Unsupervised Collaborative Learning

Nistor Grozavu
Plan

- **Introduction**
- **The problem**
  - Consensus Clustering & Collaborative Clustering
- **Collaborative clustering**
  - Horizontal collaboration
  - Vertical collaboration
- **Topological Collaborative Clustering**
- **Diversity Analysis**
  - The problem
  - Proposed solutions
- **Real applications**
- **Conclusions**
Introduction

Classification

The labels and the number of classes are known

Clustering

The labels and the number of classes are unknown

Difficulties

- The objects are not labeled, …
- We need to use a similarity measure (for which variables?)
- Do we need to know a priori the number of classes?
- How to characterize clusters?
Introduction : Clustering

- Grouping together of “similar” objects
  - Hard Clustering -- Each object belongs to a single cluster
  - Soft Clustering -- Each object is probabilistically assigned to clusters

In general, the formalization of the Clustering problem is determined by the following components:

- Data representation (categorical, binary, graph…)
- The affinity measures (similarity, distance,..)
- The objective function
- The optimization procedure
- Data distribution …
The principle of the Fusion

\[ C_1 \]
\[ c_i^1 \quad c_i^2 \]
\[ : \quad : \]
\[ c_i^1 \quad c_i^2 \]

Algo fusion

\[ C_{\text{fus}} \]
\[ c_i^f \quad c_i^f \]
\[ c_i^f \quad c_i^f \]
\[ : \quad : \]
\[ c_i^f \]

The principle of the Collaboration

\[ C_1 \]
\[ c_i^1 \quad c_i^c \]
\[ : \quad : \]
\[ c_i^1 \quad c_i^c \]

\[ C_{\text{col}} \]
\[ c_i^c \quad c_i^c \]

- 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;
Collaboration : principle
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, 2002)
- Compute the normalized mutual information (NMI) for each dataset;
- Compute the mean of the NMI for a set of $r$ classes (labels) - ANMI;

\[
NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}
\]

\[
\phi^{(ANMI)}(\Lambda, \hat{\lambda}) = \frac{1}{r} \sum_{q=1}^{r} \phi^{(NMI)}(\hat{\lambda}, \lambda^{(q)})
\]
Distributed Data Clustering (DDC):

- KDEC – Density estimation based Distributed Clustering;
- Compute the densities for each local DB:
  \[
  \hat{\phi}_{K,h}[S](\bar{x}) = \sum_{i=1}^{N} K \left( \frac{d(\bar{x}, \bar{x}_i)}{h} \right)
  \]
- Send these densities to a « helper site » which will build the global clustering and send these information to other local sites.
Bennani et al., 2009 (Fusion)

- Fusion of several classifications using Relational Analysis approach (AR)

1st fusion of SOM clustering: RA consensus

**Coding Type 1 (NxC)**

- SOM_1: X_{11} ... X_{1C}
- O_1: X_{11} ... X_{1C}
- O_2: X_{21} ... X_{2C}
- ...
- O_N: X_{N1} ... X_{NC}

**Coding Type 2 (NxD)**

- SOM_1: X_{11} ... X_{1D}
- O_1: X_{11} ... X_{1D}
- O_2: X_{21} ... X_{2D}
- ...
- O_N: X_{N1} ... X_{ND}

... ...

**Coding Type K (NxE)**

- SOM_1: X_{11} ... X_{1E}
- O_1: X_{11} ... X_{1E}
- O_2: X_{21} ... X_{2E}
- ...
- O_N: X_{N1} ... X_{NE}

Fusion of C SOM clustering results by RA from Type 1 coding

<table>
<thead>
<tr>
<th>RA_1</th>
<th>O_1</th>
<th>X_{1E}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_N</td>
<td>X_{NE}</td>
<td></td>
</tr>
</tbody>
</table>

Fusion of D SOM clustering results by RA from Type 2 coding

<table>
<thead>
<tr>
<th>RA_2</th>
<th>O_1</th>
<th>X_{1E}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_N</td>
<td>X_{NE}</td>
<td></td>
</tr>
</tbody>
</table>

... ...

Fusion of E SOM clustering results by RA from Type K coding

<table>
<thead>
<tr>
<th>RA_K</th>
<th>O_1</th>
<th>X_{1E}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_N</td>
<td>X_{NE}</td>
<td></td>
</tr>
</tbody>
</table>

... ...
Distributed Data Clustering (DDC):

- KDEC – Density estimation based Distributed Clustering;

- Compute the densities for each local DB:

\[ \hat{\Phi}_{K,h}[S](\bar{x}) = \sum_{i=1}^{N} K \left( \frac{d(\bar{x}, \bar{x}_i)}{h} \right) \]

- Send these densities to a "helper site" which will build the global clustering and send these information to other local sites.
Fuzzy C-Means Clustering (FCM):

- For each dataset, build granular prototypes using the partitions matrix;
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.
## Horizontal collaboration

<table>
<thead>
<tr>
<th>ID</th>
<th>Att1</th>
<th>Att2</th>
<th>Att3</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>id9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dataset [1]

<table>
<thead>
<tr>
<th>ID</th>
<th>Att4</th>
<th>Att5</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dataset [2]

<table>
<thead>
<tr>
<th>ID</th>
<th>Att6</th>
<th>Att7</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dataset [P]

---

Same samples
### Vertical Collaboration

<table>
<thead>
<tr>
<th>ID</th>
<th>Att 1</th>
<th>Att 2</th>
<th>Att 3</th>
<th>Att 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Dataset [1]

<table>
<thead>
<tr>
<th>ID</th>
<th>Att 1</th>
<th>Att 2</th>
<th>Att 3</th>
<th>Att 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Dataset [2]

### Same variables

... 

### Dataset [P]

<table>
<thead>
<tr>
<th>ID</th>
<th>Att 1</th>
<th>Att 2</th>
<th>Att 3</th>
<th>Att 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Nxd₃

Nxd₁

Nxd₂
The problem

- How to improve the local clustering derived out of a set of distant clustering results without sharing the initial data?
Collaborative FCM (Pedrycz, 2002)
The distance function between the $i$th prototype and the $k$th pattern in the same subset is denoted by $d_{ik}^{2[ii]}$, $i = 1, 2, \ldots, c$, $k = 1, 2, \ldots, N$ and $ii = 1, 2, \ldots, P$. 

$$\sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik}[ii]d_{ik}^{2[ii]}$$
Each entry of the collaborative matrix describes the intensity of the interaction. In general, $\alpha[i ii, kk]$ assumes nonnegative values.

Collaboration in the clustering scheme represented by the matrix of collaboration levels between the subsets; the partition matrices generated for each data set are shown.
General collaborative clustering scheme  (Pedrycz, 2002)

Given: subsets of patterns $X_1, X_2, \ldots, X_P$

Select: distance function, number of clusters ($c$), termination criterion, and collaboration matrix $\alpha[ii, jj]$.

Compute: initiate randomly all partition matrices $U[1], U[2], \ldots, U[P]$

**Phase I**

For each data

repeat

compute prototypes $\{v_i[ii]\}, i = 1, 2, \ldots, c$ and partition matrices $U[ii]$ for all subsets of patterns

until a termination criterion has been satisfied

**Phase II**

repeat

For the given matrix of collaborative links $\alpha[ii, jj]$, compute prototypes and partition matrices $U[ii]$ using (7.4) and (7.7)

until a termination criterion has been satisfied
The intensity of collaboration:

$$\delta = \|U[ii] - U_{\text{ref}[ii]}\|$$

$U[ii]$ to denote the partition matrix produced independently of other sources.

Consistency measure:

$$\phi[ii, jj] = \|U[ii] - U[jj]\|$$

indicates the structural differences between the partition matrices defined over two data sets (ii and jj, respectively).
Topological Collaborative Clustering
Prototype based Clustering (SOM)

\[ K_{\delta(i,j)} = \exp\left(-\frac{\delta^2(i,j)}{T^2}\right) \]

Neighborhood function

\[ w^* = \arg\min_{w} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{|W|} K_{\delta(j,\chi(x_i))} \left\| x_i - w_j \right\|^2 \right\} \]

Assignment function

\[ \chi(x_i) = \arg\min_{j} \left( \left\| x_i - w_j \right\|^2 \right) \]

N : number of observations x, |W| : number of prototypes w
Horizontal Collaboration

\[ w^* = \arg \min_w \left\{ R_{SOM}^{[ii]}(\chi, w) + R_{Col - H}^{[ii]}(\chi, w) \right\} \]

Collaboration term

\[ R_{Col - H}^{[ii]}(\chi, w) = \sum_{jj=1, jj \neq ii}^P \sum_{i=1}^N \sum_{j=1}^{|w|} \left( K_{\delta}(j, \chi(x_i)) - K_{\delta}(j, \chi(x_i)) \right)^2 \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2 \]

\[ R_{SOM}^{[ii]}(\chi, w) = \sum_{i=1}^N \sum_{j=1}^{|w|} K_{\delta}(j, \chi(x_i)) \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2 \]

\( N^{[ii]} \) : the number of observations on the datasets [ii], \( P \) : the number of datasets
Algorithm 1: The horizontal collaboration algorithm

Fix the collaboration matrix $\alpha_{[ji]}^{[jj]}$

1. Local step:
   For each dataset $BD[ii]$, $ii = 1$ to $P$ :
   
   Find the prototypes minimizing the classical SOM objective function:
   
   $$w^* = \arg \min_w R_{SOM}^{[ii]}(\chi, w)$$

2. Collaboration step:
   For the horizontal collaboration of the $[ii]$-th map with the $[jj]$-th map:
   
   Update the prototypes of the $[ii]$-th map minimizing the objective function of the horizontal collaboration:

   $$w_{jk}^*[ii] = \frac{\sum_{i=1}^{N} K_{\sigma(j, \chi(x_i))}^{[ii]} x_{ik}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j, \chi(x_i))}^{[ii]} - K_{\sigma(j, \chi(x_i))}^{[jj]} \right)^2 x_{ik}^{[ii]}}{\sum_{i=1}^{N} K_{\sigma(j, \chi(x_i))}^{[ii]} + \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j, \chi(x_i))}^{[ii]} - K_{\sigma(j, \chi(x_i))}^{[jj]} \right)^2}$$
Vertical Collaboration

\[
w^* = \arg\min_{w} \left\{ R_{SOM}^{[ii]} (\chi, w) + R_{Col \_ V}^{[ii]} (\chi, w) \right\}
\]

Collaboration term

\[
R_{Col \_ V}^{[ii]} (\chi, w) = \sum_{jj=1, jj \neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \sum_{j=1}^{|w|} \left( K_{\delta (j, \chi (x_i))}^{[ji]} - K_{\delta (j, \chi (x_i))}^{[jj]} \right)^2 \left\| w_j^{[ii]} - w_j^{[ii]} \right\|^2
\]

\[
R_{SOM}^{[ii]} (\chi, w) = \sum_{i=1}^{N} \sum_{j=1}^{|w|} K_{\delta (j, \chi (x_i))}^{[ji]} \left\| x_i^{[ii]} - w_j^{[ii]} \right\|^2
\]

\[N^{[ii]} : \text{the number of observations on dataset [ii], } P : \text{the number of datasets sites}\]
Algorithm 2: Vertical Collaboration algorithm

1. Local step:
For each dataset $BD[ii]$, $ii = 1$ to $P$:

Find the prototypes minimizing the classical SOM objective function:

$$w^* = \arg \min_w \left[ R_{SOM}^{[ii]}(\chi, w) \right]$$

2. Collaboration step:
For the vertical collaboration of the $[ii]$-th map with the map $[jj]$:

Update the prototypes of the $[ii]$-th map minimizing the objective function of the vertical collaboration:

$$w^*_j[k] = \frac{\sum_{i=1}^{N^{[ii]}} K_{\sigma(j, \chi(x_i))}^{[ii]} x_i^{[ii]} + \sum_{j=1, jj \neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j, \chi(x_i))}^{[ii]} - K_{\sigma(j, \chi(x_i))}^{[jj]} \right)^2 w_{j[k]}^{[jj]} \sum_{i=1}^{N^{[ii]}} K_{\sigma(j, \chi(x_i))}^{[ii]} + \sum_{j=1, jj \neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j, \chi(x_i))}^{[ii]} - K_{\sigma(j, \chi(x_i))}^{[jj]} \right)^2}{\sum_{i=1}^{N^{[ii]}} K_{\sigma(j, \chi(x_i))}^{[ii]} + \sum_{j=1, jj \neq ii}^{P} \sum_{i=1}^{N^{[ii]}} \alpha_{[ii]}^{[jj]} \left( K_{\sigma(j, \chi(x_i))}^{[ii]} - K_{\sigma(j, \chi(x_i))}^{[jj]} \right)^2}$$
Ilustrative example

- **Waveform dataset**
  - 5000 samples
  - 40 variables where 19 variables are Gaussian noisy
  - 3 classes
Horizontal Collaboration (waveform)
Horizontal Collaboration (waveform)

The prototypes of the 1st dataset before the collaboration: SOM1

\[ 75.71\% \]

The prototypes of the 2nd dataset before the collaboration: SOM2

\[ 79.61\% \]

The prototypes of the 1st dataset after the collaboration with the SOM2 map: SOM12

\[ 76.21\% \]
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%
Validation de l’approche collaboratif sur différents bases de données

Collaboration horizontale

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DB..</th>
<th>Purity</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>wdbc</td>
<td>SOM1</td>
<td>94.9550</td>
<td>1.9993</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>97.2777</td>
<td>2.0749</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>islet 5x5</td>
<td>SOM1</td>
<td>81.2081</td>
<td>12.6149</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>95.1220</td>
<td>14.4591</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>madelon</td>
<td>SOM1</td>
<td>60.8879</td>
<td>15.5896</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>62.6402</td>
<td>15.5065</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spam</td>
<td>SOM1</td>
<td>83.8603</td>
<td>3.4582</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>85.7205</td>
<td>2.5580</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Collaboration verticale

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DB..</th>
<th>Purity</th>
<th>QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>wdbc</td>
<td>SOM1</td>
<td>96.7153</td>
<td>90.5413</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>97.8723</td>
<td>67.6035</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>islet 5x5</td>
<td>SOM1</td>
<td>98.8506</td>
<td>8.1904</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>98.4615</td>
<td>8.7671</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>madelon</td>
<td>SOM1</td>
<td>69.7198</td>
<td>612.3251</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>69.8718</td>
<td>611.5365</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spam</td>
<td>SOM1</td>
<td>76.2624</td>
<td>61.8324</td>
</tr>
<tr>
<td></td>
<td>SOM2</td>
<td>70.4306</td>
<td>48.2763</td>
</tr>
<tr>
<td></td>
<td>SOM12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOM21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic Collaborative Clustering
Probabilistic Clustering

Generative Topographic Mapping [Bishop 95]

\[ y = y(z, W) = W\Phi(z) \]

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

\[ \mathcal{L}(W, \beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(x_n|z_i, W, \beta) \right\} \implies \text{EM Algorithm} \]
E & M steps

E step - Computing posterior probabilities

\[ r_{in} = \frac{p(z_i | x_n, W_{old}, \beta_{old})}{\sum_{i'=1}^{K} p(x_n | z_i', W_{old}, \beta_{old})} \]

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**

\[ R_{H}^{[i\bar{i}]}(W) = R_{Quantiz}(W) + R_{Collab}(W) \]

- **Prototype based Clustering**

\[
R_{Quantiz}(W) = \sum_{j \neq i} P \alpha_{[i\bar{i}]} \sum_{i=1}^{N} \sum_{j=1}^{\mid w \mid} \mathcal{K}_{\sigma(j, \chi(x_i))} \| x_{i}^{[i\bar{i}]} - w_{j}^{[i\bar{i}]} \|^2
\]

\[
R_{Collab}(W) = \sum_{j \neq i} P \beta_{[i\bar{i}]} \sum_{i=1}^{N} \sum_{j=1}^{\mid w \mid} \left( \mathcal{K}_{\sigma(j, \chi(x_i))}^{[i\bar{i}]} - \mathcal{K}_{\sigma(j, \chi(x_i))}^{[j\bar{j}]} \right)^2 \| x_{i}^{[i\bar{i}]} - w_{j}^{[i\bar{i}]} \|^2
\]

- **Probabilistic Clustering**

\[
\mathcal{L}_{hor}^{[i\bar{i}]} = \mathbb{E}[\mathcal{L}_{comp}(W^{[i\bar{i}]}, \beta^{[i\bar{i}]})] - \sum_{[\bar{j}]=1, [\bar{j}] \neq [i\bar{i}]}^{P} \alpha_{[i\bar{i}]} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\beta_{[i\bar{i}]}}{2} (r_{in}^{[i\bar{i}]} - r_{in}^{[\bar{j}]})^2 \| x_n - W^{[i\bar{i}]} \phi^{[i\bar{i}]}(z_i) \|^2
\]
Collaborative Generative Topographic Mapping

Horizontal approach

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

Vertical approach

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Map</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waveform</td>
<td>$GTM_1$</td>
<td>86.44</td>
</tr>
<tr>
<td></td>
<td>$GTM_2$</td>
<td>86.52</td>
</tr>
<tr>
<td></td>
<td>$GTM_{1\rightarrow2}$</td>
<td>87.16</td>
</tr>
<tr>
<td></td>
<td>$GTM_{2\rightarrow1}$</td>
<td>87.72</td>
</tr>
<tr>
<td>Wdbc</td>
<td>$GTM_1$</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>$GTM_2$</td>
<td>96.34</td>
</tr>
<tr>
<td></td>
<td>$GTM_{1\rightarrow2}$</td>
<td>96.08</td>
</tr>
<tr>
<td></td>
<td>$GTM_{2\rightarrow1}$</td>
<td>96.15</td>
</tr>
<tr>
<td>Isolet</td>
<td>$GTM_1$</td>
<td>87.17</td>
</tr>
<tr>
<td></td>
<td>$GTM_2$</td>
<td>86.83</td>
</tr>
<tr>
<td></td>
<td>$GTM_{1\rightarrow2}$</td>
<td>87.29</td>
</tr>
<tr>
<td></td>
<td>$GTM_{2\rightarrow1}$</td>
<td>85.87</td>
</tr>
<tr>
<td>SpamBase</td>
<td>$GTM_1$</td>
<td>52.05</td>
</tr>
<tr>
<td></td>
<td>$GTM_2$</td>
<td>51.68</td>
</tr>
<tr>
<td></td>
<td>$GTM_{1\rightarrow2}$</td>
<td>52.41</td>
</tr>
<tr>
<td></td>
<td>$GTM_{2\rightarrow1}$</td>
<td>52.17</td>
</tr>
</tbody>
</table>
Collaborative Clustering
Diversity analysis
Diversity: why?

Studied in Consensus clustering

Dataset X containing 15 samples

<table>
<thead>
<tr>
<th></th>
<th>Algo1</th>
<th>Algo2</th>
<th>Algo3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
</tr>
</tbody>
</table>

Accuracy

10/15 = 0.667
10/15 = 0.667
10/15 = 0.667
Diversity: why?

Studied in Consensus clustering

Dataset X containing 15 samples

Algo1

Algo2

Algo3

Fusion

Accuracy

10/15 = 0.667
10/15 = 0.667
10/15 = 0.667
11/15 = 0.773

Majority vote rule
Diversity: why?

Studied in Consensus clustering

Dataset X containing 15 samples

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algo1</td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td>Algo2</td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td>Algo3</td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td>Fusion</td>
<td>11/15 = 0.773</td>
</tr>
</tbody>
</table>

Accuracy calculated with the Majority vote rule.
Diversity: why?

Studied in Consensus clustering

Dataset X containing 15 samples

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Fusion Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algo1</td>
<td>10/15 = 0.667</td>
<td>11/15 = 0.773</td>
</tr>
<tr>
<td>Algo2</td>
<td>10/15 = 0.667</td>
<td></td>
</tr>
<tr>
<td>Algo3</td>
<td>10/15 = 0.667</td>
<td></td>
</tr>
<tr>
<td>Fusion</td>
<td>10/15 = 0.667</td>
<td></td>
</tr>
</tbody>
</table>

- Majority vote rule

Nxd

Gamma

Correct

Wrong
Diversity: why?

Studied in Consensus clustering

Dataset X containing 15 samples

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correct</th>
<th>Wrong</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algo1</td>
<td><img src="image" alt="Correct" /></td>
<td><img src="image" alt="Wrong" /></td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td>Algo2</td>
<td><img src="image" alt="Correct" /></td>
<td><img src="image" alt="Wrong" /></td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td>Algo3</td>
<td><img src="image" alt="Correct" /></td>
<td><img src="image" alt="Wrong" /></td>
<td>10/15 = 0.667</td>
</tr>
<tr>
<td><strong>Fusion</strong></td>
<td><img src="image" alt="Correct" /></td>
<td><img src="image" alt="Wrong" /></td>
<td>11/15 = 0.773</td>
</tr>
</tbody>
</table>

Majority vote rule
Diversity : why?

Studied in Consensus clustering

Dataset X containing 15 samples

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Correct (10/15)</th>
<th>Wrong (10/15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algo1</td>
<td>0.667</td>
<td>10/15</td>
<td>5/15</td>
</tr>
<tr>
<td>Algo2</td>
<td>0.667</td>
<td>10/15</td>
<td>5/15</td>
</tr>
<tr>
<td>Algo3</td>
<td>0.667</td>
<td>10/15</td>
<td>5/15</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.773</td>
<td>11/15</td>
<td>4/15</td>
</tr>
</tbody>
</table>

- Majority vote rule

Fusion accuracy: 11/15 = 0.773

- Majority vote rule

Fusion accuracy: 10/15 = 0.667

- Majority vote rule

Fusion accuracy: 8/15 = 0.533

- Majority vote rule
Collaborative clustering

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GTM1&lt;-2</th>
<th>GTM2&lt;-1</th>
<th>GTM3&lt;-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
</tr>
<tr>
<td>X2</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
</tr>
<tr>
<td>X3</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
<td>[Diagram]</td>
</tr>
</tbody>
</table>

Accuracy:
- X1: 8/15 = 0.533
- X2: 12/15 = 0.8
- X3: 11/15 = 0.733
# Diversity measures

<table>
<thead>
<tr>
<th>index</th>
<th>formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand index</td>
<td>(\text{Rand} = \frac{a_{00} + a_{11}}{a_{00} + a_{01} + a_{10} + a_{11}})</td>
</tr>
<tr>
<td>Adjusted Rand index</td>
<td>(\text{AdjustedRand} = \frac{a_{00} + a_{11} - n_c}{a_{00} + a_{01} + a_{10} + a_{11} - n_c})</td>
</tr>
<tr>
<td>Jaccard index</td>
<td>(\text{Jaccard} = \frac{a_{11}}{a_{01} + a_{10} + a_{11}})</td>
</tr>
<tr>
<td>Wallace’s coefficient</td>
<td>(W_{P_1 \to P_2} = \frac{a_{11}}{a_{11} + a_{10}}) and (W_{P_2 \to P_1} = \frac{a_{11}}{a_{11} + a_{01}})</td>
</tr>
<tr>
<td>Adjusted Wallace index</td>
<td>(\text{AW}<em>{P_1 \to P_2} = \frac{W</em>{P_1 \to P_2} - W_{i_{P_1 \to P_2}}}{1 - W_{i_{P_1 \to P_2}}})</td>
</tr>
<tr>
<td>Normalized Mutual Information</td>
<td>(\text{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}{N} + \sum_j n_j \log \frac{n_j N}{N}})</td>
</tr>
<tr>
<td>Variation of Information</td>
<td>(\text{VI} = -2 \sum_{ij} \frac{n_{ij} \log \frac{n_{ij} N}{n_i n_j}}{N} - \sum_i \frac{n_i \log \frac{n_i N}{N}}{N} - \sum_j \frac{n_j \log \frac{n_j N}{N}}{N})</td>
</tr>
</tbody>
</table>
Diversity measures on waveform datasets

![Graphs showing diversity measures on different datasets](image)

### Table 1: Diversity measure on the waveform subsets

<table>
<thead>
<tr>
<th>Subset</th>
<th>Relevant datasets</th>
<th>Relevant vs Noisy datasets</th>
<th>Noisy datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity index</td>
<td>db2/db3</td>
<td>db2/db8</td>
<td>db7/db8</td>
</tr>
<tr>
<td>Rand</td>
<td>0.6707</td>
<td>0.5539</td>
<td>0.543</td>
</tr>
<tr>
<td>Adjusted Rand</td>
<td>0.2625</td>
<td>0.00008</td>
<td>0.00002</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.3429</td>
<td>0.2017</td>
<td>0.2008</td>
</tr>
<tr>
<td>Wallace’s coefficient</td>
<td>0.5079</td>
<td>0.3332</td>
<td>0.33</td>
</tr>
<tr>
<td>Adjusted Wallace</td>
<td>0.5135</td>
<td>0.3383</td>
<td>0.35</td>
</tr>
<tr>
<td>Normal Mutual Information</td>
<td>0.262</td>
<td>0.0002</td>
<td>0.0003</td>
</tr>
<tr>
<td>Variation of Information</td>
<td>2.334</td>
<td>3.1577</td>
<td>3.168</td>
</tr>
</tbody>
</table>
Collaborative clustering

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

X1

\[
\begin{array}{c}
\text{Correct} \\
\text{Wrong}
\end{array}
\]

\[
\begin{array}{c}
\text{Accuracy} \\
\text{Diversity}
\end{array}
\]

X1-X2 = 0.956
X2-X3 = 0.678

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

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
(b) Wdbc dataset

dx X represents the Diversity and axe Y - the Accuracy gain
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
Collaborative Clustering and Consensus Clustering

real applications
The authors would like to thank CESBIO (Danielle Ducrot, Claire Marais-Sicre, Olivier Hagolle, Mireille Huc and Jordi Inglada) for providing the land-cover maps and the geometrically and radiometrically corrected Formosat-2 images.
Images : Strasbourg satellite image (2)  Projet COCLICO

Before collaboration

After collaboration
System for searching visual information
Multimodal information: prediction

Application context:
- A Wikipedia dataset containing a set of 20,000 images from wikipedia and transformed into numerical values by Xerox research center.

Results:
- Reduce the dimensionality of this dataset of size 20,000 x 12,800 into 20,000 x 10
- Patent (THALES, Paris 13 University) No: 08 06947
  Inventors: BENHADDA H., BENNANI Y., LEBBAH M., GROZAVU N.
Recall: topological learning

\[ \mathbb{R}_{SOM}(\chi, W) = \sum_{i=1}^{N} \sum_{j=1}^{W} K_{j, \chi(x_i)} \| x_i - w_j \|^2 \]
Conventional search engine

Example of image search using a traditional search engine

*(France flag)*

Research time: **0,11s**

Browsing the collection of images by user: **15 days**

(13.700.000 images / 21 images/pages = 652.380 pages * 2 s = 1.304.760s (362h or 15 days))
Map of images

Wikipedia (19,000 x 6400) x 2
Hierarchical SOM (3D view)

\[ \chi(M_d) = \left( \| x_i - w_j \|^2 \right) \]
Intelligent system based on the Topological Learning

Feature extraction from images
Principle

Figure 3.0: General Schema of the Fusion Procedure
Relational Analysis (Marcotorchino al. 1980)

1- Pairwise comparison principle

We denote \( F(R, X) \) - the linear criterion measuring the adequacy between the data \( R \) and the solution \( X \), the mathematical formulation of the problem is:

\[
\max_{X} F(R, X)
\]

Under the linear constraints generated by the properties of \( X \).

2- (0,1) Linear programming modelling

Linear coding  Complete Disjunctive Coding  Relational Coding
Coding et Fusion

1st fusion of SOM clustering: RA consensus

Coding Type 1 (NxC)

SOM_1 SOM_2 ... SOM_C
O_1 : x_11 x_12 ... x_1C
O_2 : x_21 x_22 ... x_2C
...
O_N : x_N1 x_N2 ... x_NE

Fusion of C SOM clustering results by RA from Type 1 coding

Coding Type 2 (NxD)

SOM_1 SOM_2 ... SOM_D
O_1 : x_11 x_12 ... x_1D
O_2 : x_21 x_22 ... x_2D
...
O_N : x_N1 x_N2 ... x_NE

Fusion of D SOM clustering results by RA from Type 2 coding

Coding Type K (NxK)

SOM_1 SOM_2 ... SOM_K
O_1 : x_11 x_12 ... x_1K
O_2 : x_21 x_22 ... x_2K
...
O_N : x_N1 x_N2 ... x_NK

Fusion of E SOM clustering results by RA from Type K coding

RA_1
O_1 X_1E
...
O_N X_NE

RA_2
...

RA_K
...
O_N X_NE
Dimensionality reduction

Figure 3.7: Pre-processing of the images dataset

Figure 3.8: Dimensionality Reduction
\[ \chi(x_i) = \arg \min_{j} \left( \| x_i - w_j \|^2 \right) \]
Demonstration video

BREVET (THALES, Université Paris 13) N°: 08 06947


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