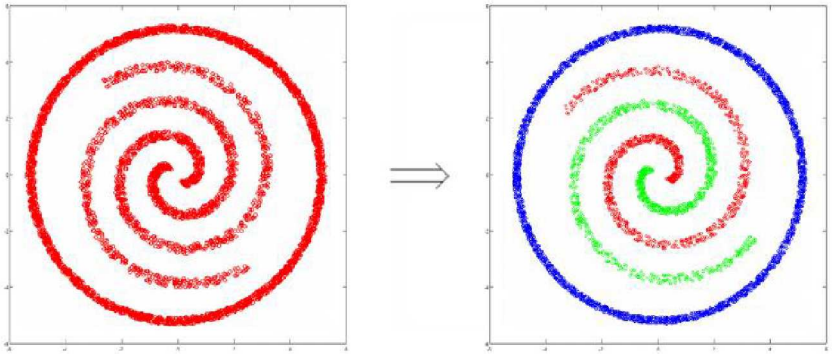
Db.	b/a sel/cl	method SOM	*lwo*-SOM	*lwd*-SOM
Isolet	b.	0.7786±0.05	0.7975±0.04	0.7792±0.047
	Sel f.	0.7409±0.052	0.7863±0.043	0.7608±0.041
	Sel cl.	0.6786±0.061	0.7821±0.047	0.7796±0.048
wdbc	b.	0.8941±0.042	0.9203±0.037	0.9052±0.041
	Sel f.	0.8923±0.047	0.9152±0.04	0.9023±0.043
	Sel cl.	0.891±0.046	0.9145±0.041	0.9014±0.042
Spam	b.	0.8958±0.041	0.8669±0.041	0.8568±0.043
	Sel f.	0.8579±0.039	0.8754±0.04	0.8727±0.043
	Sel cl.	0.6184±0.044	0.8564±0.041	0.8534±0.042
made-lon	b.	0.6541±0.041	0.6803±0.04	0.6752±0.039
	Sel f.	0.6608±0.038	0.7017±0.041	0.6914±0.04
	Sel cl.	0.6524±0.052	0.7163±0.042	0.7089±0.047

Part II
Reducing the number of rows

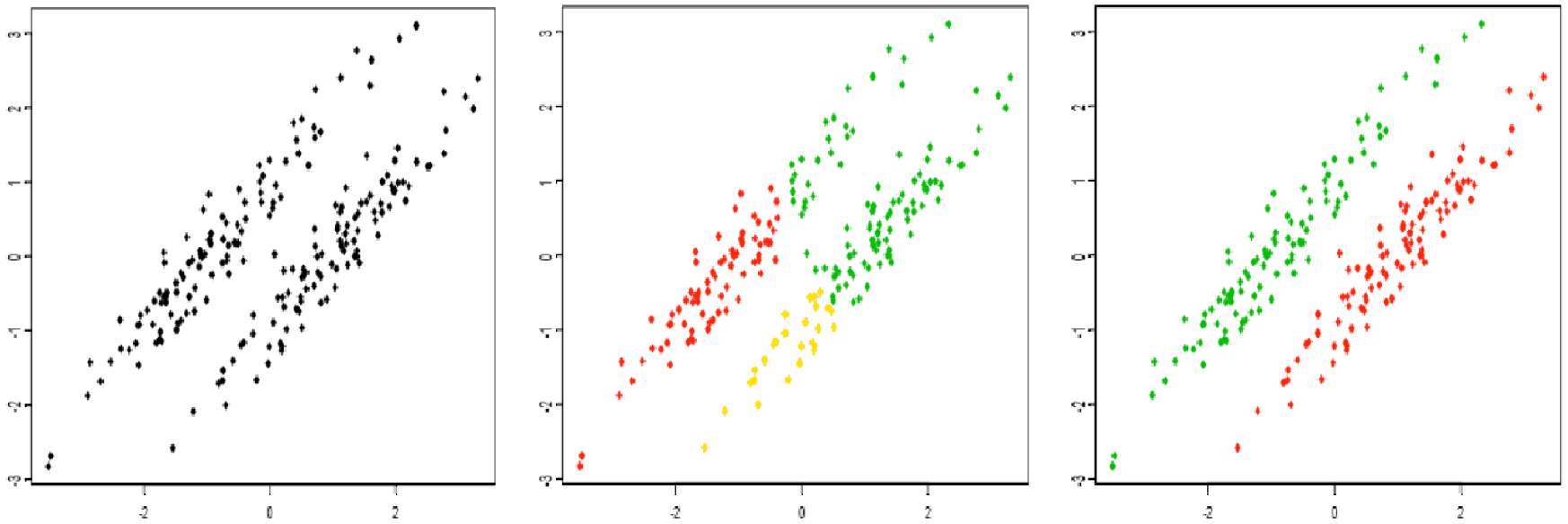
Topological clustering

Clustering





Main Problem



Choosing the right number of cluster is a difficult problem

Model selection

- **Internal and External Quality Indexes**
- **Penalized Likelihood Methods:**
  - Akaike's Information Criterion (AIC)
  - Bayesian information criterion (BIC)
- **Stability-based Methods:**
  - The best model is defined as the most stable clustering under perturbation of data, parameters, initializations, etc . . .
- **In practice** : Try $K : K_{min} \ldots K_{max}$ and choose $K^* = argmin\ R_k(w)$

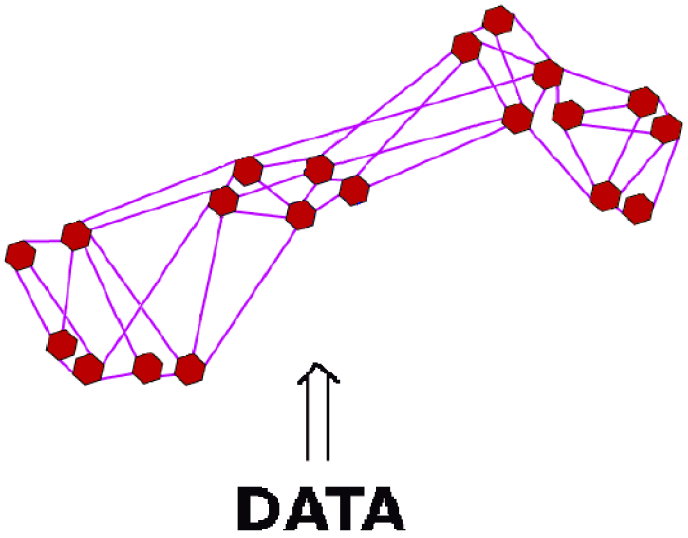
Our approach

Automatic detection of clusters boundaries

1. Clustering using neighborhood
   - Algorithm 1 : Simultaneous 2 Levels - Self Organizing Map
2. Clustering using density
   - Algorithm 2 : Density-based Simultaneous 2 Levels - Self Organizing Map

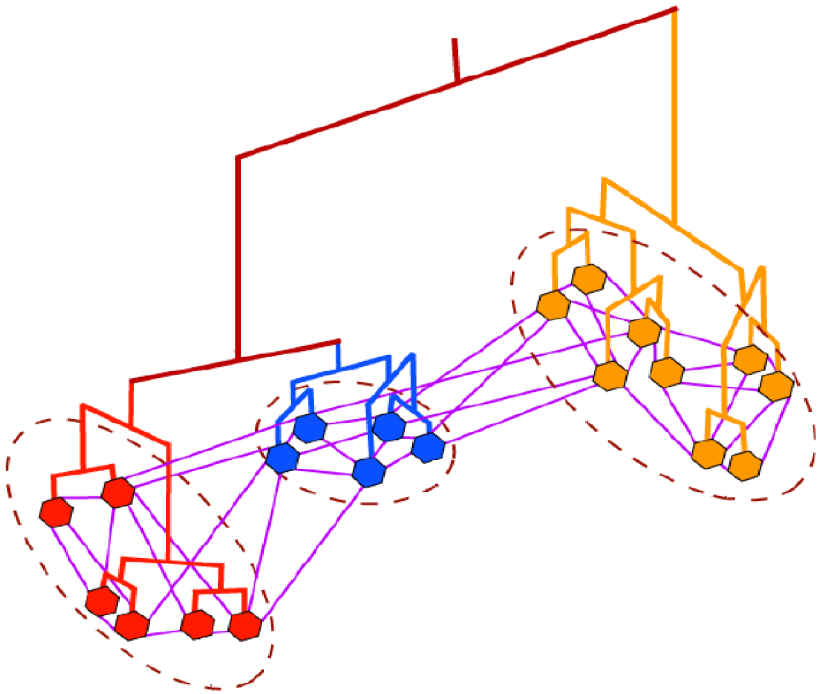
The number of cluster is automatically detected

Two-level clustering



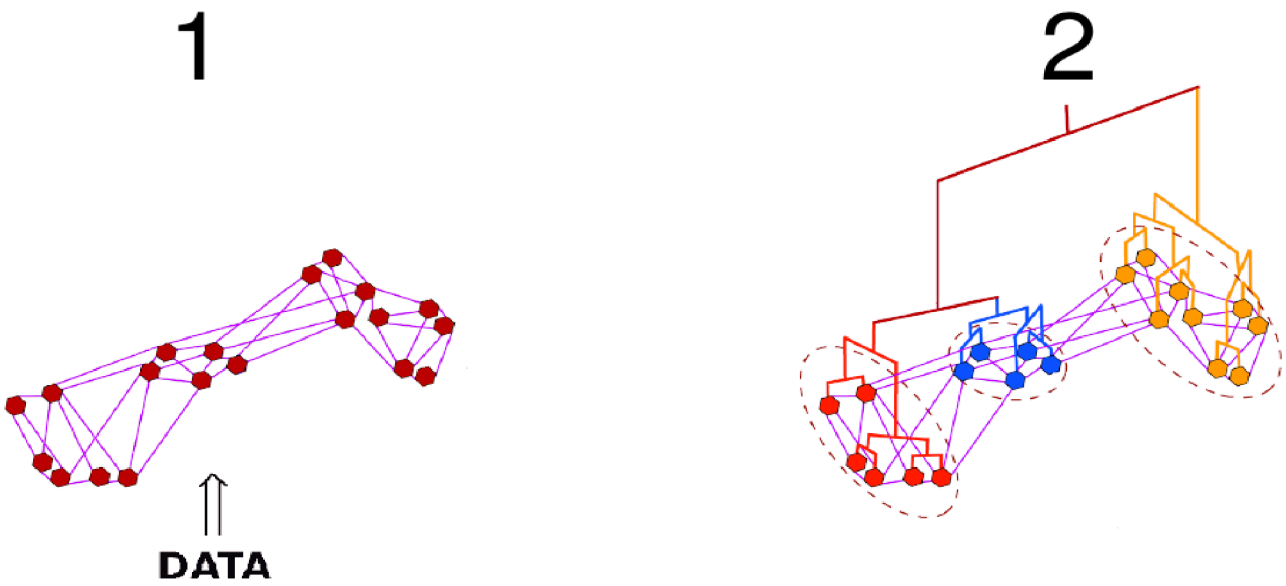
First level : vector quantization (SOM)

Two-level clustering



Second level : clustering (HAC, Kmeans,...)

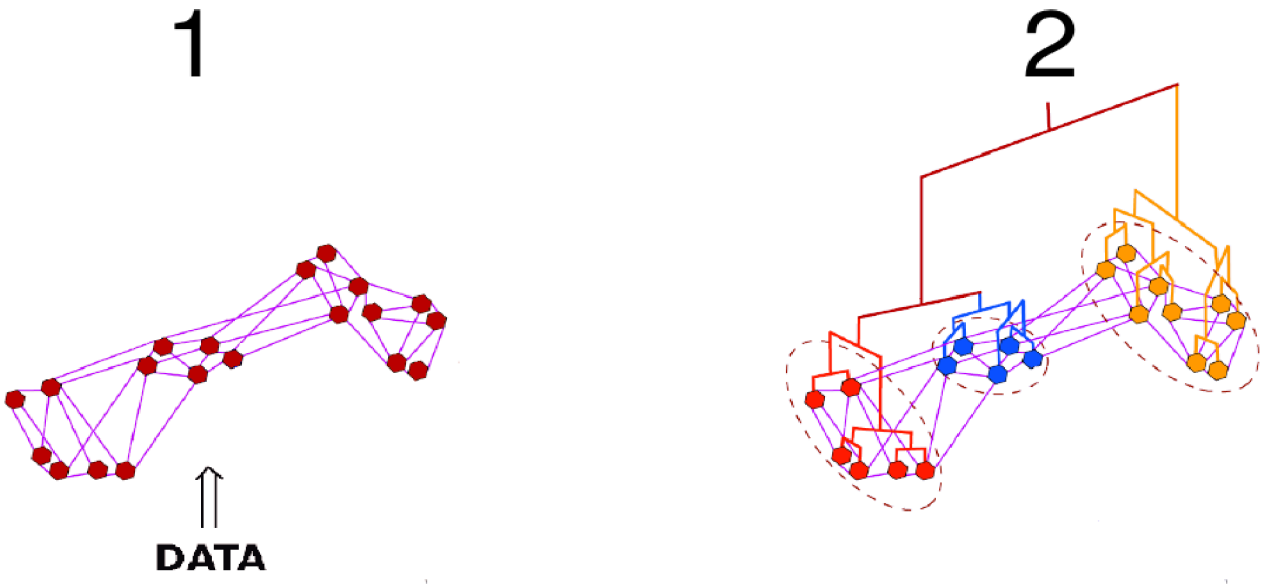
Two-level clustering



Advantages

- Increase speed
- Suitable for the analysis of big databases in high dimensions

Two-level clustering

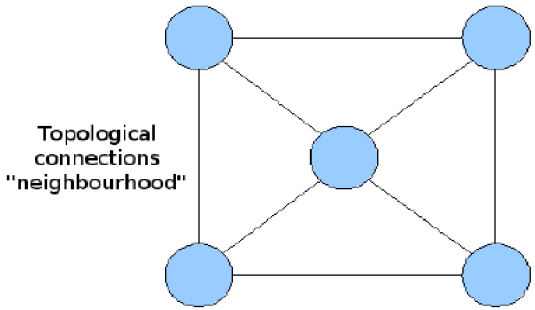


Inconvenient

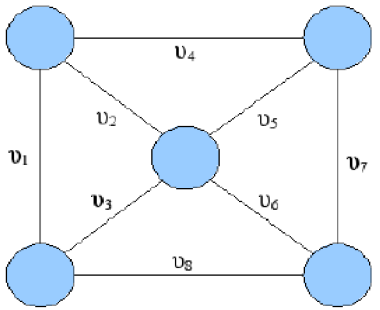
- Lost of information during the first level
- The detection of the number of cluster is still difficult

Learning neighborhood

SOM
Self-Organizing Map



S2L-SOM
Simultaneous 2 Levels - SOM



$\upsilon_1,\ldots,\upsilon_8$ are estimated during learning

Cost function:

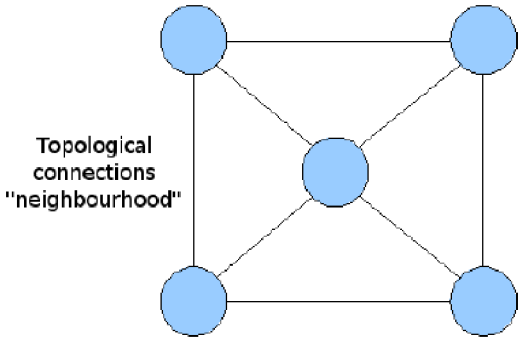
- Cost of the prototypes' estimation.
- Cost of the neighborhood values' estimation:

$$\widetilde{R}(v) = \sum_{k=1}^{N} \sum_{i,j=1}^{M} \left[v_{ij} - \mathbb{1}_{\{w_i, w_j \text{ bmus of } x^{(k)}\}} \right]^2$$

Learning neighborhood

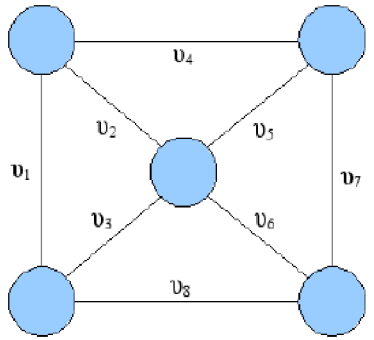
SOM

Self-Organizing Map



S2L-SOM

Simultaneous 2 Levels - SOM

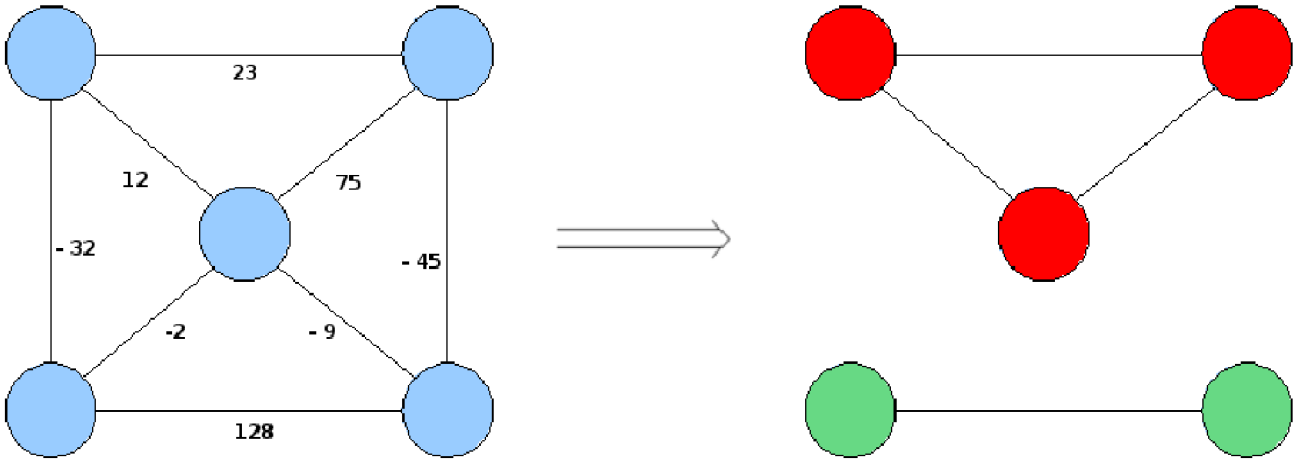


$\upsilon_1,\ldots,\upsilon_8$ are estimated during learning

Learning values

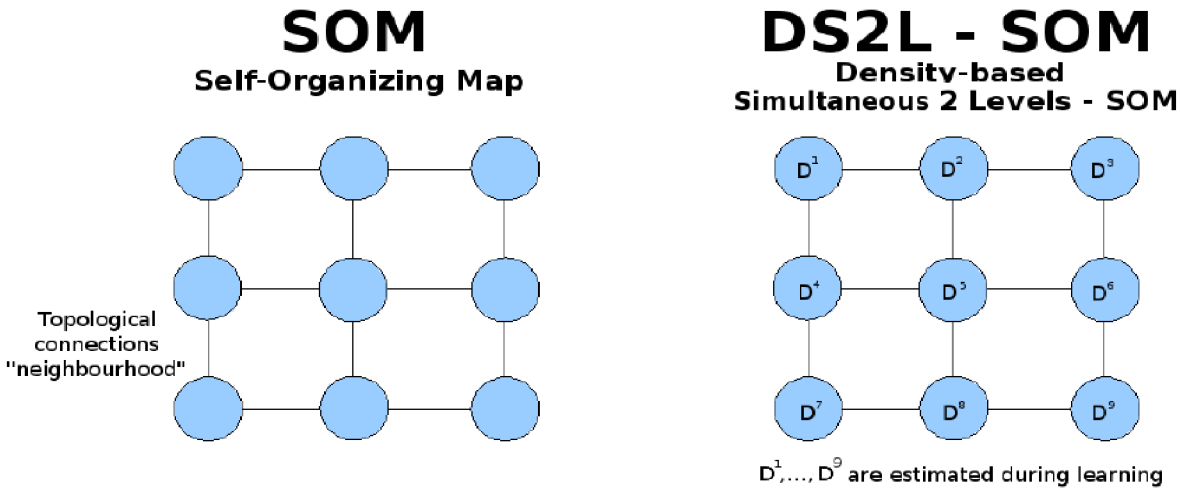
For each data, we propose to increase during the learning the value of the connection between the two closest prototypes and to decrease the values of the other connections from the best prototype.

Clustering from neighborhood



- At the end of the clustering process, we can detect clusters as a sets of prototypes which are linked together by neighborhood connections with positive values.
- Thus, the number of clusters is easy to determine.

Learning density

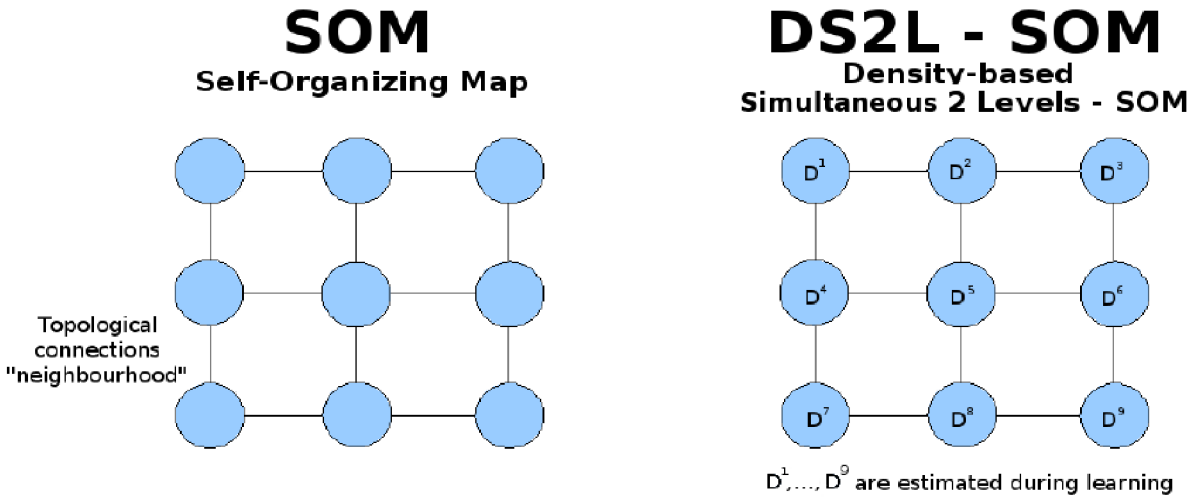


Cost functions:

- Cost of the prototypes' estimation.
- Cost of the density values' estimation :

$$\widetilde{R}(D) = \sum_{k=1}^{N} \sum_{i=1}^{M} \left[D^i - e^{-\frac{\|w_j - x(k)\|^2}{2\sigma^2}} \right]^2$$

Learning density

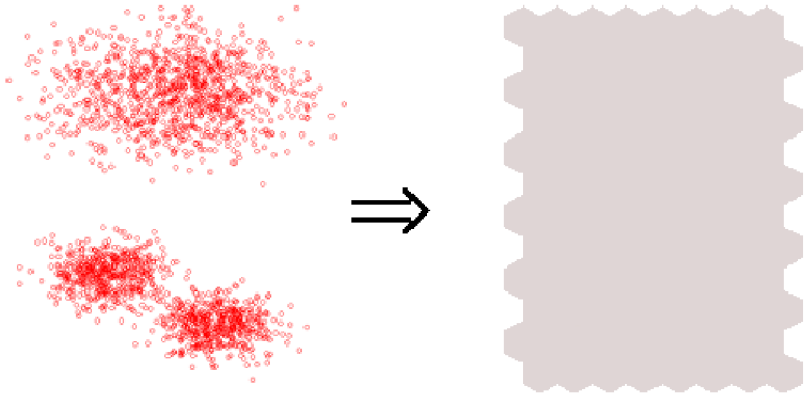


Estimation of density

Compute a function D^i for each prototype i by placing a Gaussian over each data point $x = 1..N$ and summing all Gaussians on i.

$$D^i = \sum_{k=1}^{N} e^{-\frac{\|w_j - x^{(k)}\|^2}{2\sigma^2}}$$

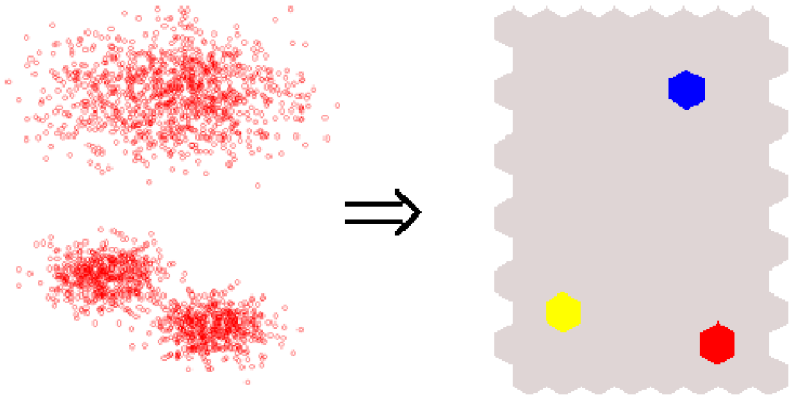
Learning density



Step 1

An enriched SOM is computed from the data

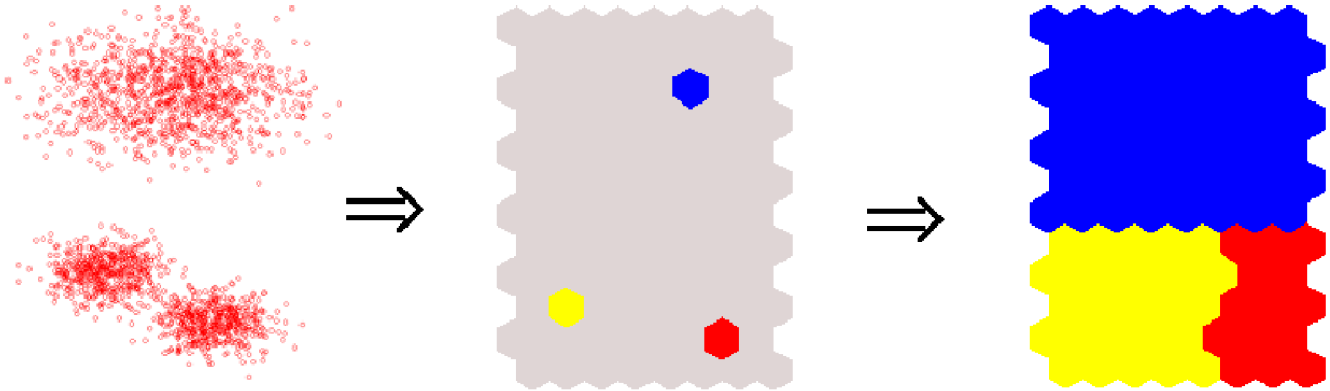
Learning density



Step 2

A local density maximum define a cluster

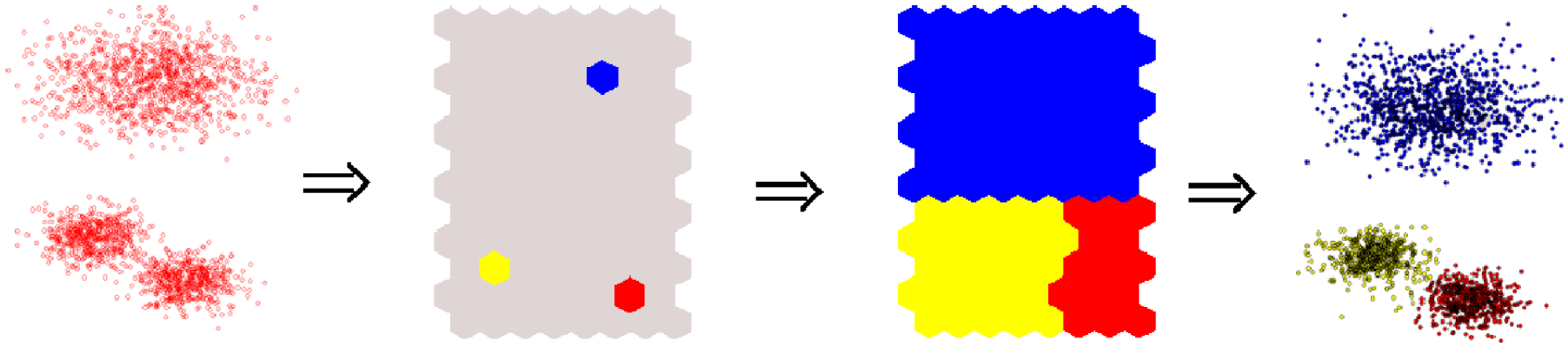
Learning density



Step 3

Each prototype is associated to a density maximum

Learning density

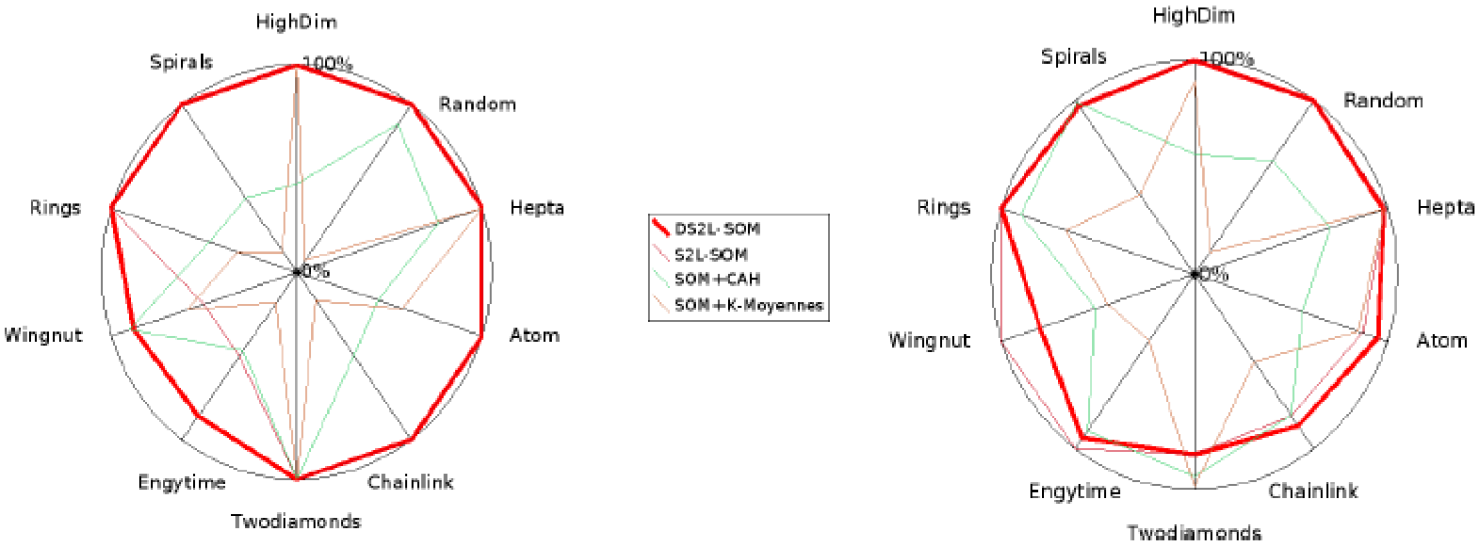


Step 4

Data clustering from prototypes clustering

Evaluation fo the algorithm

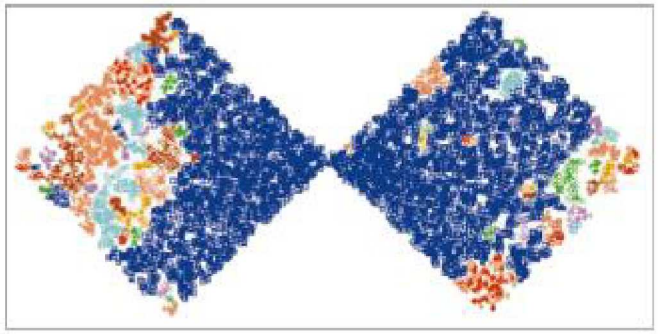
Quality

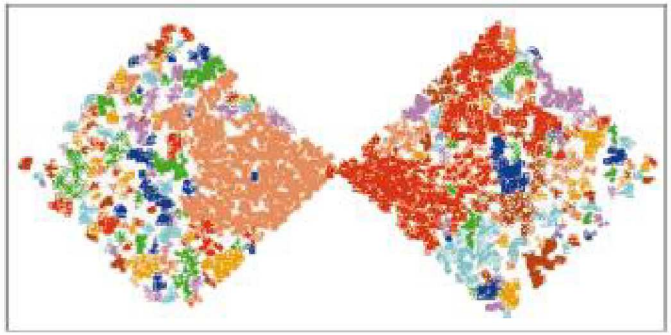


Stability

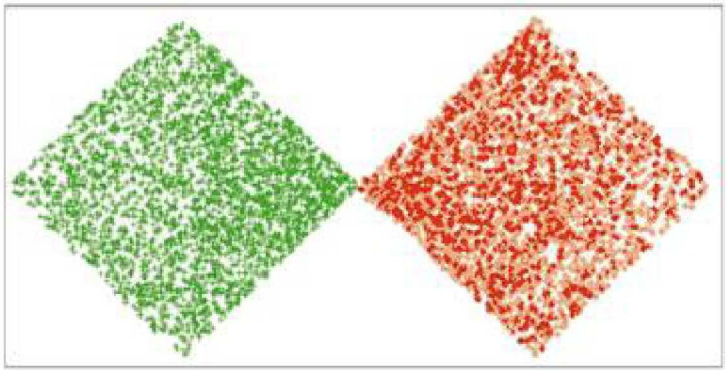
Evaluation of clustering quality and stability according to the database and the method used

Comparisons



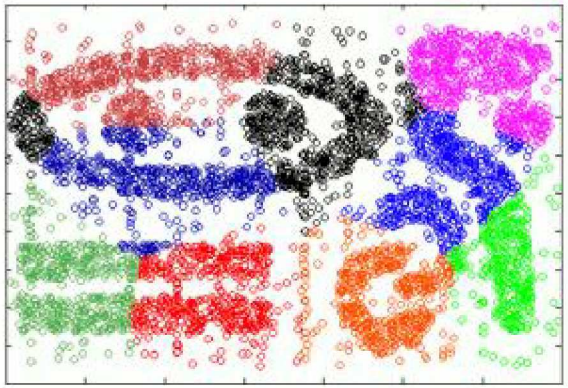


DBSCAN

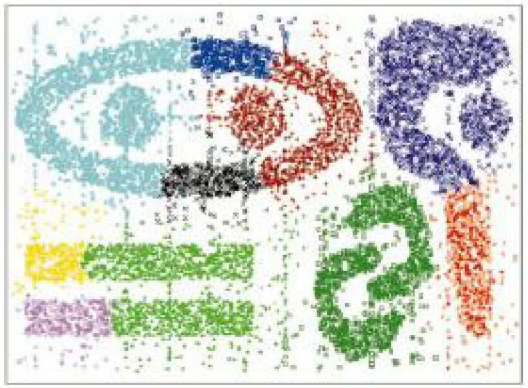


DS2L-SOM

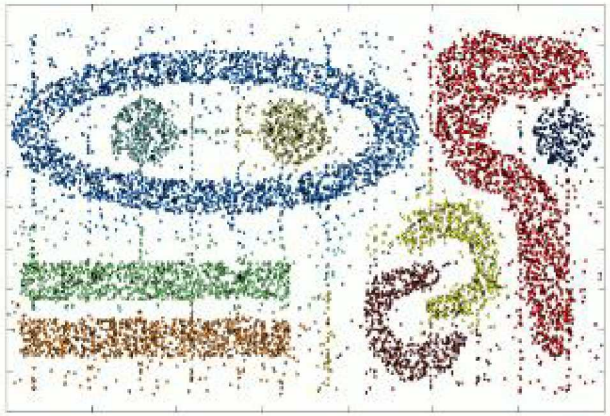
Comparisons



Spectral



CURE



DS2L-SOM

Conclusions

Dimensionality reduction through clustering and features selection

- Reducing de number of columns
  - Variables selection through weighting distance (lwd-SOM)
  - Variables selection through weighting observations (lwo-SOM)

- Reducing de number of rows
  - A Simultaneous Two-Level Clustering Algorithm for Automatic Model Selection (S2L-SOM)
  - A Local Density-Based Simultaneous Two-Level Algorithm for Topographic Clustering (DS2L-SOM)

Thank you …

Questions ?

younes@lipn.univ-paris13.fr
Part I: grozavu@lipn.univ-paris13.fr
Part II: cabanes@lipn.univ-paris13.fr