Computer Science Laboratory of
Paris 13 University

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

Diversity Analysis in Collaborative Clustering

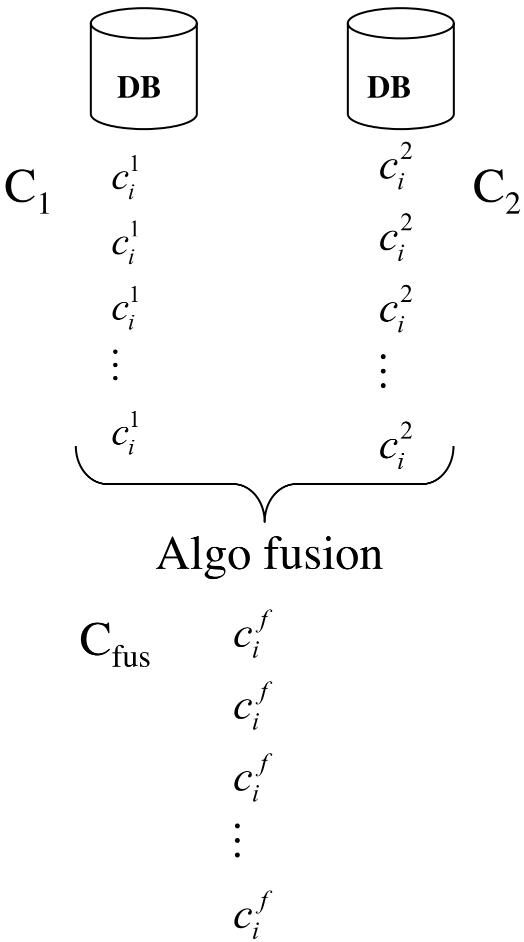
Nistor Grozavu, Guénël Cabanes, Younès Bennani

Plan

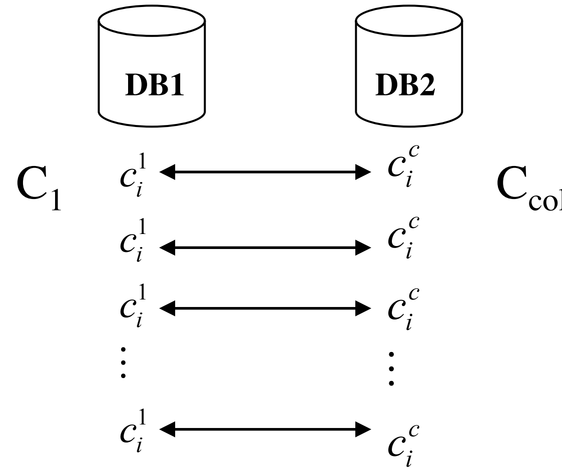
- Introduction
- The problem of the collaborative clustering
  - **Horizontal collaboration**
  - **Vertical collaboration**
- Topological Collaborative Clustering
- Diversity Analysis
  - **The problem**
  - **Proposed solutions**
- Conclusions & Future works

Introduction - Fusion vs Collaboration

The principle of the Fusion



The principle of the Collaboration



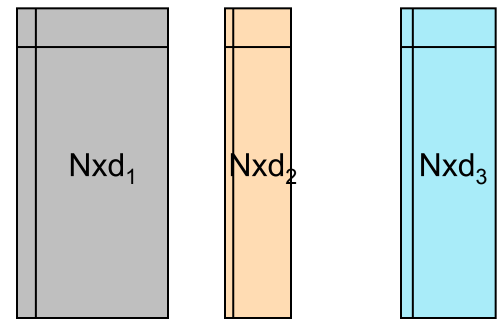
- Collaborate the datasets of different size;
- Use the same clustering method +
  a collaboration step;
- Use this schema for different datasets or for the multi-views datasets;

Collaborative Clustering

Three main types of collaboration :

1. Horizontal

All datasets are described by the same observations but in different spaces
Of description (different variables).

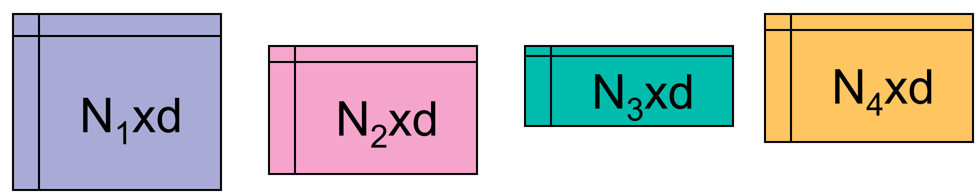


2. Vertical

All the datasets have the same variables (same description space),
but have different observations.

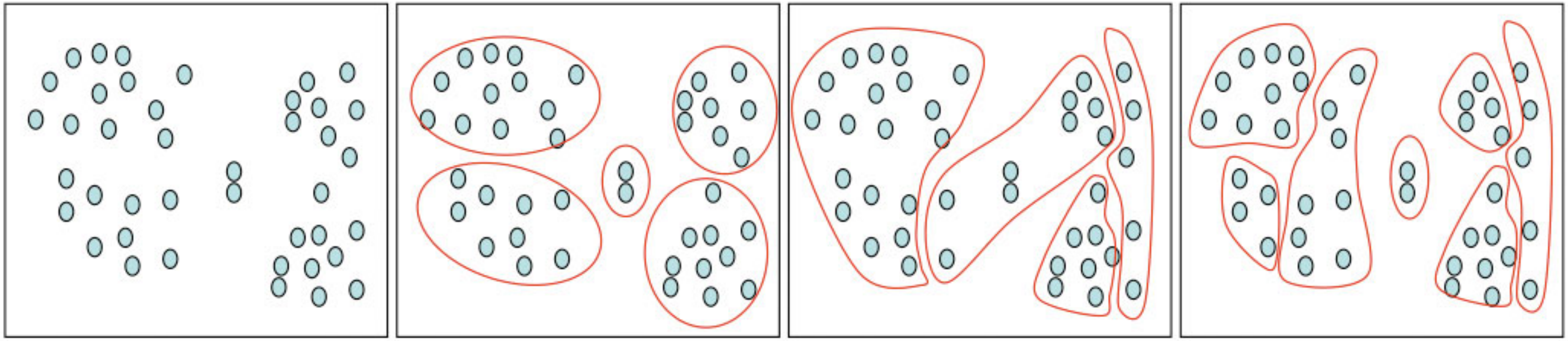
3. Hybrid

Combination between 1 & 2.



The problem

Horizontal collaboration vs Vertical collaboration

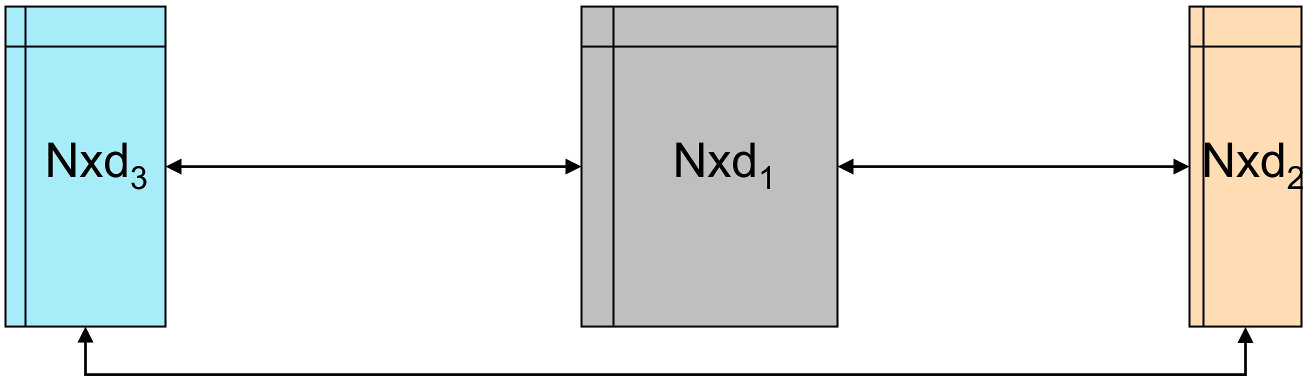


Group of houses

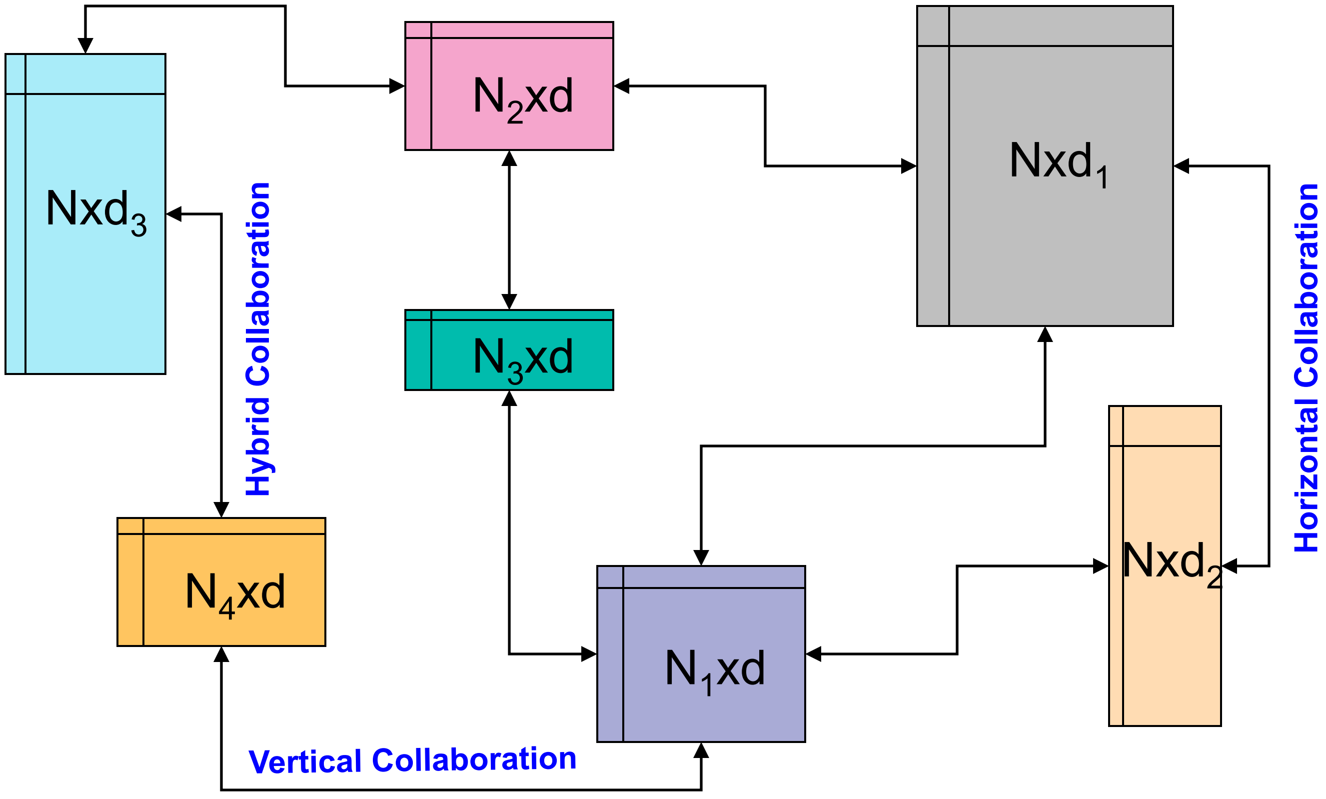
Clustered based on geographic distance

Clustered based on value

Clustered based on size and value



The problem



- **How to improve the local clustering derived out of a set of distant clustering results without sharing the initial data ?**

The problem

- The collaborative clustering is an emerging problem
- Some works (fusion & collaboration) :
  - Pedrycz & Rai 2008 (Collaboration);
  - Costa da Silva & Klusch, 2006 (Collaboration);
  - Wemmert & al., 2007 (Collaborative and Fusion);
  - Cleuziou et al., 2009 (Horizontal Collaboration);
  - Forestier et al., 2009 (Fusion/Collaboration);
  - Grozavu et al., 2009 (Fusion, Collaboration);
  - Strehl & Ghosh, 2002 (Fusion).

- Collaborative Topological Learning uses the principle of the Collaborative Fuzzy c-means (Pedrycz & Rai, 2008)

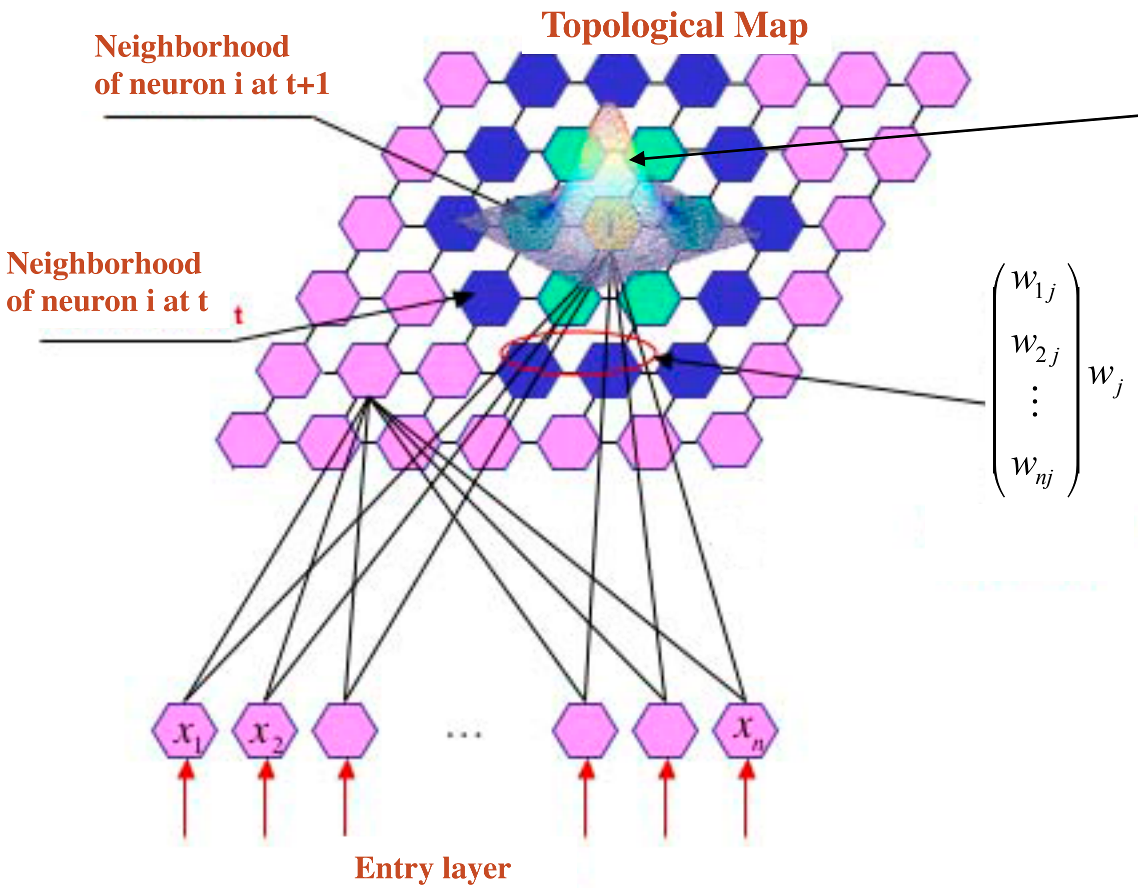
  + self-organization

  + the neighborhood between clusters using SOM (Self Organizing Maps)

Topological Collaborative Clustering

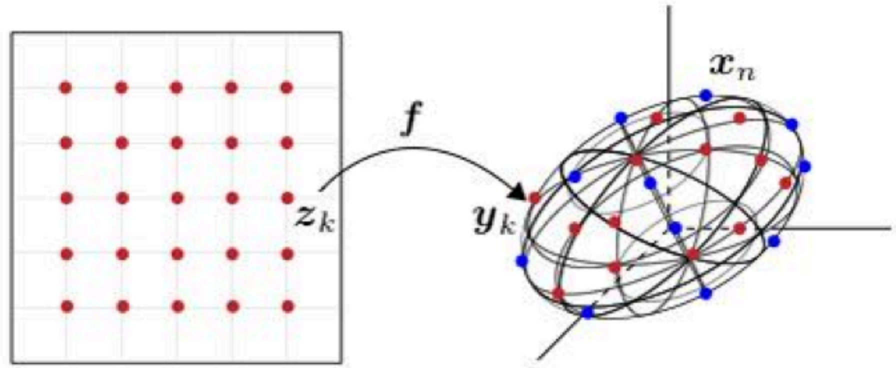
Base model : Kohonen Self-Organizing Map's (SOM)

$$\Re_{SOM}(\chi, W) = \sum_{i=1}^{N} \sum_{j=1}^{|W|} K_{j,\chi(x_i)} \left\| x_i - w_j \right\|^2$$



Probabilistic Clustering

Generative Topographic Mapping [Bishop 95]



$$\boxed{y = y(z, W) = W\Phi(z)}$$

$$\boxed{p(x_n|z, W, \beta) = \mathcal{N}(y(z, W), \beta)}$$

$$\mathcal{L}(W, \beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(x_n|z_i, W, \beta) \right\} \Longrightarrow \boxed{\textbf{EM} \text{ Algorithm}}$$

E & M steps

E step - Computing posterior probabilites

$$
\begin{aligned}
r_{in} &= p(z_i|x_n, W_{old}, \beta_{old}) \\
&= \frac{p(x_n|z_i, W_{old}, \beta_{old})}{\sum_{i'=1}^{K} p(x_n|z_i', W_{old}, \beta_{old})}
\end{aligned}
$$

M step - Updating parameters

$$\mathbb{E}[\mathcal{L}_{comp}(W, \beta)] = \sum_{n=1}^{N} \sum_{i=1}^{K} r_{in} \ln\{p(x_n|z_i, W, \beta)\}$$

$$\Phi^T G \Phi W_{new}^T = \Phi^T R X$$

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{i=1}^{K} r_{in} \|x_n - W^{new} \phi(z_i)\|^2$$

Topological Collaborative Clustering

Collaborative Clustering : **local step + collaboration step**

$$R_H^{[ii]}(W) = R_{Quantiz}(W) + R_{Collab}(W)$$

- Prototype based Clustering

$$R_{Quantiz}(W) = \sum_{jj=1, jj \neq ii}^{P} \alpha_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} \mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

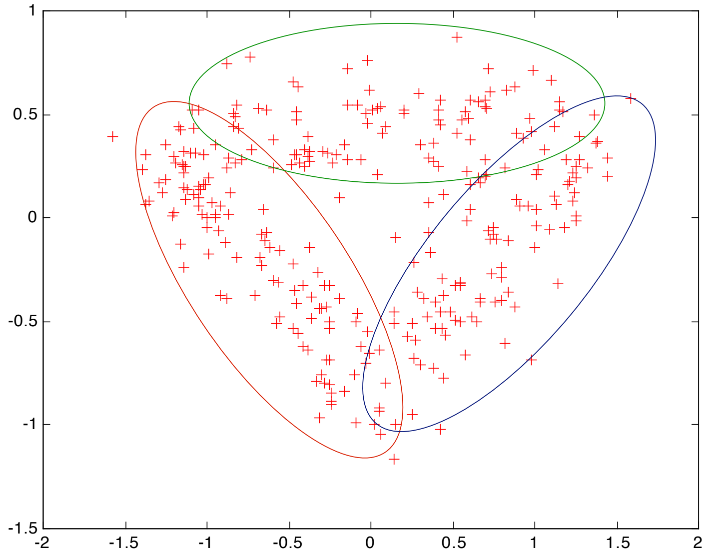
$$R_{Collab}(W) = \sum_{jj=1, jj \neq ii}^{P} \beta_{[ii]}^{[jj]} \sum_{i=1}^{N} \sum_{j=1}^{|w|} \left(\mathcal{K}_{\sigma(j,\chi(x_i))}^{[ii]} - \mathcal{K}_{\sigma(j,\chi(x_i))}^{[jj]}\right)^2 * \|x_i^{[ii]} - w_j^{[ii]}\|^2$$

- Probabilistic Clustering

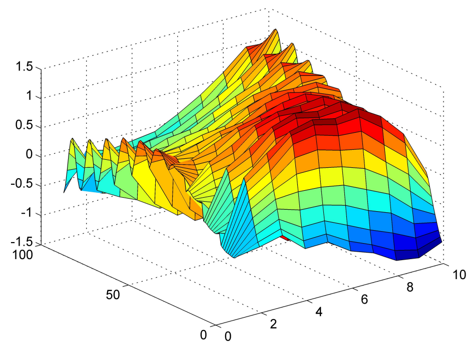
$$\mathcal{L}^{hor}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1, [jj] \neq [ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N} \sum_{i=1}^{K} \frac{\beta^{[ii]}}{2} (r_{in}^{[ii]} - r_{in}^{[jj]})^2 \|x_n - W^{[ii]} \phi^{[ii]}(z_i)\|^2$$

Experimental results (1)

- Waveform dataset
  - 5000 samples
  - 40 variables where 19 variables are Gaussian noisy
  - 3 classes

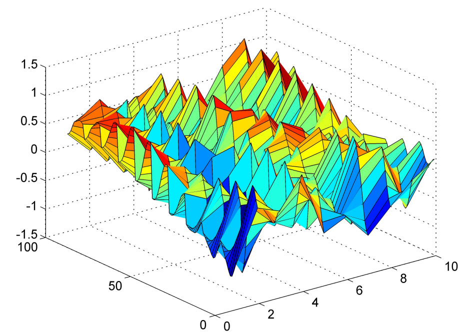


Horizontal Collaboration (waveform)



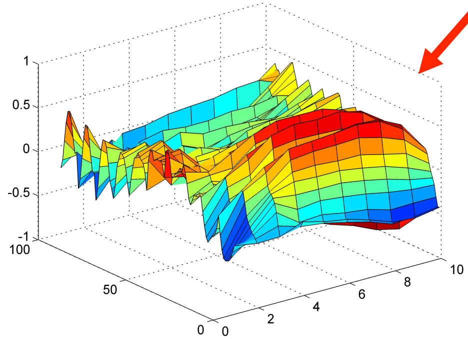
The prototypes of the 1st map obtained from the 1st dataset before the collaboration : SOM1

75.71%



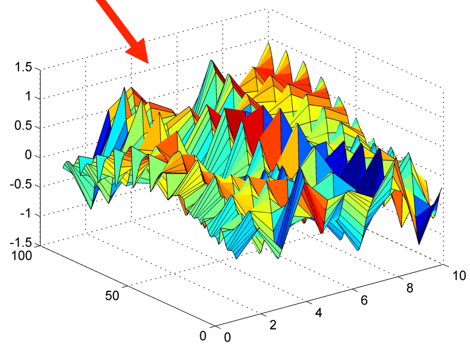
The prototypes of the map from the 3rd dataset before the collaboration : SOM3

47.19%



The prototypes of the map obtained from the 1st dataset after the collaboration with SOM3 : SOM13

62.47%

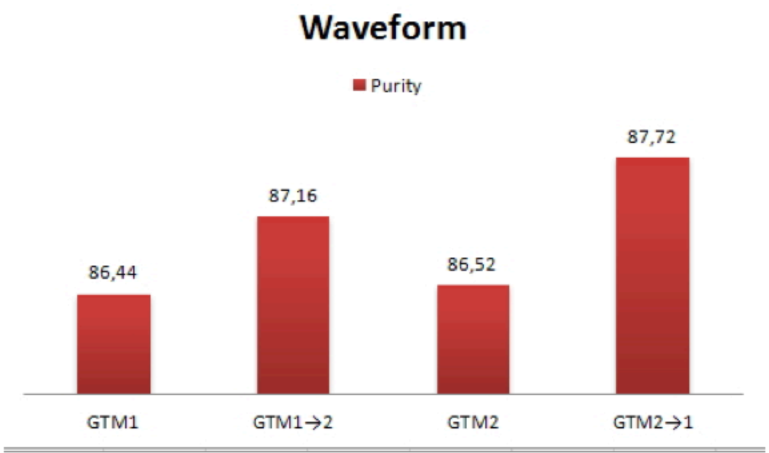


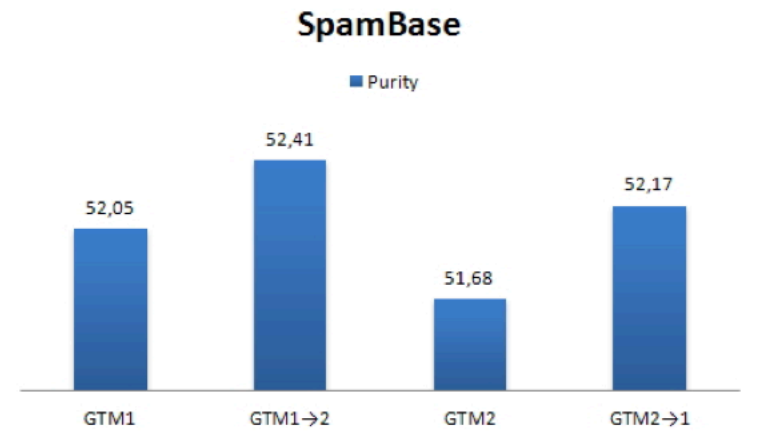
The prototypes of the map obtained from the 3rd dataset after the collaboration with SOM1 : SOM31

54.63%

Experimental results (2)

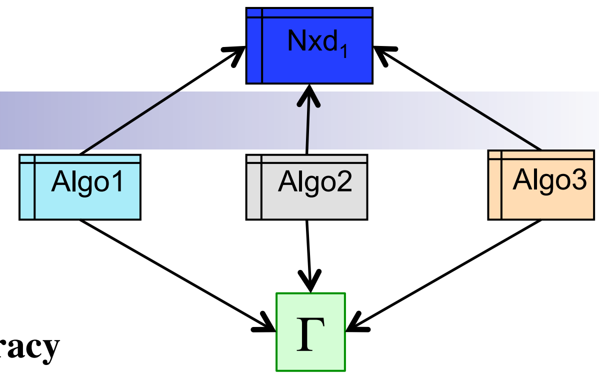
Dataset	Map	Purity
Waveform	GTM_1	86.44
	GTM_2	86.52
	$GTM_{1\to2}$	87.16
	$GTM_{2\to1}$	87.72
Wdbc	GTM_1	96
	GTM_2	96.34
	$GTM_{1\to2}$	96.08
	$GTM_{2\to1}$	96.15
Isolet	GTM_1	87.17
	GTM_2	86.83
	$GTM_{1\to2}$	87.29
	$GTM_{2\to1}$	85.87
SpamBase	GTM_1	52.05
	GTM_2	51.68
	$GTM_{1\to2}$	52.41
	$GTM_{2\to1}$	52.17





Diversity analysis

Diversity : why?

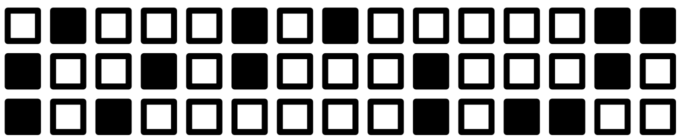


Studied in Consensus clustering

Dataset X containing 15 samples

accuracy

Algo1
Algo2
Algo3

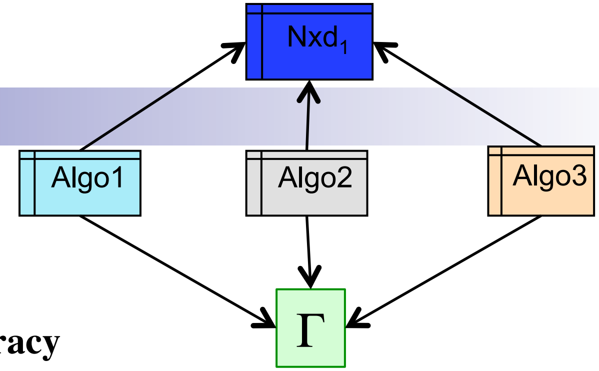


10/15 = **0.667**
10/15 = **0.667**
10/15 = **0.667**

Correct

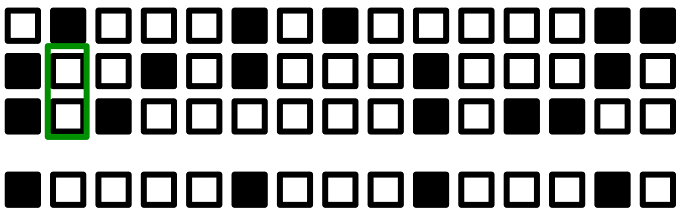
Wrong

Diversity : why?



Studied in Consensus clustering

Dataset X containing 15 samples

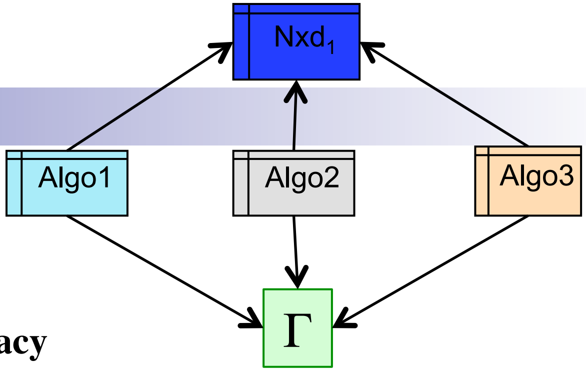


	accuracy
Algo1	10/15 = **0.667**
Algo2	10/15 = **0.667**
Algo3	10/15 = **0.667**
Fusion	11/15 = **0.773**

Correct Wrong

Majority vote rule

Diversity : why?



Studied in Consensus clustering

Dataset X containing 15 samples

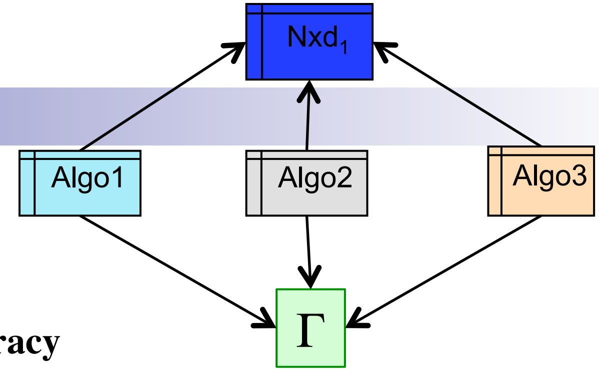
		accuracy
Algo1	□ ■ □ □ □ ■ □ ■ □ □ □ □ □ ■ ■	10/15 = **0.667**
Algo2	■ □ □ ■ □ ■ □ □ □ ■ □ □ □ ■ □	10/15 = **0.667**
Algo3	■ □ ■ □ □ □ □ □ □ ■ □ ■ ■ □ □	10/15 = **0.667**
Fusion	■ □ □ □ □ ■ □ □ □ ■ □ □ □ ■ □	11/15 = **0.773**

□ Correct ■ Wrong

Majority vote rule

Algo1	■ □ ■ ■ ■ ■ □ □ □ □ □ □ □ □ □	10/15 = **0.667**
Algo2	■ □ ■ ■ ■ ■ □ □ □ □ □ □ □ □ □	10/15 = **0.667**
Algo3	■ □ ■ ■ ■ ■ □ □ □ □ □ □ □ □ □	10/15 = **0.667**

Diversity : why?



Studied in Consensus clustering

Dataset X containing 15 samples

□ Correct ■ Wrong

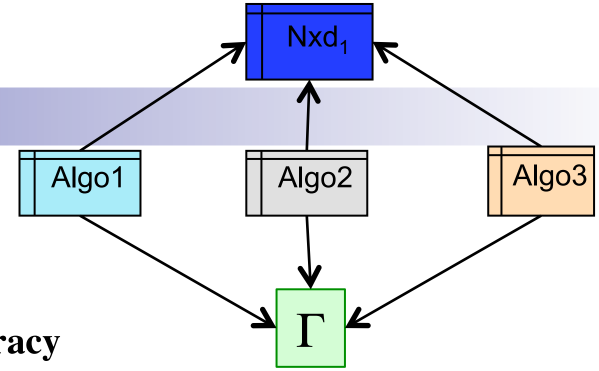
		accuracy
Algo1	□■□□□■□■□□□□□■■	10/15 = **0.667**
Algo2	■□□■□■□□□■□□□■□	10/15 = **0.667**
Algo3	■□■□□□□□□■□■■□□	10/15 = **0.667**
Fusion	■□□□□■□□□■□□□■□	11/15 = **0.773**

Majority vote rule

Algo1	■□■■■■□□□□□□□□□	10/15 = **0.667**
Algo2	■□■■■■□□□□□□□□□	10/15 = **0.667**
Algo3	■□■■■■□□□□□□□□□	10/15 = **0.667**
Fusion	■□■■■■□□□□□□□□□	10/15 = **0.667**

Majority vote rule

Diversity : why?

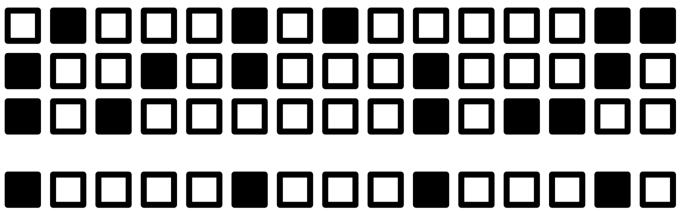


Studied in Consensus clustering

Dataset X containing 15 samples

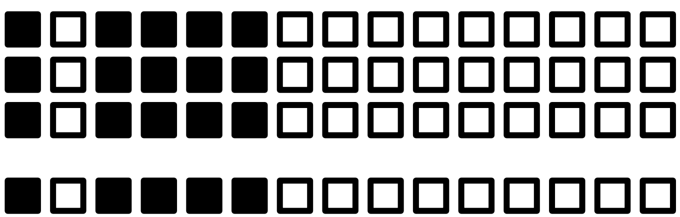
accuracy

Correct   Wrong



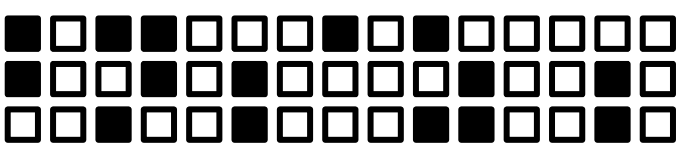
	accuracy
Algo1	10/15 = **0.667**
Algo2	10/15 = **0.667**
Algo3	10/15 = **0.667**
Fusion	11/15 = **0.773**

Majority vote rule



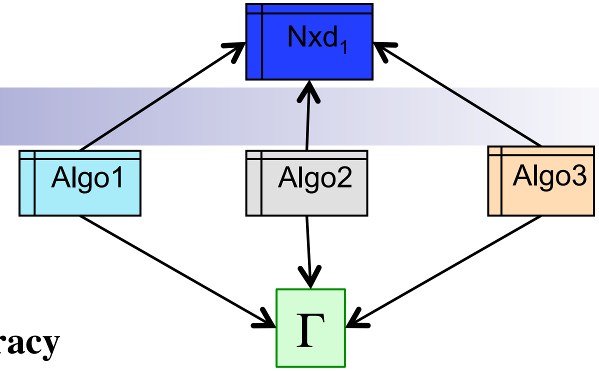
	accuracy
Algo1	10/15 = **0.667**
Algo2	10/15 = **0.667**
Algo3	10/15 = **0.667**
Fusion	10/15 = **0.667**

Majority vote rule



	accuracy
Algo1	10/15 = **0.667**
Algo2	10/15 = **0.667**
Algo3	10/15 = **0.667**

Diversity : why?



Studied in Consensus clustering

Dataset X containing 15 samples

□ Correct ■ Wrong

		accuracy
Algo1	□■□□□■□■□□□□□■■	10/15 = **0.667**
Algo2	■□□■□■□□□■□□□■□	10/15 = **0.667**
Algo3	■□■□□□□□□■□■■□□	10/15 = **0.667**
Fusion	■□□□□■□□□■□□□■□	11/15 = **0.773**

Majority vote rule

Algo1	■□■■■■□□□□□□□□□	10/15 = **0.667**
Algo2	■□■■■■□□□□□□□□□	10/15 = **0.667**
Algo3	■□■■■■□□□□□□□□□	10/15 = **0.667**
Fusion	■□■■■■□□□□□□□□□	10/15 = **0.667**

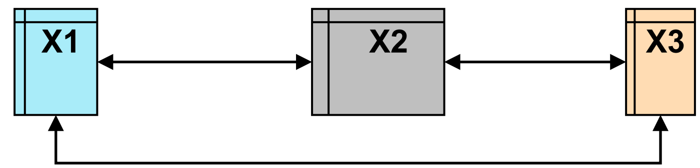
Majority vote rule

Algo1	■□■■□□□■□■□□□□□	10/15 = **0.667**
Algo2	■□□■□■□□□□■□□■□	10/15 = **0.667**
Algo3	□□■□□■□□□■■□□■□	10/15 = **0.667**
Fusion	■□■■□■□□□■■□□■□	8/15 = **0.533**

Majority vote rule

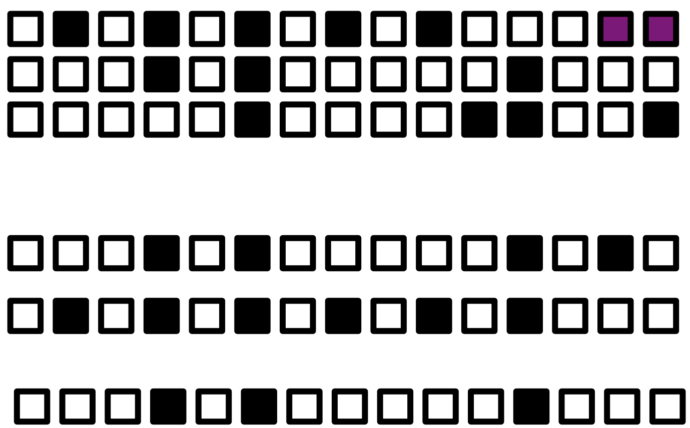
Diversity (2)

Collaborative clustering



Dataset X1 containing 15 samples
Dataset X2 containing 15 samples
Dataset X3 containing 15 samples

□ Correct ■ Wrong

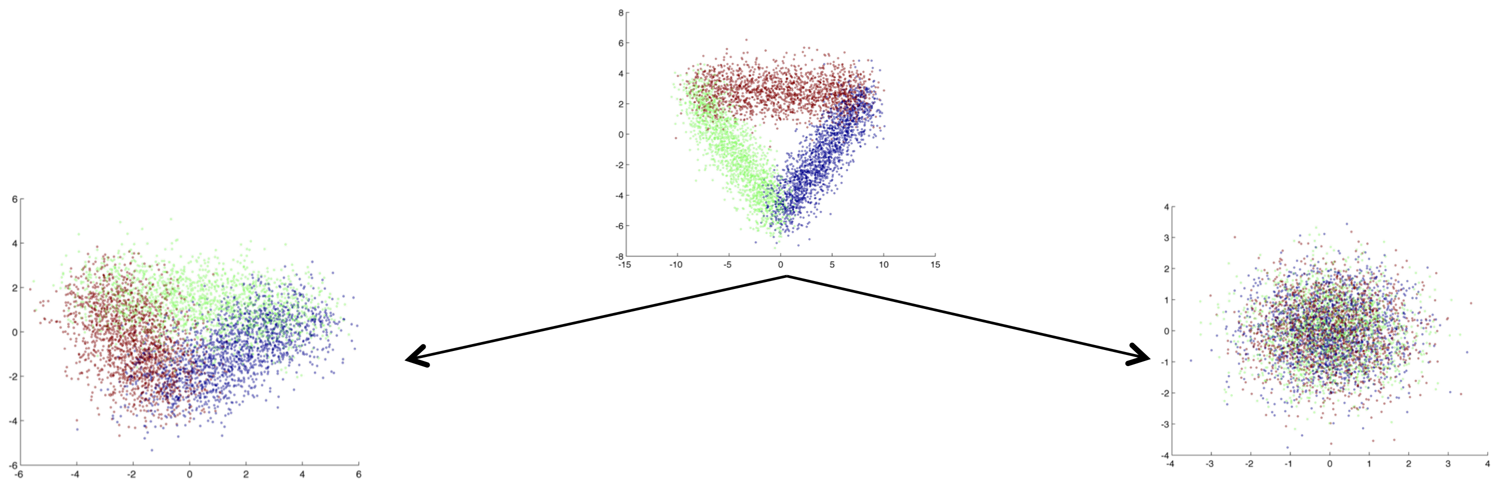


	accuracy
X1	8/15 = **0.533**
X2	12/15 = **0.8**
X3	11/15 = **0.733**
GTM1<-2	11/15 = **0.733**
GTM2<-1	10/15 = **0.6**
GTM3<-2	12/15 = **0.8**

Diversity measures

index	formula
Rand index	$Rand = \frac{a_{00}+a_{11}}{a_{00}+a_{01}+a_{10}+a_{11}}$
Adjusted Rand index	$AdjustedRand = \frac{a_{00}+a_{11}-n_c}{a_{00}+a_{01}+a_{10}+a_{11}-n_c}$
Jaccard index	$Jaccard = \frac{a_{11}}{a_{01}+a_{10}+a_{11}}$
Wallace's coefficient	$W_{P1\rightarrow P2} = \frac{a_{11}}{a_{11}+a_{10}}$ and $W_{P2\rightarrow P1} = \frac{a_{11}}{a_{11}+a_{01}}$
Adjusted Wallace index	$AW_{P1\rightarrow P2} = \frac{W_{P1\rightarrow P2} - Wi_{P1\rightarrow P2}}{1 - Wi_{P1\rightarrow P2}}$
Normalized Mutual Information	$NMI = \frac{-2\sum_{ij} n_{ij} log \frac{n_{ij}N}{n_i n_j}}{\sum_i n_i log \frac{n_i}{N} + \sum_j n_j log \frac{n_j}{N}}$
Variation of Information	$VI = -2\sum_{ij} \frac{n_{ij}}{N} log \frac{n_{ij}N}{n_i n_j} - \sum_i \frac{n_i}{N} log \frac{n_i}{N} - \sum_j \frac{n_j}{N} log \frac{n_j}{N}$

Diversity measures on waveform datasets



(a) dataset db1

(b) dataset db6

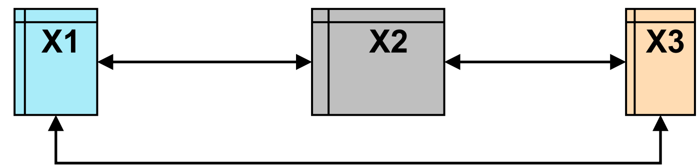
Table 1: Diversity measure on the waveform subsets

Subset	Relevant datasets		Relevant vs Noisy datasets		Noisy datasets	
Diversity index	db2/db3	db3/db4	db2/db8	db4/db9	db7/db8	db9/db10
Rand	0.6707	0.7042	0.5539	0.555	0.543	0.5553
Adjusted Rand	0.2625	0.3356	0.00008	0.0002	0.00002	0.00004
Jaccard	0.3429	0.3869	0.2017	0.2008	0.2	0.2003
Wallace's coefficient	0.5079	0.5578	0.3332	0.3342	0.33	0.3334
Adjusted Wallace	0.5135	0.5581	0.3383	0.3347	0.35	0.3411
Normal Mutual Information	0.262	0.3072	0.0002	0.0006	0.0003	0.0004
Variation of Information	2.334	2.1918	3.1577	3.1631	3.168	3.1664

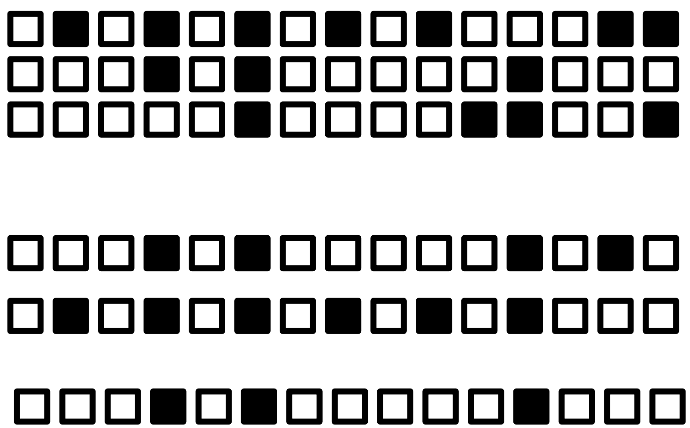
Diversity (2)

Collaborative clustering

Dataset X1 containing 15 samples
Dataset X2 containing 15 samples
Dataset X3 containing 15 samples



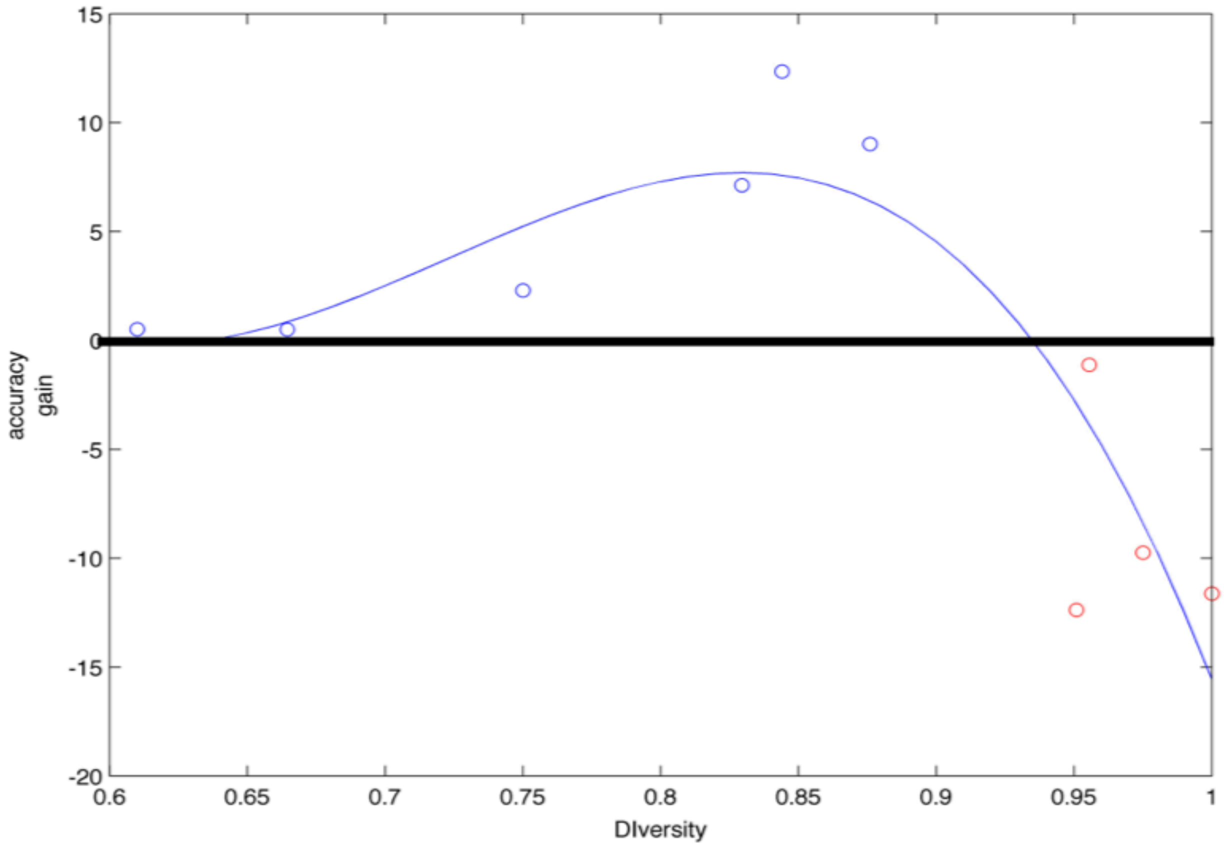
□ Correct ■ Wrong



	accuracy	diversity
X1	8/15 = **0.533**	X1-X2 =**0.956**
X2	12/15 = **0.8**	X2-X3 = **0.678**
X3	11/15 = **0.733**	
GTM1<-2	11/15 = **0.733**	
GTM2<-1	10/15 = **0.6**	
GTM3<-2	12/15 = **0.8**	

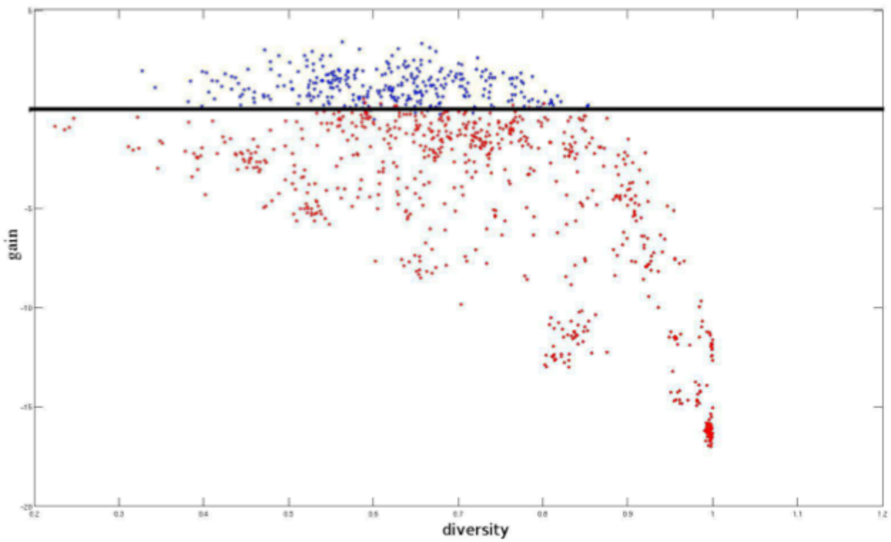
Need to study the local quality.

Results : 10 waveform sub-sets

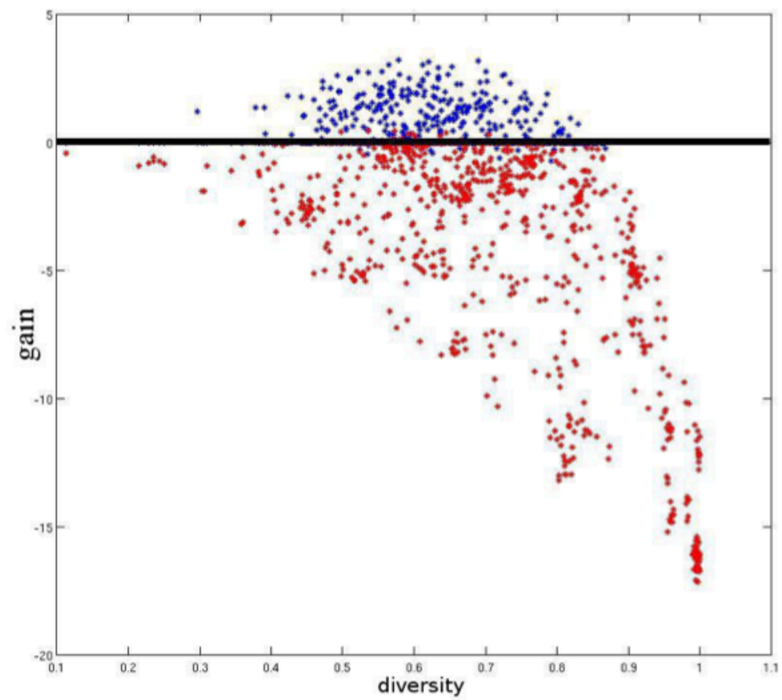


The plot of diversity and the accuracy difference after collaboration

Results : 1-1.000 waveform sub-sets



(a) waveform subset 1

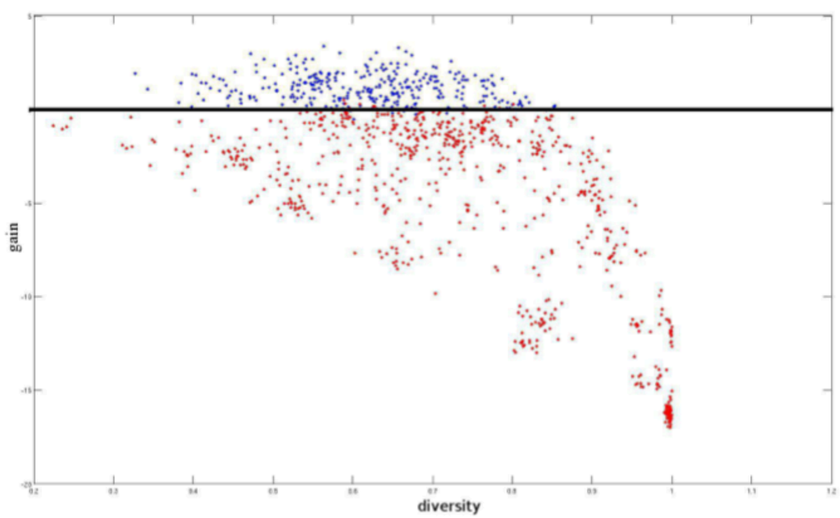


(b) waveform subset 2

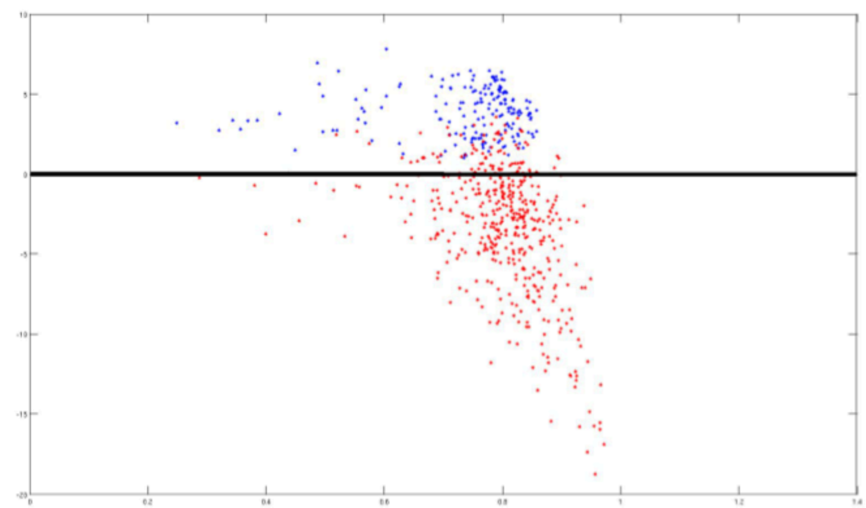
Waveform datasets: Collaboration results between a fixed subset and 1000 randomly subsets (axe X represents the Diversity and axe Y - the Accuracy gain)

Collaboration results (1)

Collaboration results between a fixed subset and 1000 randomly subsets



(a) Waveform dataset

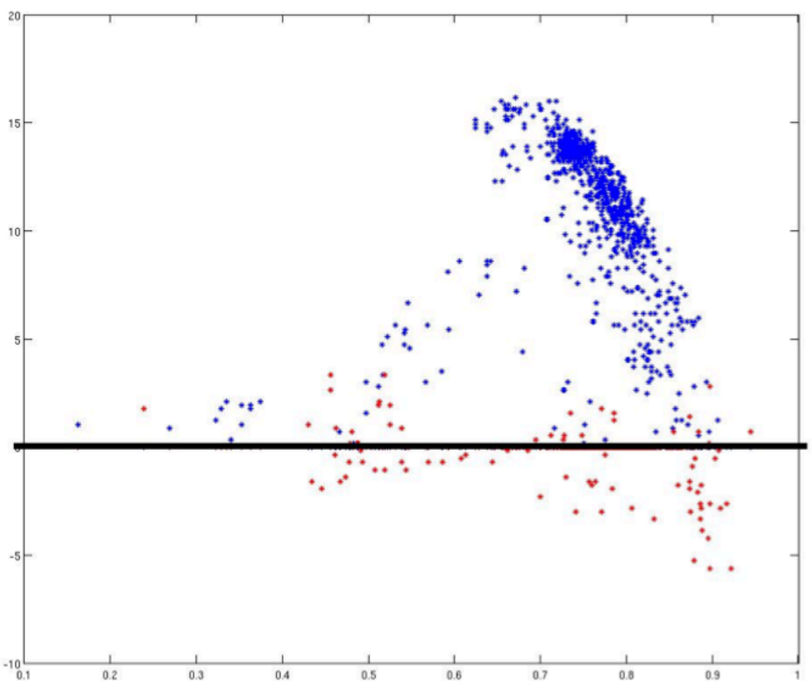


(d) Wdbc dataset

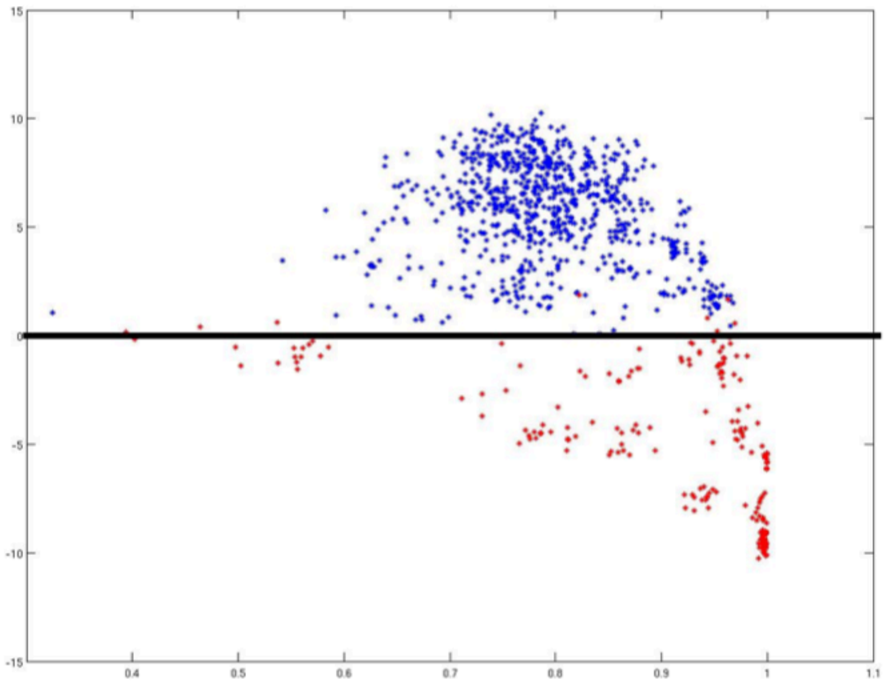
axe X represents the Diversity and axe Y - the Accuracy gain

Collaboration results (2)

Collaboration results between a fixed subset and 1000 randomly subsets



(b) SpamBase dataset



(c) Isolet dataset

axe X represents the Diversity and axe Y - the Accuracy gain

Images : Strasbourg satellite image (1)

Projet COCLICO

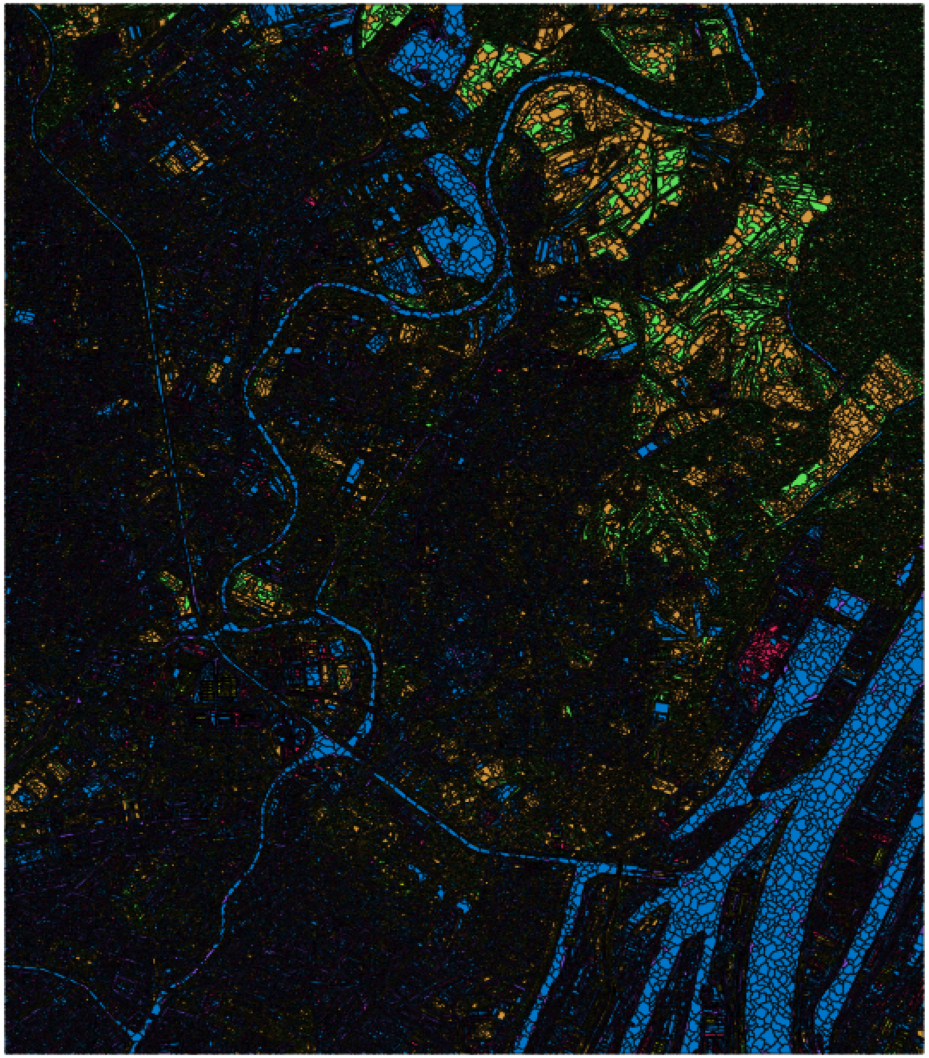
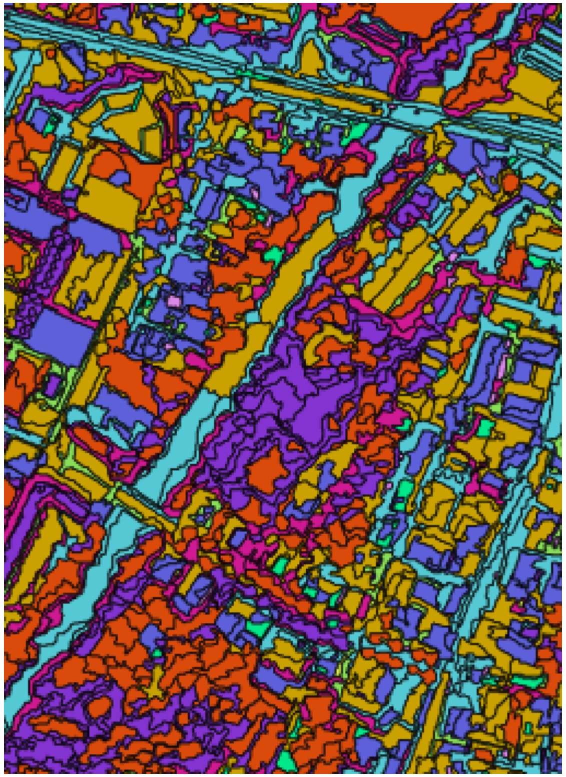


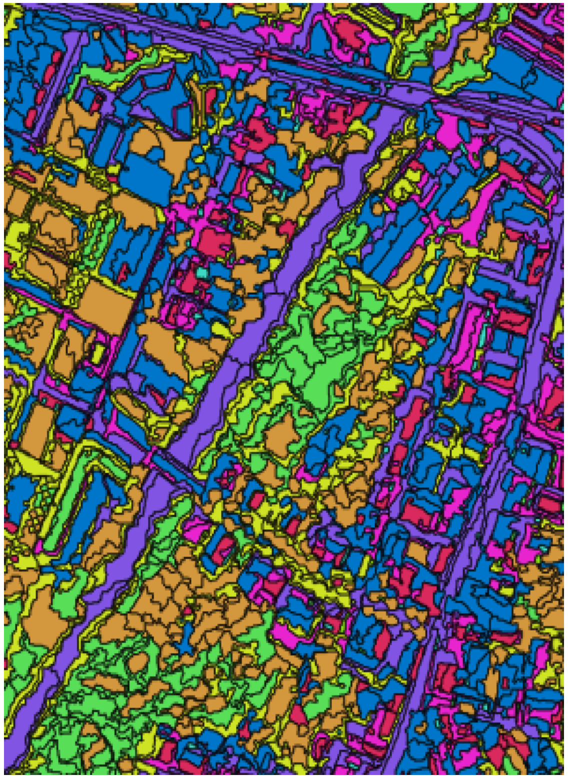
Image Ville Environnement

The authors would like to thank CESBIO (Danielle Ducrot, ClaireMarais-Sicre, Olivier Hagolle, Mireille Huc and Jordi Inglada) for providing the land-cover maps and the geometrically and radiometrically corrected Formosat-2 images.

Before collaboration



After collaboration



Conclusions & Future works

- The collaborative clustering allows:
  - An interaction between different datasets
  - Reveal underlying structures and patterns within data sets.

- During the collaboration step, where is no need of data, the algorithm requires only the clustering results of other datasets.
  - obtain a new classification that is as close as possible to that which would have obtained if we had centralized datasets and then make a partition.

- The quality of the local clustering algorithm is very important for the collaboration's quality improvement regarding the diversity index
  - Overall, the variability of the collaboration's quality increase with the diversity

- Create a «*helper site*» which will build the global clustering and send these information to other local sites

- Use the diversity for Selective Collaborative Clustering

Collaborative Generative Topographic Mapping

Horizontal approach

$$\mathcal{L}^{hor}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1,[jj]\neq[ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N} \sum_{i=1}^{K} \frac{\beta^{[ii]}}{2} (r_{in}^{[ii]} - r_{in}^{[jj]})^2 \|x_n - W^{[ii]}\phi^{[ii]}(z_i)\|^2$$

Vertical approach

$$\mathcal{L}^{ver}[ii] = \mathbb{E}[\mathcal{L}_{comp}(W^{[ii]}, \beta^{[ii]})] - \sum_{[jj]=1,[jj]\neq[ii]}^{P} \alpha_{[ii]}^{[jj]} \sum_{n=1}^{N[ii]} \sum_{i=1}^{K} r_{in} \frac{\beta^{[ii]}}{2} \|W^{[ii]}\phi^{[ii]}(z_i) - W^{[jj]}\phi^{[jj]}(z_i)\|^2$$