

Apprentissage spectral pour la caractérisation et la reconnaissance du Phytoplancton par cytométrie en flux

Guillaume WACQUET

Équipe Extraction de l'Information et Apprentissage

LISIC, ULCO EA 4491

Calais, France

Guillaume.Wacquet@lisic.univ-littoral.fr



Calais, le 06 Mai 2010



Plan

Le Phytoplancton

- ✘ *Présentation générale*
- ✘ *Les enjeux*

La cytométrie en flux

- ✘ *Présentation générale*
- ✘ *Analyse cytométrique*
- ✘ *Difficultés de discrimination*

La classification

- ✘ *Classification supervisée*
- ✘ *Classification non supervisée*
- ✘ *Classification semi-supervisée*

Le Phytoplancton

Présentation générale

- Étymologie : du grec « phyton » et « plagkton » qui signifient respectivement « plante » et « errant »
=> Plancton végétal microscopique qui erre au gré des courants (entre 1 μm et plusieurs mm)
 - Organisme photosynthétique : fabrication de sa biomasse grâce à l'absorption des sels minéraux et du carbone inorganique (sous forme de CO_2) sous l'effet de la lumière
 - 1% de la biomasse d'organismes photosynthétiques de la planète
 - Environ 6000 espèces au niveau mondial dont 70 toxiques ou nuisibles (en particulier en zones côtières)
-
-

Présentation générale



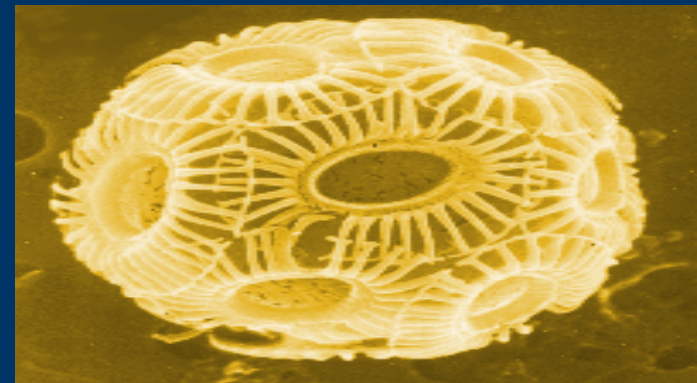
Cyanobactérie (3,8 Ga)
taille : entre 1 et 10 microns



Diatomée (185 Ma)
taille : entre 20 et 200 microns



Dinoflagellé (420 Ma)
taille : plusieurs dizaines de microns



Coccolithophoridé (150 Ma)
taille : entre 10 et 30 microns

Les enjeux

Importance de l'étude de l'environnement des milieux littoraux (notamment les écosystèmes marins) :

- Sur les plans écologique et climatique :
 - *Préservation de l'environnement, biodiversité*
 - 50% de l'activité photosynthétique totale de la planète
 - 45% de la production primaire mondiale
- Sur le plan économique :
 - *Tourisme, Transport, Ressources exploitables et Nourriture*
 - Premier maillon de la chaîne alimentaire dans l'écosystème marin
 - Environ 70 espèces toxiques ou nuisibles

La cytométrie en flux

Présentation générale

- Création des premiers cytomètres en flux dans les années 1950
- Outil de mesures permettant d'obtenir certaines propriétés physiques et optiques des particules par un écoulement à grande vitesse devant un faisceau laser
- Obtention de signaux temporels → Analyse de la diffusion et de la fluorescence des particules afin de classer la population suivant plusieurs critères

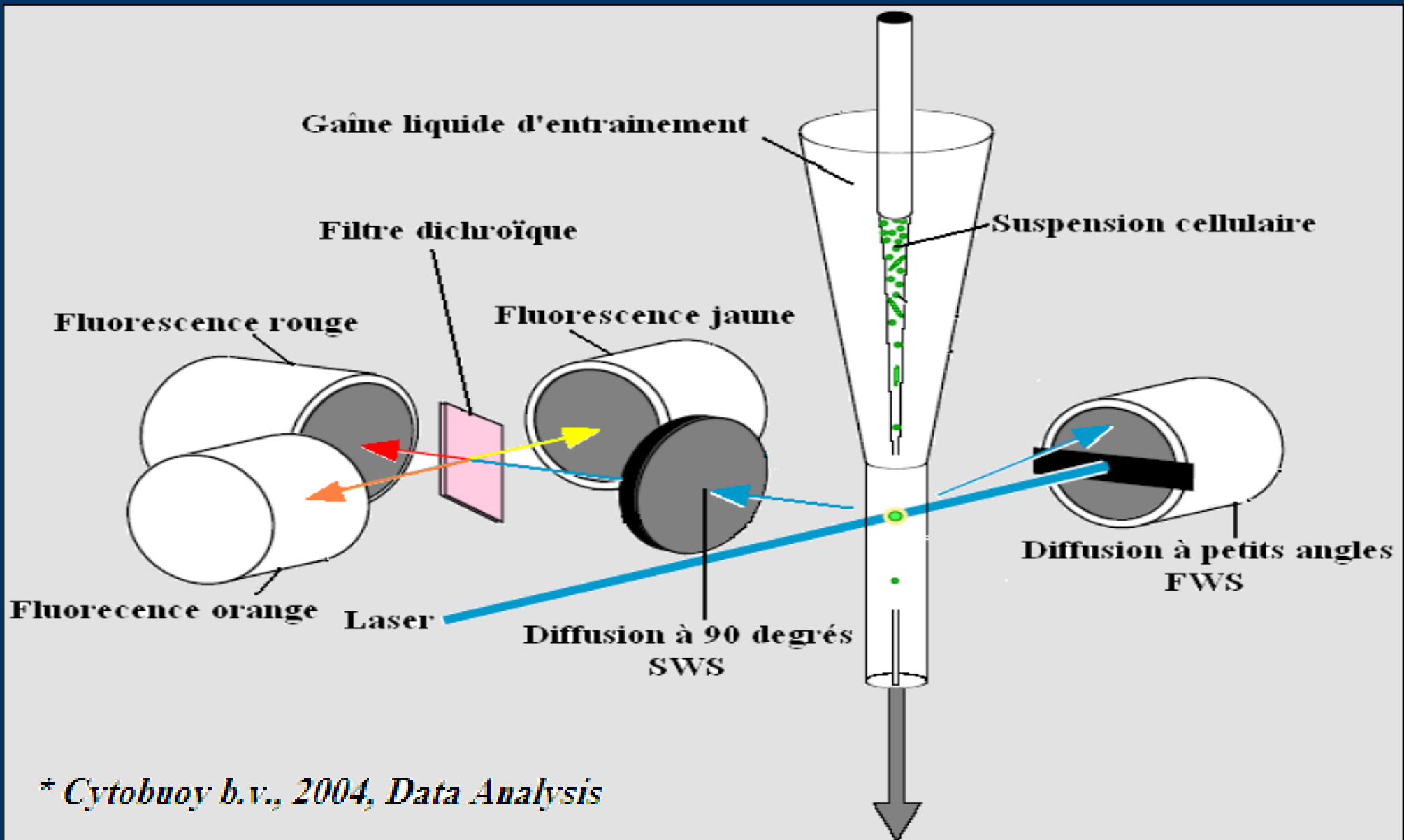


CytoSub©



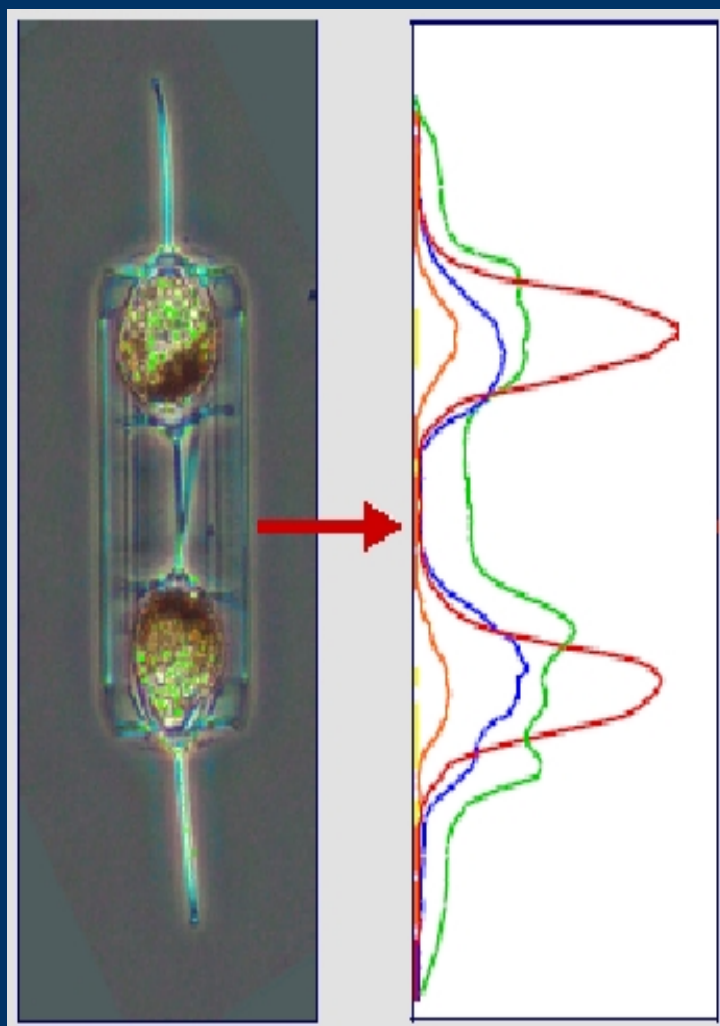
CytoSense©

Analyse cytométrique*



* *Cytobuoy b.v., 2004, Data Analysis*

Analyse cytométrique



Courbes cytométriques :

- ✕ 8 signaux par cellule
 - ✕ Acquisition réalisée dans des conditions expérimentales identiques (fréquences échantillonnage égales, seuils de détections égaux, etc.)
- 1 signal de diffusion à petits angles **FWS**, qui correspond à la structure externe de la cellule
 - 2 signaux de diffusion à 90° **SWS** (en haute et basse sensibilité), qui correspondent à la structure interne
 - 2 signaux de fluorescence rouge **FLR** (en haute et basse sensibilité), qui caractérisent les pigments de chlorophylle
 - 1 signal de fluorescence orange **FLO**, qui caractérise des pigments spécifiques
 - 2 signaux de fluorescence jaune **FLY**, qui caractérisent des pigments spécifiques

Difficultés de discrimination

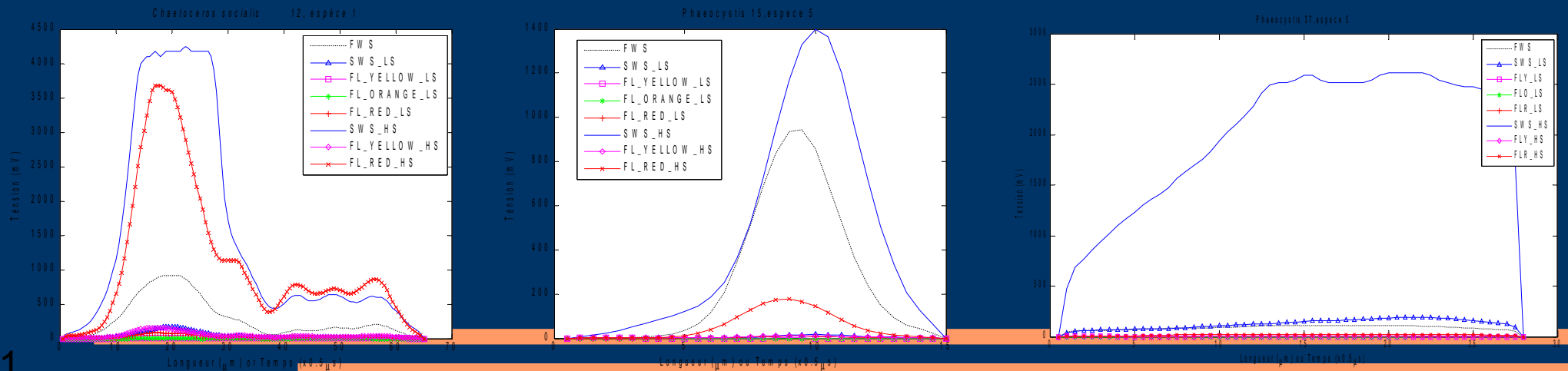
- **Entre espèces différentes :**

Grande diversité de taille et de contenu pigmentaire (en qualité mais aussi en quantité), fonction du groupe auquel elles appartiennent mais également des conditions du milieu naturel

- **Pour une même espèce :**

Structure interne (noyau et chloroplastes) variable pour une même espèce selon son état (cellule, colonie, cycle de vie, etc.)

=> Variabilité de la position des pics sur les courbes et variabilité des intensités des signaux de fluorescence

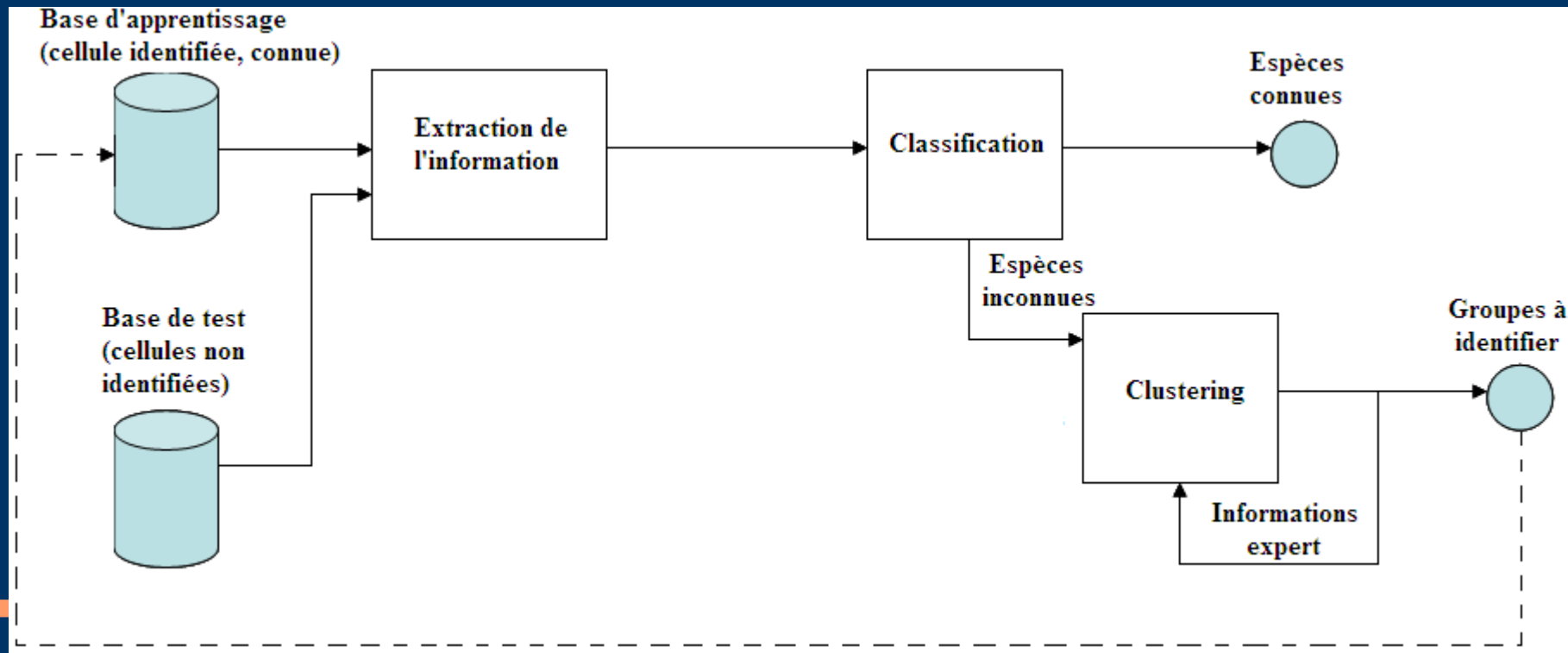


La classification

La classification

3 modes de classification :

- Classification supervisée : cellules étiquetées (base d'exemples de chaque espèce)
- Classification non supervisée (clustering) : cellules non étiquetées
- Classification semi-supervisée : clustering avec apports d'informations *a priori*

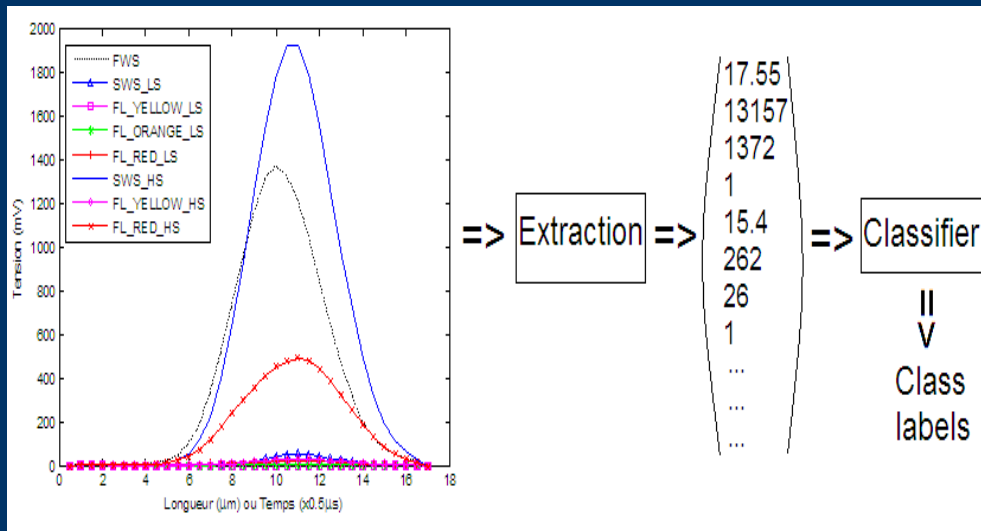


La classification supervisée

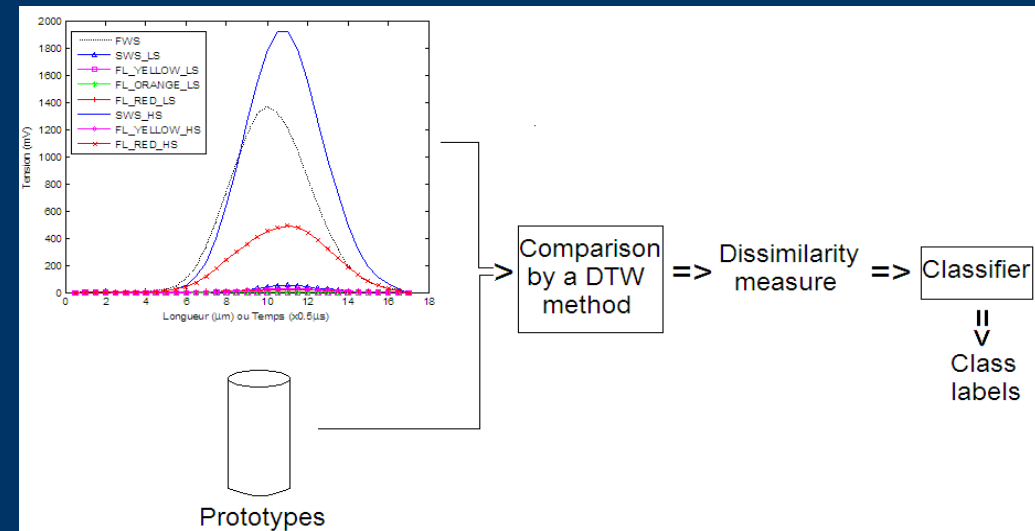
Classification supervisée

2 approches :

- Une approche basée attributs : extraction de paramètres sur les courbes cytométriques
- Une approche basée signaux : mesure globale de dissimilarité entre les formes des signaux



Approche basée attributs

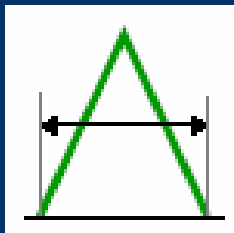


Approche basée signaux

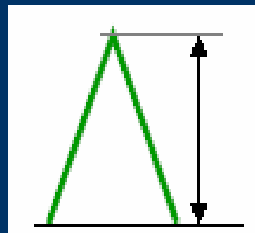
Classification supervisée

Approche basée attributs

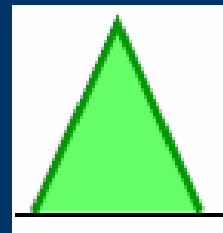
- Approche classique : chaque cellule phytoplanctonique est décrite par des attributs numériques
- 4 attributs extraits de chaque signal : longueur, hauteur, intégrale et le nombre de pics



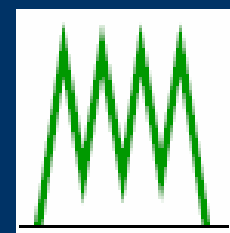
Longueur



Hauteur



Intégrale



Nombre de pics

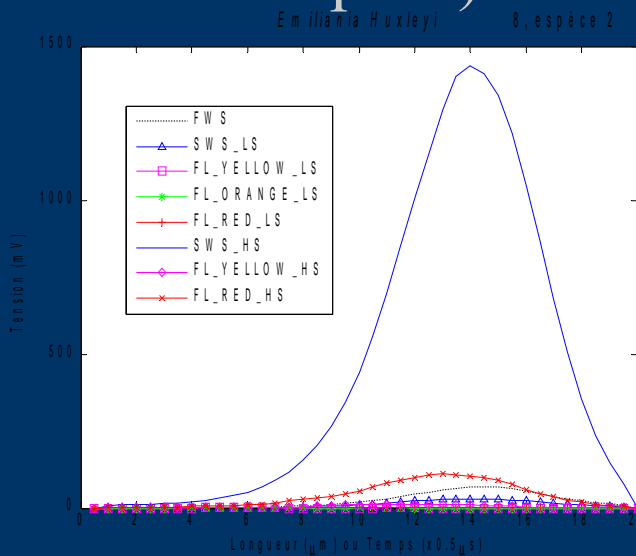
=> Chaque cellule décrite par 32 attributs extraits des 8 signaux

=> Application de classifieurs sur les points en 32-dimensions

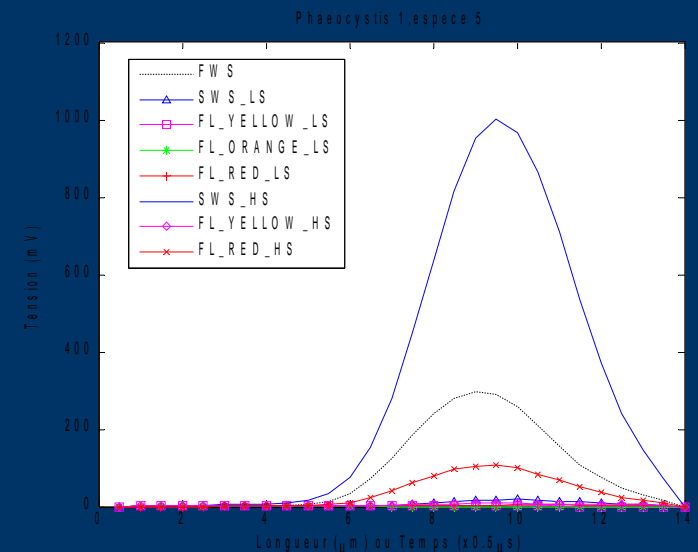
Classification supervisée

Approche basée signaux

- Approche basée sur la dissimilarité des formes des signaux
- Comparaison des signaux temporels décrivant un individu par rapport à des profils de référence
- Chaque comparaison donne une valeur de dissimilarité (associée à un label d'espèce)



\Rightarrow Valeur de \leq
dissimilarité ???



\Rightarrow Application de classifieurs sur les vecteurs de dissimilarités

Classification supervisée

Approche basée signaux

- Appariement élastique conjoint (DTW)

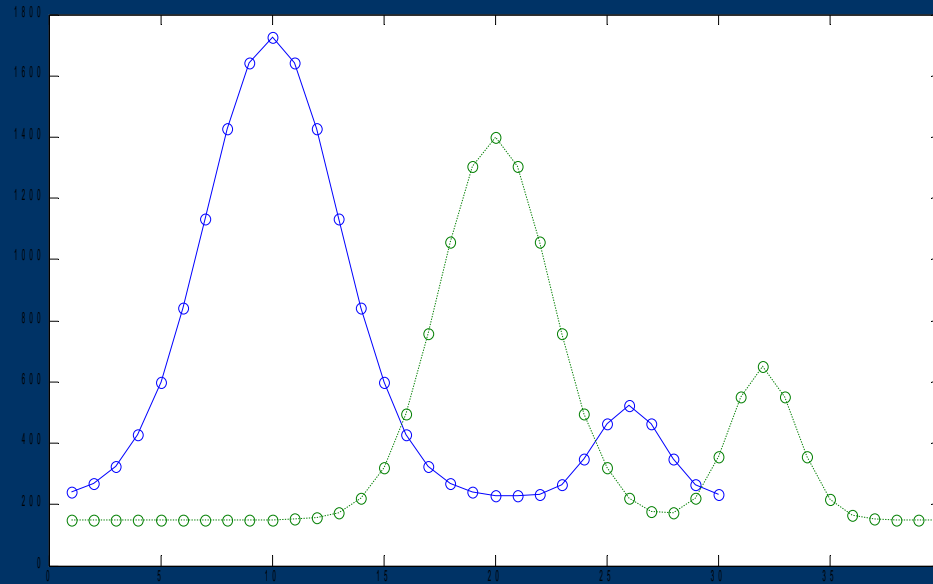
- Faire correspondre au mieux les points de l'un et l'autre signal
- Les signaux peuvent être de durées et de fréquences d'échantillonnage différentes
- Mesure traduit la quantité de distorsion géométrique nécessaire à la superposition des deux courbes, en tolérant des déformations temporelles locales.

- Extension de l'algorithme

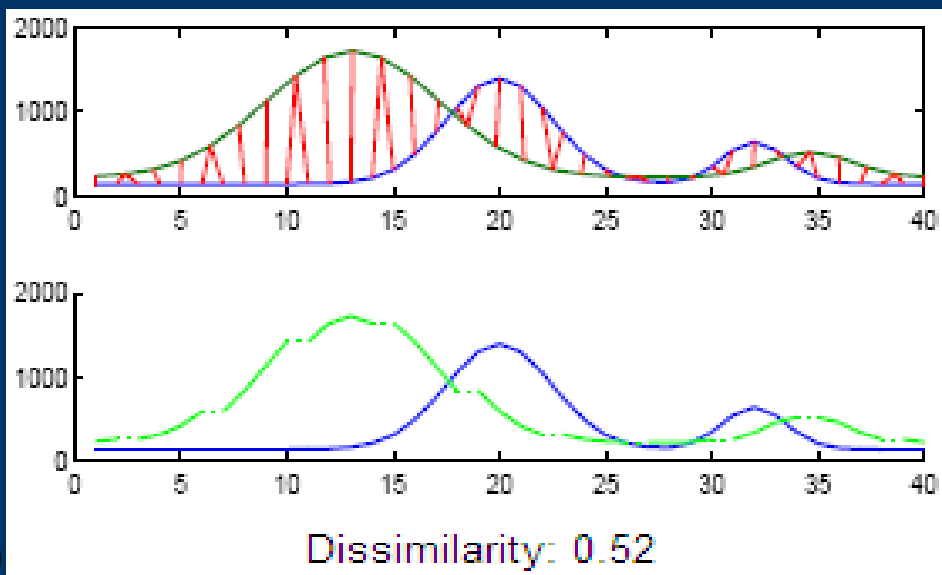
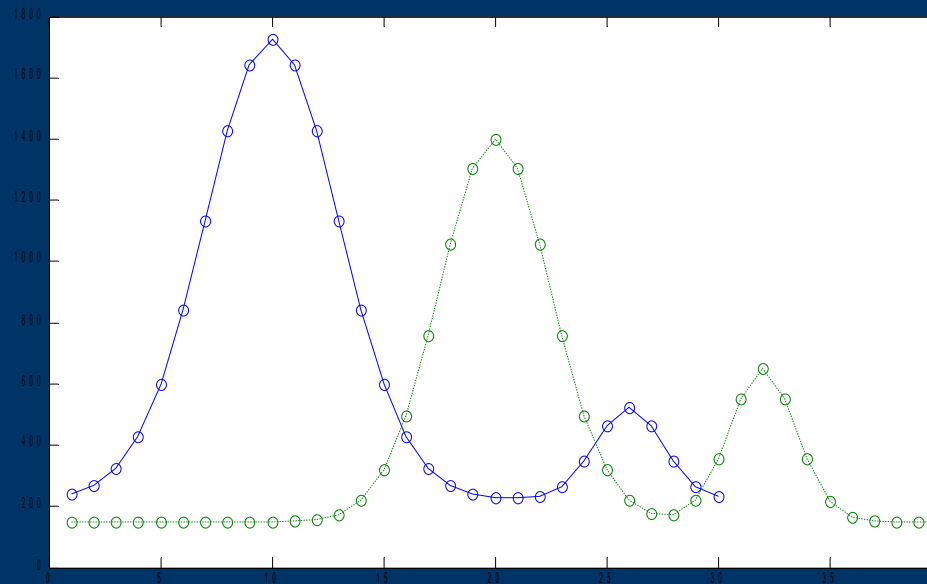
- Mesure de dissimilarité (dans $[0,1]$) entre 2 séquences de mesures temporelles
- Dissimilarité = moyenne des distances entre points appariés

⇒ Algorithme souple car il permet d'apparier des points décalés dans le temps.

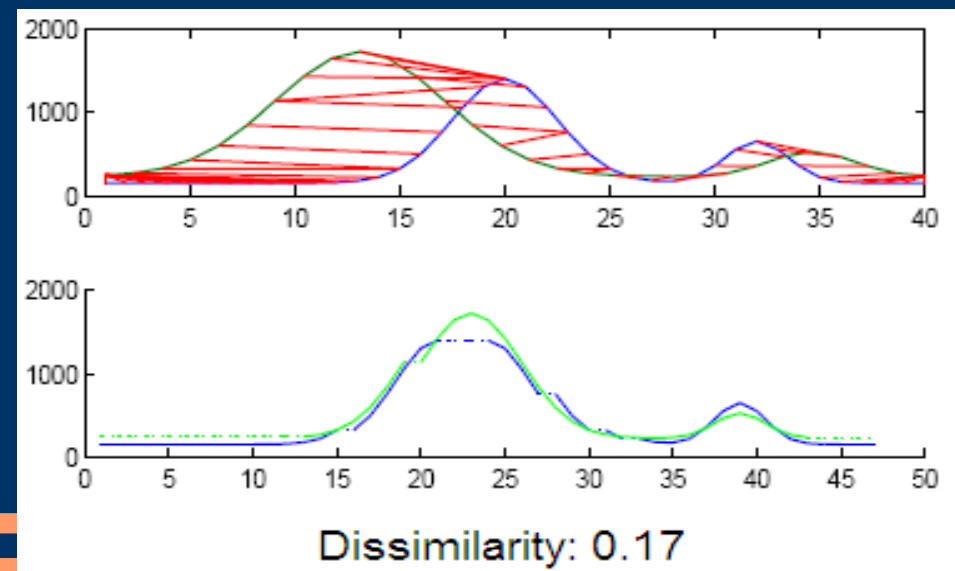
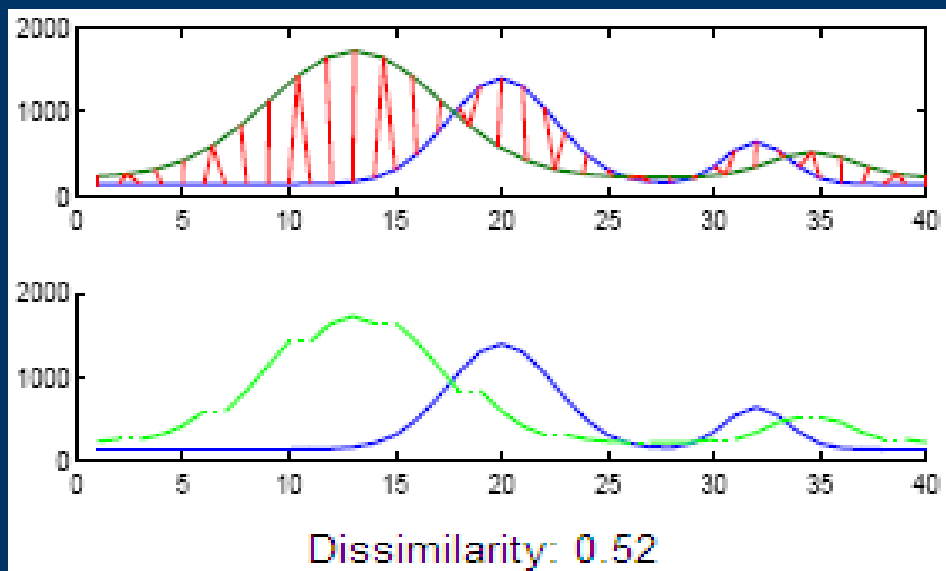
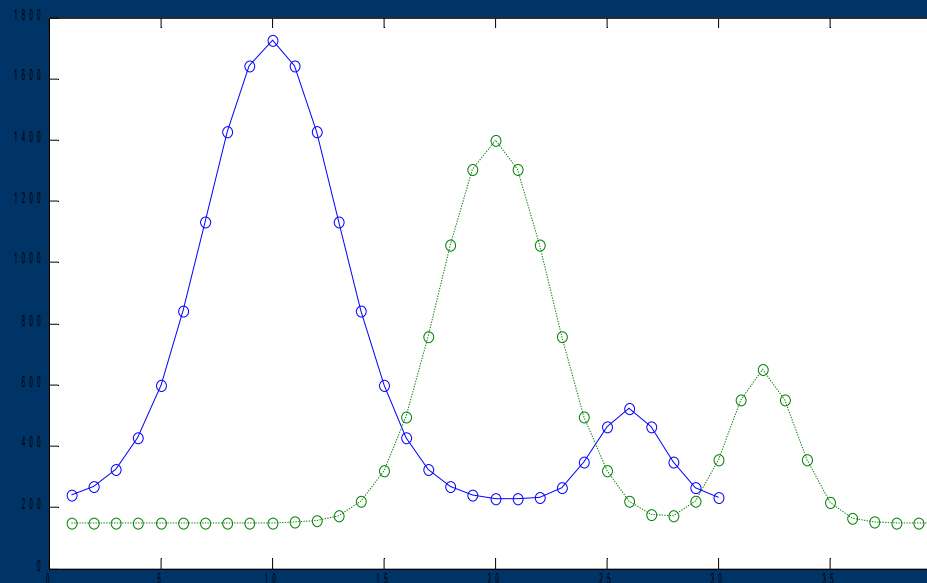
Classification supervisée



Classification supervisée



Classification supervisée



Classification supervisée

Données expérimentales

- Cellules provenant d'un unique échantillon de culture
- 7 espèces phytoplanctoniques distinctes : *Chaetoceros socialis*, *Emiliana Huxleyi*, *Lauderia annulata*, *Leptocylindrus minimus*, *Phaeocystis globosa*, *Skeletonema costatum* et *Thalassiosira rotula*
- Chaque espèce représentée par 100 cellules phytoplanctoniques (étiquetées par les biologistes du LOG – UMR LOG 8187)

=> 7 espèces x 100 cellules = 700 cellules

Classification supervisée

Classification basée attributs :

Training Fold	Fold 1	Fold 2	Fold 3	Fold 4	Mean	Std
MLP	96,9	94,8	96	94,8	95,6	1,1
1-NN	93,7	90,2	93,7	92,5	92,5	1,7
SVM2	95	91,2	90	93,9	92,5	2,4
SVM1	90	87,4	91	92,5	90,2	2,2
NBAYES	88,9	89,1	88	86,7	88,2	1,1
TREE	88	86,3	88,8	86,9	87,5	1,1

Classification basée signaux :

Training Fold	Fold 1	Fold 2	Fold 3	Fold 4	Mean	Std
MLP	98,2	97,3	97,3	96,7	97,3	0,7
1-NN	98,2	95,4	96,1	97,1	96,7	1,3
SVM1	98,8	95,6	95,6	96,1	96,5	1,6
SVM2	92,3	93,5	93,3	92,9	93	0,6
NBAYES	89,9	89,9	90,8	88	89,6	1,2
TREE	88,7	88,2	90,4	88,7	89	1

La classification non supervisée

Classification non supervisée

Taux de bonne reconnaissance élevé en supervisé

- => Information contenue dans les attributs est suffisante pour discriminer les espèces
- => Même constat pour le non supervisé ???

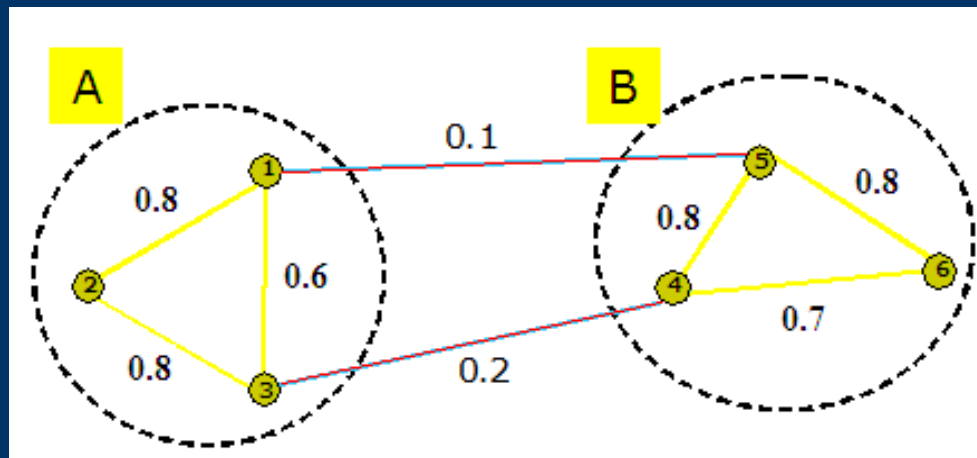
Utilisation d'algorithmes non supervisés

- Expérimentations sur des échantillons d'eau provenant du milieu naturel
- Détection de nouvelle espèce

Classification non supervisée

Spectral Clustering

- Problème de clustering = problème de segmentation de graphe



Avec : noeuds = points

liaisons entre 2 noeuds = similarités entre 2 noeuds

- Pas de suppositions sur la forme des clusters
- Facilité d'implémentation

Classification non supervisée

Spectral Clustering

- Pour une base de données composée de n points :
 - A chaque paire de points (x_i, x_j) est associée une valeur de similarité $W_{ij} \Rightarrow$ matrice de similarités W
 - W_{ij} de type Gaussien définie par :

$$W_{ij} = \exp\left(\frac{-1}{2\sigma^2} \cdot d^2(x_i, x_j)\right)$$

- Partitionner le graphe en 2 clusters \rightarrow fonction NCut
 \Rightarrow trouver 2 clusters A et B tels que :
inter-connexions minimisées et intra-connexions maximisées

$$Cut(A, B) = \sum_{i \in A, j \in B} W_{ij}$$

\rightarrow

$$Ncut(A, B) = \frac{Cut(A, B)}{Vol(A)} + \frac{Cut(B, A)}{Vol(B)}$$

Avec :

$$Vol(A) = \sum_{i \in A} \sum_{j=1}^n W_{ij}$$

et

$$Vol(B) = \sum_{i \in B} \sum_{j=1}^n W_{ij}$$

Classification non supervisée

Spectral Clustering

$$d_i = \sum_{j=1}^n W_{ij} \Rightarrow D : \text{matrice de degrés (avec } d_i \text{ sur diagonale et 0 ailleurs)}$$

$$Ncut(A, B) = \frac{Cut(A, B)}{Vol(A)} + \frac{Cut(B, A)}{Vol(B)}$$

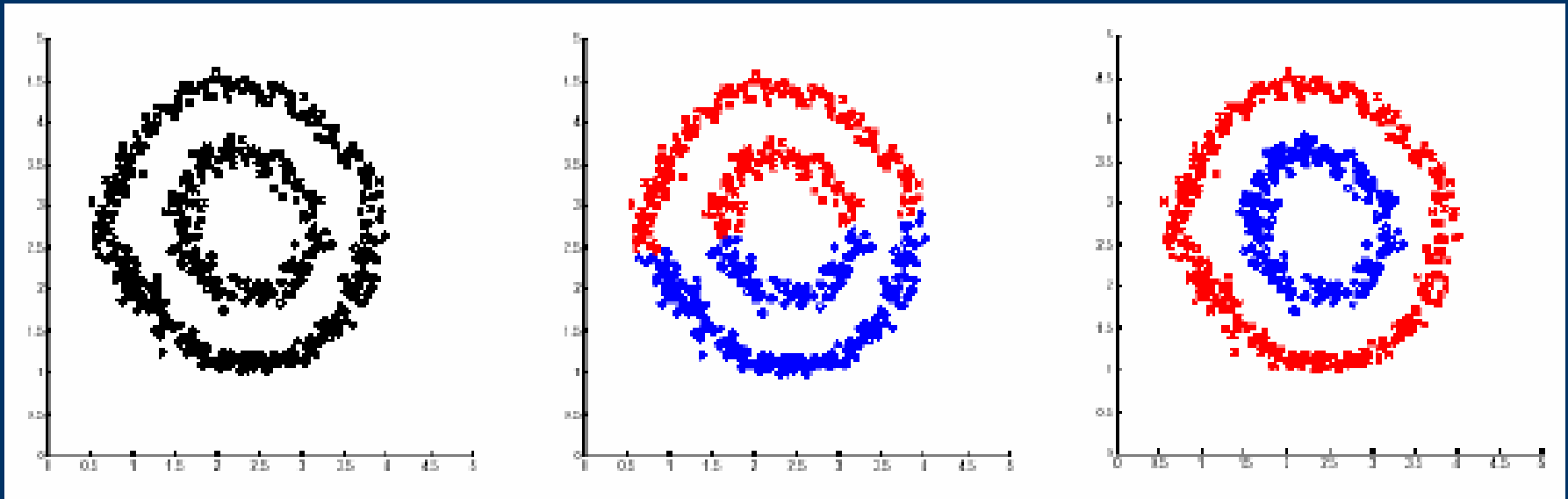
- La solution minimisant la fonction NCut est obtenue en résolvant l'équation

$$(D - W)y = \lambda Dy \rightarrow L = D^{(-\frac{1}{2})} (D - W) D^{(-\frac{1}{2})}$$

- Application des K-means sur les k vecteurs propres de L associés aux k plus grandes valeurs propres

Classification non supervisée

Spectral Clustering



Points originaux

K-means

Spectral Clustering

Classification non supervisée

Résultats

Méthode basée attributs

K-means 62.8 % de classification correcte

EM 62.6 % de classification correcte

Spectral Clustering 64.7 % de classification correcte

La classification semi-supervisée

Classification semi-supervisée

- Etiquetage de la totalité des points d'un jeu de données très long et parfois complexe (taille des échantillons, confusions)
 - Classification supervisée
 - 1 hl – MLP : 97.3 %
 - Classification non supervisée
 - Spectral Clustering : 64.7 %
- => Classification semi-supervisée
- Possibilité d'étiqueter une partie de ces points (classes exactes connues pour une partie des points) afin d'apporter des connaissances *a priori*
 - Utilisation de l'algorithme de S. Shortreed dans le cas semi-supervisé

Classification semi-supervisée

Présentation de l'algorithme semi-supervisé :

- Utilisation d'un jeu de données dont certains points sont étiquetés (au moins 1 par cluster)
- But : apprentissage d'une métrique pour donner du poids aux attributs discriminants

Constrained Cut :

$$\begin{aligned} \text{CCut}(\mathbf{C}) &= \text{Cut}(\mathbf{C}) + \rho \sum_{\ell=1, \dots, n} \mathbf{I}[\ell \text{ mal classés}] \\ &= \sum_{k=1, \dots, K} \left(\sum_{i \in C_k} \sum_{j \in C_k} S_{ij} \right) + \rho \sum_{\ell=1, \dots, n} \mathbf{I}[\ell \text{ mal classés}] \end{aligned}$$

ρ : poids positif associé au terme de pénalisation

- si $\rho=0 \rightarrow \text{CCut} = \text{Cut}$
- si $\rho=\infty \rightarrow$ Sélection des clusterings classifiant correctement les points
- Choix de ρ : nombre de données faible : $\rho=1$
nombre de données important : $\rho=1/n$

Classification semi-supervisée

Clustering = segmentation de graphe

- Extension de cette idée pour application au CCut
- Ajout de K nœuds (chacun représentant un label pour chaque cluster)
- Chaque nœud est relié aux points possédant un label de classe similaire (par exemple : nœud 1 est relié aux points faisant partie du cluster 1)

Matrice de similarités :

- $G^{(1)}$: graphe original et $G^{(2)}$: graphe augmenté
- $S^{(1)}$: matrice de similarités originale et $S^{(2)}$: matrice de similarités augmentée
- i, j indices des points originaux dans $G^{(2)}$
- $(n+1), \dots, (n+k)$ indices des nœuds supplémentaires dans $G^{(2)}$

$$S_{ij}^{(2)} = \begin{cases} S_{ij}^{(1)} & \text{si } i, j \leq n \\ \rho/2 & \text{si } i > n, j \leq n \text{ et } j \text{ est dans } C_i \\ \rho/2 & \text{si } j > n, i \leq n \text{ et } i \text{ est dans } C_j \\ 1 & \text{si } i=j > n \\ 0 & \text{sinon} \end{cases}$$

Classification semi-supervisée

					C1	C2	
S_{ij}					0	$\rho/2$	
					$\rho/2$	0	
					0	$\rho/2$	
					$\rho/2$	0	
					0	0	
C1	0	$\rho/2$	0	$\rho/2$	0	1	0
C2	$\rho/2$	0	$\rho/2$	0	0	0	1

Définition du CMNCut (Constrained Multi-Way Normalize Cut)

Matrice de similarités augmentée permet de faciliter le calcul du critère de coupe contraint $CCut(C)$ en se ramenant à un critère de coupe $Cut(C)$

$$\Rightarrow CCut(S^{(1)}) = Cut(S^{(2)})$$

$$\Rightarrow CMNCut(G^{(1)}, C) = MNCut(G^{(2)}, C)$$

$$CMNCut(C) = \sum_{k=1}^K \sum_{k' \neq k} \frac{\sum_{i \in C_k, j \in C_{k'}} S_{ij} + \frac{\rho}{2} m_{C_{k'}, C_k} + \frac{\rho}{2} m_{C_k, C_{k'}}}{Vol(C_k) + \frac{\rho}{2} n_{C_k} + \frac{\rho}{2} (n_{C_k} - m_{+, C_k}) + \frac{\rho}{2} m_{+, C_k}}$$

Classification semi-supervisée

Extension de l'algorithme (1)

- S. Shortreed → utilisation de labels partiels
- Nous → utilisation de contraintes partielles (plus générales)
 - Must-Link pour $(x_i, x_j) = ML_{i-j} = (1, 1)$
 - Cannot-Link pour $(x_i, x_j) = CNL_{i-j} = (1, 0)$
 - Données sans connaissance *a priori* = 0.5

		ML2-4 CNL3-5						
S_{ij}							0.5	0.5
							1	0.5
							0.5	1
							1	0.5
							0.5	0
ML2-4	0.5	1	0.5	1	0.5	1	0	
CNL3-5	0.5	0.5	1	0.5	0	0	1	

Classification semi-supervisée

Extension de l'algorithme (1)

- Problème 1 :

Plus il y a de contraintes, plus on ajoute de nœuds
=> matrice de similarités de taille importante
=> Temps de calcul longs

- Problème 2 :

Problème au niveau du critère de $CMNCut(C)$
→ si termes de pénalités nulles :

$$CMNCut(C) = \sum_{k=1}^K \sum_{k' \neq k} \frac{\sum_{i \in C_k, j \in C_{k'}} S_{ij} + \frac{\rho}{2} m_{C_{k'}, C_k} + \frac{\rho}{2} m_{C_k, C_{k'}}}{Vol(C_k) + \frac{\rho}{2} n_{C_k} + \frac{\rho}{2} (n_{C_k} - m_{+, C_k}) + \frac{\rho}{2} m_{+, C_k}}$$

$$CMNCut(G, C) = \sum_{k=1}^K \sum_{k' \neq k} \frac{\sum_{i \in C_k, j \in C_{k'}} S_{ij}}{Vol(C_k) + \rho n_{C_k}}$$

Or $MNCut(G, C) = \sum_{k=1}^K \sum_{k' \neq k} \frac{\sum_{i \in C_k, j \in C_{k'}} S_{ij}}{Vol(C_k)}$

Donc : $CMNCut(G, C) \neq MNCut(G, C)$

Classification semi-supervisée

Extension de l'algorithme (2)

- Pas d'ajout de nœuds dans la matrice de similarités
- Utilisation de contraintes partielles

Matrice de similarités :

- Cannot-Link (CNL) pour (x_i, x_j)
 - $S_{ij} = S_{ij} --$ et $S_{ji} = S_{ji} --$
 - $S_{ii} = S_{ii} ++$ et $S_{jj} = S_{jj} ++$
- Must-Link (ML) pour (x_i, x_j)
 - $S_{ij} = S_{ij} ++$ et $S_{ji} = S_{ji} ++$
 - $S_{ii} = S_{ii} --$ et $S_{jj} = S_{jj} --$

=> Les premiers tests montrent que cette méthode permet d'obtenir une partition pour laquelle les contraintes sont respectées

Conclusion

En supervisée :

- 2 approches :
 - Approche basée attributs
 - Approche basée signaux
- Taux de bonne reconnaissance élevé
 - => Information contenu dans les attributs (et les signaux) est suffisante pour permettre de discriminer les différentes espèces (sans sélection d'attributs)
 - => Caractère discriminant des mesures utilisées

En semi-supervisée :

- Apport d'informations *a priori*
- Actuellement :
 - perfectionnement d'algorithmes semi-supervisés
 - intégration d'autres types d'informations provenant de l'expert

Perspectives

- Recherche de prototypes représentatifs de chaque classe pour réduction de la dimension
- Intégration de degrés de confiance pour les cellules étiquetées
- Estimation du nombre de classes
- Expérimentations :
 - Classification semi-supervisée sur des échantillons d'eau provenant de culture et du milieu naturel
 - Bases de données plus importantes
 - Plus grande diversité d'espèces
 - Etiquetage plus « propre » (camera)

Caillault, E., Hébert, P-A., Wacquet, G.: *Dissimilarity-based Classification of Multidimensional Signals by Conjoint Elastic Matching: Application to Phytoplanktonic Species Recognition*. In: 11th International Conference on Engineering Applications of Neural Networks. (2009)

Membres de l'équipe Extraction de l'Information et Apprentissage :

Hamad, D., Caillault, E., Hébert P-A.

Classification semi-supervisée

2 mesures sur le clustering :

– Mesure de qualité du clustering

- Borne inférieure théorique du MNCut :

$$\text{MNCut} \geq K - \sum_{k=1, \dots, K} \lambda_k$$

où $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq -1$ sont les valeurs propres de la matrice Laplacienne

- Définition du gap :

Différence entre MNCut observé et la borne inférieure théorique, donc :

$$\text{gap} = \text{MNCut}(C) - (K - \sum_{k=1, \dots, K} \lambda_k)$$

Si Gap = 0 alors MNCut minimisé donc Clustering optimal

– Mesure de stabilité du clustering

Définition : un clustering est dit stable si tous les clustering jugés « bons » par le terme de qualité (le gap), sont proches de lui

$$\text{Gap propre} : \Delta_K = \lambda_K - \lambda_{K+1}$$

Classification semi-supervisée

Fonction coût utilisée :

Utilisation du gap (terme de qualité) régulé par le gap propre (terme de stabilité) avec α représentant le paramètre de régulation

$$J(\theta, C) = \underset{\text{Qualité}}{\text{gap}(\theta, C)} - \underset{\text{Stabilité}}{\alpha \Delta^2_K(\theta)}$$

$$J(\theta, C) = \text{MNCut}(C, \theta) - (K - \sum_{k=1, \dots, K} \lambda_k(\theta)) - \alpha \Delta^2_K(\theta)$$

Minimisation de J :

- Objectif : apprentissage des paramètres qui minimisent J
- Méthode utilisée : descente du gradient