FODOMUST: une plateforme pour la fouille de données multistratégie multitemporelle

Pierre Gançarski*, Abdoul-djawadou Salaou*

*ICube - Université de Strasbourg
{gancarski, adsalaou}@unistra.fr
http://www.icube.fr

Résumé. La plateforme FODOMUST [1] est une implantation concrète des méthodes, librairies et interfaces proposées au sein d'ICube. Elle intègre une version multisource de la méthode de classification collaborative multistratégie SAMARAH. Elle propose aussi un ensemble d'algorithmes de segmentation soit propres à ICUBE soit faisant appel à l'OTB. Enfin, trois interfaces dédiées chacune à un type de données différent permettent une interaction avec l'utilisateur. Sa principale originalité est qu'elle permet la classification, basée sur DTW (Dynamic Time Warping) de données temporelles symboliques ou numériques et de séries temporelles d'images

1 Introduction

La mise à disposition par les satellites européens de la constellation Sentinel, d'une masse considérable de données gratuites d'observation de la Terre offre une opportunité de détecter des changements lents, rapides, abruptes et/ou cycliques qui affectent les territoires. Il est pour cela nécessaire de mettre en oeuvre des méthodes de découverte des classes d'évolution des objets géographiques afin d'exploiter au mieux ce flux quasi continu d'images pour une analyse des dynamiques territoriales. En effet, une caractéristique du système Sentinel est sa pérennité autorisant des études à long terme. Or, comme dans le cadre plus général de l'augmentation massives de données et de leur complexité, connue sous le terme de « Big Data », le manque de connaissances sur les évolutions (peu ou pas d'exemple d'évolution, formalisation incomplète des classes d'évolution, ...) impose le développement d'outils d'analyse non supervisée. Dans ce contexte, le laboratoire ICube propose une solution technologique innovante pour la classification non supervisée de séries temporelles de données complexes.

2 La plateforme FODOMUST

La plateforme FODOMUST[1] est une implantation concrète des méthodes, des librairies et interfaces proposées par l'équipe Sciences des Données et Connaissances (SCD - ICube).

1. http://icube-bfo.unistra.fr/fr/index.php/Plateformes

Fouille de **Do**nnées **Mu**lti**S**tratégie multi**T**emporelle

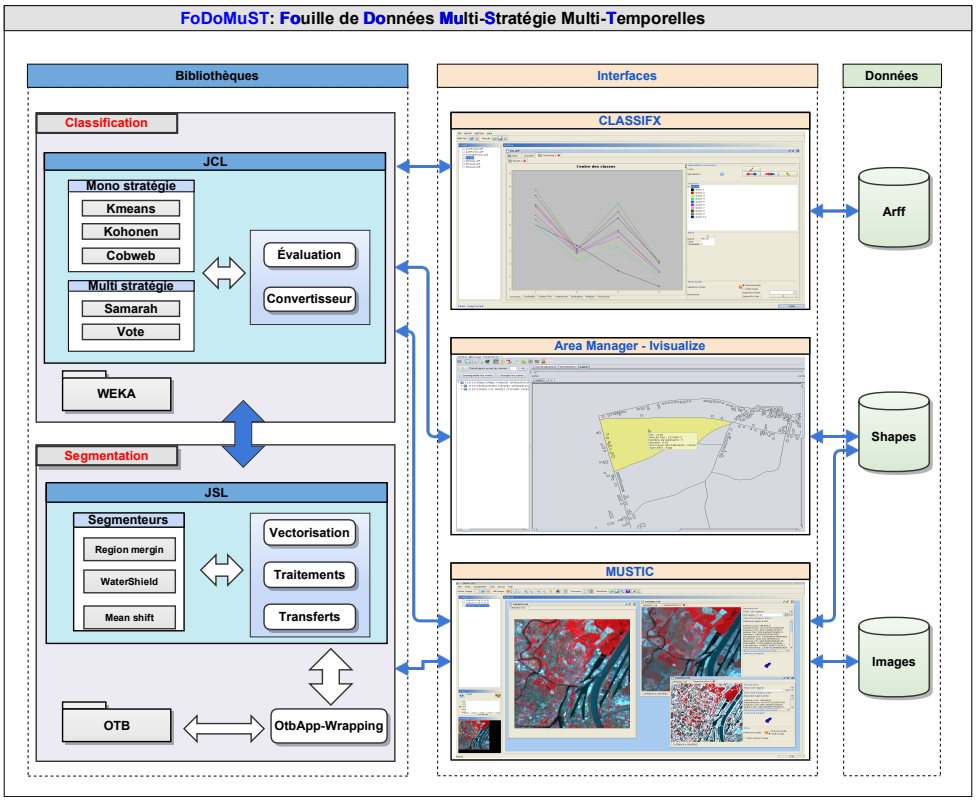

FIG. 1 – *Architecture générale de la plateforme* FODOMUST

Elle intègre un ensemble de classifieurs classiques ou flous tels que Kmeans, Cobweb, ou Kohonen en code propre. Elle intègre surtout la méthode de classification multistratégie SAMARAH (cf Section 3.1.3). Mais sa *principale originalité* est qu'elle permet la classification, basée sur DTW (Dynamic Time Warping) de données temporelles symboliques ou numériques et de séries temporelles d'images. Trois interfaces dédiées chacune à un type de données différent permettent une interaction avec l'utilisateur.

La plateforme (Fig. 1) est donc composée de deux librairies développées par l'équipe :

- JCL est une bibliothèque de classifieurs
- JSL est une bibliothèque d'algorithmes de segmentation soit propres à ICUBE soit proposés par l'Orfeo Tool Box (OTB) [2].

et de trois interfaces dédiées chacune à une famille d'applications différentes :

- CLASSIFX dédiée à l'analyse et la classification de données (temporelles) ARFF
- MUSTIC dédiée à l'analyse et la classification (de séries temporelles) d'images
- IVISUALIZE dédiée à la classification de séries temporelles de données géographiques

2. https ://www.orfeo-toolbox.org/

3 Classification (de séries) d'images

3.1 JCL : Java Clustering Library

3.1.1 Données

Quel que soit le type initial des données, celles-ci sont transformées en un modèle attributs-valeurs propre à JCL. Les trois interfaces de FoDoMuST intègrent les mécanismes nécessaires à cette traduction pour les types de données simples - entier, réel, symbolique - ou construits - tableaux, structures - au format ARFF (CLASSIFX), pour les images aux formats classiques TIF ou BMP (MUSTIC) mais aussi pour des séries temporelles de données ARFF (CLASSIFX), d'images (MUSTIC) ou de données géographiques (IVISUALIZE).

3.1.2 Distance inter-objets et moyenne

Afin de pourvoir appliquer les algorithmes de classification basés sur une distance, une telle mesure doit être définie pour chaque type d'attributs. Ainsi, une fonction de distance peut être choisie, voire (re)définie. Par exemple, pour les types :

- simples numériques et structurés (tableau ou structure) : la distance euclidienne, éventuellement pondérée, est utilisée
- symboliques : une matrice de similarité doit être définie (une interface dédiée à la gestion de ces matrices est intégrée dans les différentes interfaces)
- temporels : l'utilisateur peut choisir entre la distance euclidienne et DTW.

Une méthode de calcul de moyenne doit généralement être définie, par exemple pour Kmeans. Pour les distances euclidiennes, la moyenne euclidienne est utilisée. Pour DTW, la moyenne DBA (DTW Barycenter Averaging) est implantée et est utilisable (Petitjean et al. (2011)).

3.1.3 SAMARAH

Une version multisource (images de même résolution) de la méthode SAMARAH proposée dans (Gançarski et Wemmert (2007)) est implantée et est applicable à tout type de données. Cette méthode consiste à faire collaborer des méthodes de classification différentes afin qu'elles raffinent mutuellement leurs résultats jusqu'à obtenir des résultats "similaires" de qualité, ces résultats pourront alors être unifiés (Fig. 2).

Le processus consiste en trois étapes principales :

- Classifications initiales
- Collaboration pour un raffinement itératif :
  - Détection de conflits entre les classifications proposées par les différentes méthodes
  - Résolution locale des conflits par modifications de clusters (scission, fusion, ...)
  - Évaluation et prise en compte globale des résolutions locales
- Unification par un algorithme de vote adapté

La figure 3 présente le panel de configuration de SAMARAH via CLASSIFX

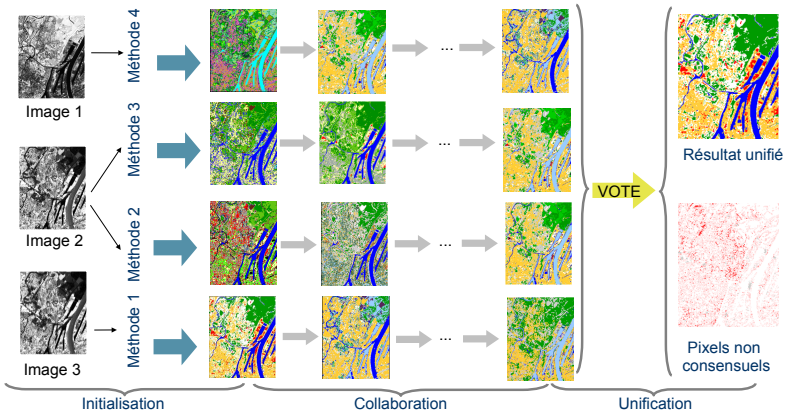

FIG. 2 – SAMARAH *: Classification collaborative multistratégie*

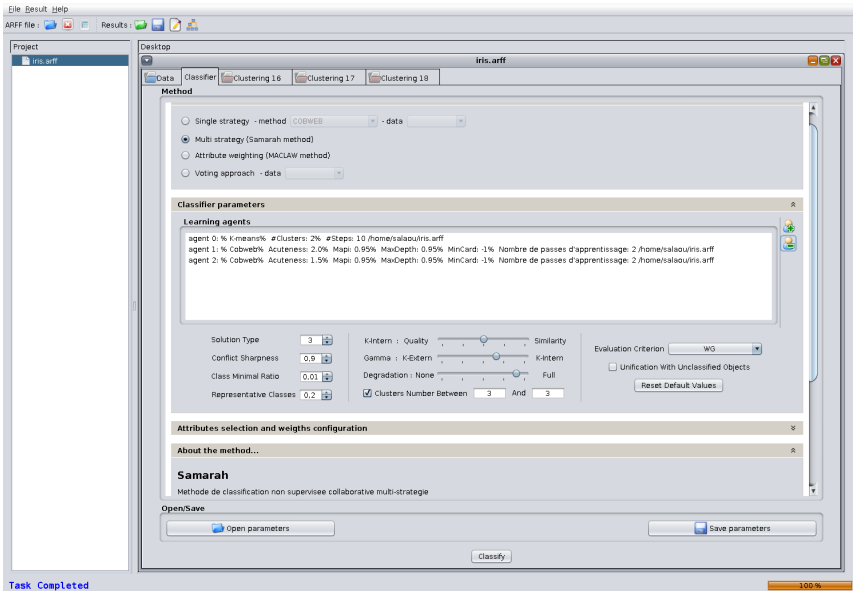

FIG. 3 – CLASSIFX *: Une interface pour la classification collaborative multistratégie*

3.2 MUSTIC

L'analyse d'images de télédétection à des résolutions [3] comprises entre 5 et 500 m se fait généralement au niveau des pixels à partir des valeurs radiométriques de ceux-ci, éventuellement complétées par des informations spatiales ou texturales. Avec les images à très haute

3. À une résolution de N mètres, un pixel couvre une zone de (N $\times$N) m^2

résolution spatiale (THR) proche du mètre, les méthodes par pixels ont montré leurs limites : les objets d'intérêt doivent être reconstruits avant analyse. Les méthodes *orientées objets (ou régions)* segmentent l'image en zones homogènes puis les caractérisent (forme, texture…) avant de les classifier en utilisant éventuellement des connaissances du domaine (Fig. 4). Enfin, la complexité de la scène dans une image THR ne permet plus de définir exhaustivement les classes présentes voire même de donner un nombre suffisant d'exemples. De fait, ces méthodes sont principalement non supervisées.

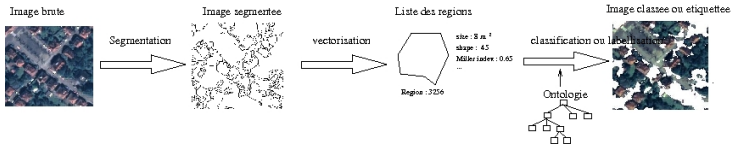

FIG. 4 – *Approche orientée régions pour l'analyse d'images*

La figure 5 présente l'interface MUSTIC avec, sous le bandeau regroupant les principaux outils de traitement d'images (contraste, découpe, etc.), en partant de la gauche vers la droite :

- un panel permettant la gestion des fichiers,
- un panel présentant les résultats d'une classification non supervisée de l'image,
- un panel montrant d'une part, une segmentation effectuée sur l'image et d'autre part, un extrait des caractéristiques associées à chacun des segments construits. Cet ensemble de régions peut alors être directement classifié ("Configurer un classifieur").

4 Conclusion

La version actuelle de FODOMUST démontre la pertinence et le bénéfice à utiliser un processus collaboratif pour la classification de séries temporelles d'images (Petitjean et al. (2012)). Les méthodes d'apprentissage supervisé font l'hypothèse que les données d'apprentissage décrivent de manière suffisante et complète les classes auxquelles elles sont rattachées. Dans le cas de l'analyse temporelle, le manque d'exemple d'évolution et de formalisation incomplète des classes d'évolution rend cette hypothèse peu réaliste. Les méthodes d'apprentissage actif sont une solution à ce problème bien que n'ayant encore été que très peu appliquées dans ce cadre. La figure 6 présente une proposition d'extension pour un apprentissage collaboratif actif distribué (ACOLLAD) facilement déployable sur des grilles et clusters de calcul.

Références

Gançarski, P. et C. Wemmert (2007). Collaborative multi-step mono-level multi-strategy classification. *Multimedia Tools and Applications 35*(1), 1–27.

Petitjean, A., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping with applications to clustering. *Pattern Recognition 44*, 678–693.

Petitjean, F., C. Kurtz, N. Passat, et P. Gançarski (2012). Spatio-temporal reasoning for the classification of satelitte image time series. *Pattern Recognition Letters 33*(13), 1805–1815.

Fouille de **Do**nnées **Mu**lti**S**tratégie multi**T**emporelle

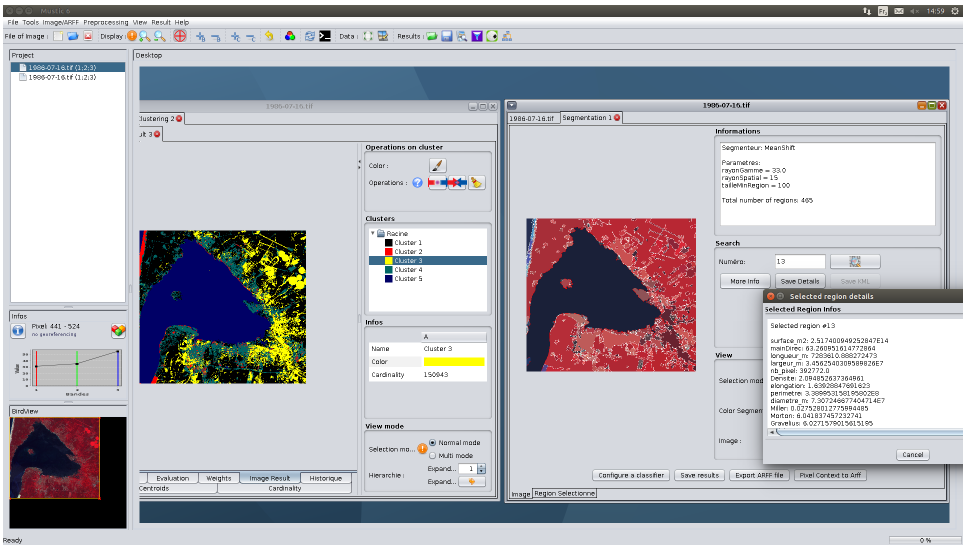

FIG. 5 – MUSTIC *: Une interface pour l'analyse d'images orientée régions*

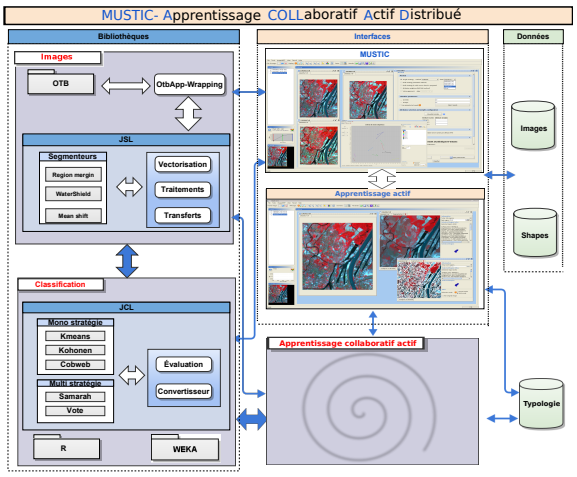

FIG. 6 – ACOLLAD *: Vers une classification collaborative interactive*

Summary

The platform FODOMUST is a concrete implementation of a multi-source version of the collaborative classification method SAMARAH developed within ICube. Its main innovation is that it enables the classification, based on DTW (Dynamic Time Warping), of temporal symbolic or numerical data and time series of images