FODOMUST - Une plateforme de clustering collaboratif sous contraintes incrémental de séries temporelles

Pierre Gançarski, Baptiste Lafabregue, Abdoul-djawadou Salaou, Harrison Vernier

ICube, CNRS - Université de Strasbourg
{gancarski, lafabregue, adsalaou, harrison.vernier}@unistra.fr

Résumé. La plateforme FODOMUST [1] est une implantation concrète des méthodes, librairies et interfaces dédiées au clustering de données complexes proposées au sein d'ICube. Elle intègre une version multisource de la méthode de clustering collaboratif multistratégie SAMARAH. Sa principale originalité est qu'elle offre à l'utilisateur les méthodes et les interfaces permettant le clustering sous contraintes incrémental de données temporelles symboliques ou numériques. L'interface MULTICUBE dédiée à l'analyse de séries temporelles offre aussi un accès à un ensemble d'algorithmes de segmentation soit propres à ICube soit faisant appel à l'OTB diffusée par le CNES.

1 Introduction

Les méthodes d'apprentissage supervisé classiquement utilisées font l'hypothèse que les données d'apprentissage décrivent de manière suffisante et complète les classes auxquelles elles sont rattachées. En d'autres termes, ces méthodes nécessitent que les classes recherchées soient parfaitement connues et définies et que l'expert soit capable de fournir un jeu de données d'apprentissage suffisant tant en nombre qu'en qualité. Or, face aux flux quasi-continus de nouvelles données issues de capteurs de plus en plus nombreux et variés, cette hypothèse n'est plus réaliste dans bien des domaines. En effet, la révolution technologique de la haute fréquence d'acquisition y est encore trop récente pour que les connaissances thématiques se soient adaptées. Ainsi, bien souvent il n'existe pas de typologies (ou nomenclatures) des changements réellement utilisables pour ce type d'analyse supervisée et donc de données d'apprentissage de qualité associées. Pour pallier ce manque, des approches récentes basées sur un apprentissage non supervisé font l'hypothèse que même en l'absence de connaissances formalisées, celles-ci peuvent néanmoins être représentées en partie à travers des contraintes (de comparaison, d'étiquetage ou de structure) opérables (Basu et al., 2008; Dao et al., 2017; Lampert et al., 2018). Ces contraintes a priori plus aisées à définir, peuvent alors être utilisées pour guider le processus de clustering afin de produire des clusters plus proches des « intuitions » de l'expert c'est-à-dire des classes thématiques potentielles. Ainsi, la méthode collaborative SAMARAH (Gançarski et Wemmert, 2007; Forestier et al., 2010) développée par ICube a été complétée par une prise en compte incrémentale de contraintes (Lampert et al., 2019).

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

Cet article présente l'architecture globale de la plateforme FODOMUST (**Fo**uille de **Do**nnées **Mu**lti**S**tratégie multi**T**emporelle) implantant la méthode SAMARAH (Section 2) ainsi que les fondements scientifiques de cette méthode (Section 3). La plateforme et les méthodes présentées sont génériques et peuvent s'appliquer à tout type de données mono ou multi-dates (symboliques, numériques, structurées ...). Néanmoins pour mieux illustrer nos propos nous nous placerons dans le domaine de la télédétection. En effet, avec le lancement et la mise en production des satellites européens de la constellation Sentinel ou franco-israélien Venμs, des données satellitaires arrivent maintenant en flux quasi continus. Ces données temporelles massives devraient permettre des avancées fortes dans différentes disciplines des Sciences de la Terre et de l'Environnement (Flamary et al., 2018) pour l'étude et la modélisation des phénomènes complexes (dynamiques agricoles ou urbaines, la déforestation, actions anthropiques sur la biodiversité...). Nous présenterons donc plus en détail l'interface MULTICUBE dédiée à l'analyse de séries temporelles d'images (Section 2.3).

2 La plateforme FODOMUST

La plateforme FODOMUST (Fig. 1) est une implantation concrète des méthodes, librairies et interfaces proposées par l'équipe Sciences des Données et Connaissances (SDC) d'ICube.

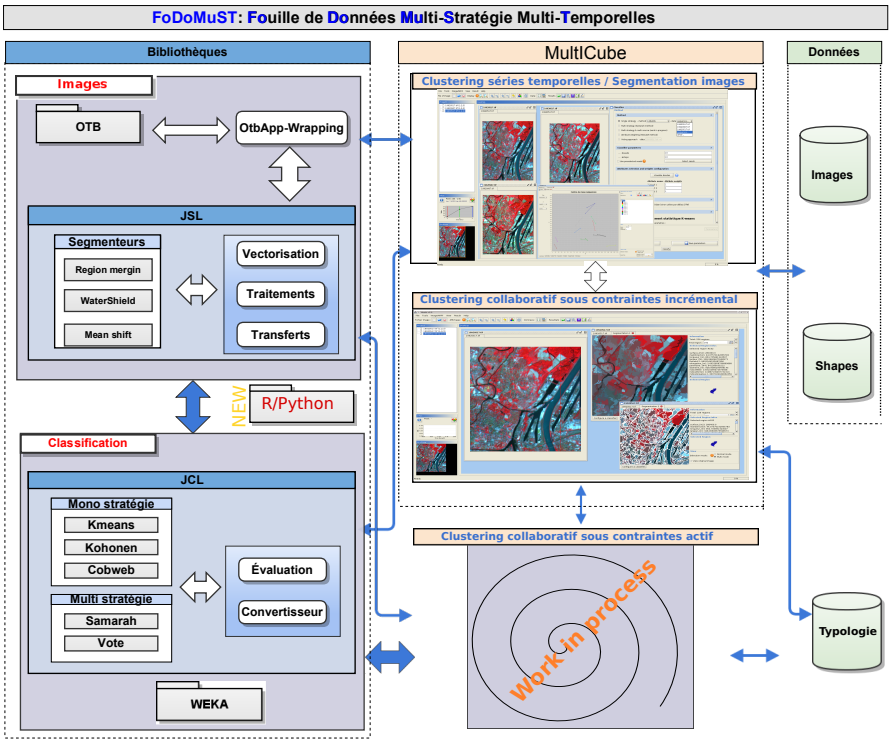

FIG. 1 – *Architecture générale de la plateforme* FODOMUST

2.1 Principales composantes

Le cœur de la plateforme est composé de deux librairies :

- JCL qui regroupe un ensemble de classifieurs non supervisés classiques ou flous tels que Kmeans, Cobweb, CHA ou Kohonen en code propre ainsi que la méthode de clustering collaborative multistratégie sous contraintes incrémentale SAMARAH ;
- JSL qui regroupe des algorithmes de segmentation soit propres à ICube soit proposés par l'Orfeo Tool Box (OTB) [2].

et de trois niveaux d'interfaçage (liés à la complexité des données) permettant une interaction simplifiée avec l'expert :

- CLASSIFX dédiée à l'analyse de données numériques ou catégorielles (ARFF ou CSV)
- IVISUALIZE dédiée à la classification de données issues de BDs Topo IGN.
- MULTICUBE dédiée à l'analyse et la classification de séries temporelles d'images (seule représentée sur la figure)

Enfin, un développement récent permet à l'utilisateur d'accéder directement à quasi toutes les fonctionalités de JCL ou JSR via Python ou R.

Dans cet article, nous présentons uniquement l'interface MULTICUBE (Multistratégie, Multirésolution, Multitemporel) qui permet d'appliquer la méthode SAMARAH sur des séries temporelles d'images pour des applications privilégiées en télédétection.

2.2 JCL : Java Clustering Library

2.2.1 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 monodates ou multidates simples (entier, réel, symbolique) ou construits (tableaux, structures) au format ARFF ou CSV via CLASSIFX, au format images TIF, BMP ... via MULTICUBE) ou plus spécialisées comme des données géographiques issues de bases de données topographiques de l'IGN via IVISUALIZE.

2.2.2 Distance inter-objets et moyenne

Afin de pouvoir appliquer les algorithmes de classification basés sur une distance, une mesure de similarité peut être (re)définie pour chaque type d'attributs. Par exemple :

- pour les données simples numériques et structurées (tableau ou structure) : la distance euclidienne, éventuellement pondérée, est proposée par défaut ;
- pour les données symboliques : une matrice de similarité doit être définie via une interface dédiée à la gestion de ces matrices ;
- pour les données temporelles : l'utilisateur peut choisir entre la distance euclidienne et DTW. Pour les distances euclidiennes, la moyenne euclidienne est utilisée. Pour DTW, la moyenne DBA (DTW Barycenter Averaging) est implantée (Petitjean et al., 2011).

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

2.2.3 SAMARAH

La version de base de SAMARAH multisource (images de même résolution) (Gançarski et Wemmert, 2007) a été complétée par un mécanisme de prise en compte des contraintes.

2.3 L'interface MULTICUBE

L'analyse d'images de télédétection à des résolutions comprises entre 5 et 500 m se fait généralement au niveau des pixels à partir des valeurs radiométriques de ceux-ci. Avec les images à très haute résolution spatiale proche du mètre, ces méthodes 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 classifier les régions ainsi obtenues en utilisant éventuellement des connaissances du domaine. La plateforme FODOMUST offre des outils permettant de créer de telles régions.

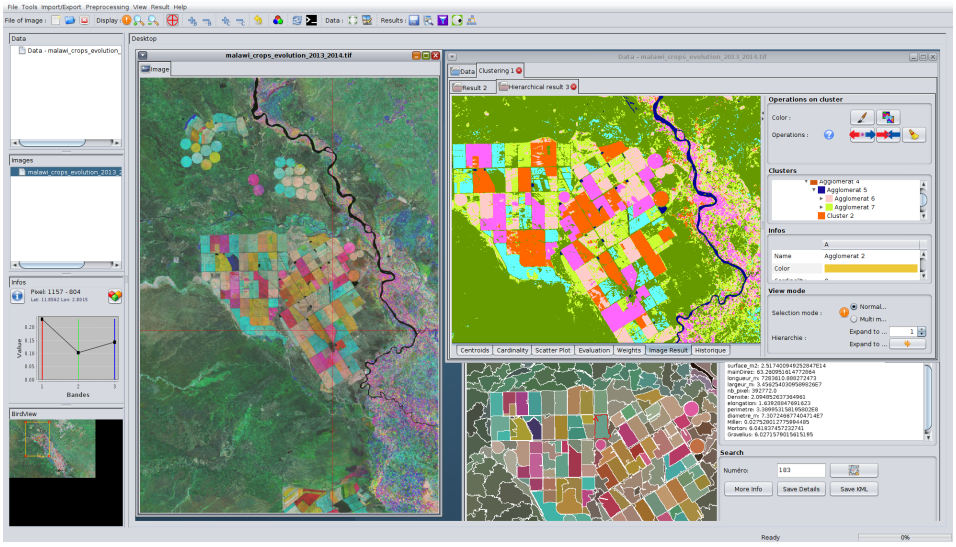

FIG. 2 – MULTICUBE *: Une interface pour l'analyse d'images orientée régions*

La figure 2 présente l'interface MULTICUBE avec :

- un bandeau (bord supérieur de la figure) regroupant les principaux outils de traitement d'images (contraste, découpe, segmentation…)
- un panel (à gauche) permettant la gestion des fichiers et des données chargées
- un panel (au centre) présentant l'image en fausses couleurs
- un panel (en haut à droite) montrant le résultat d'une classification non supervisée de l'image en deux étapes : ici, l'application d'un algorithme Kmeans (25 clusters) suivi d'un regroupement hiérarchique ascendant
- un panel (en bas à droite) montrant d'une part, une segmentation effectuée sur l'image et d'autre part, un extrait des caractéristiques associées à chacun des segments construits. Ces régions peuvent alors être directement classifiées ou simplement sauvegardées.

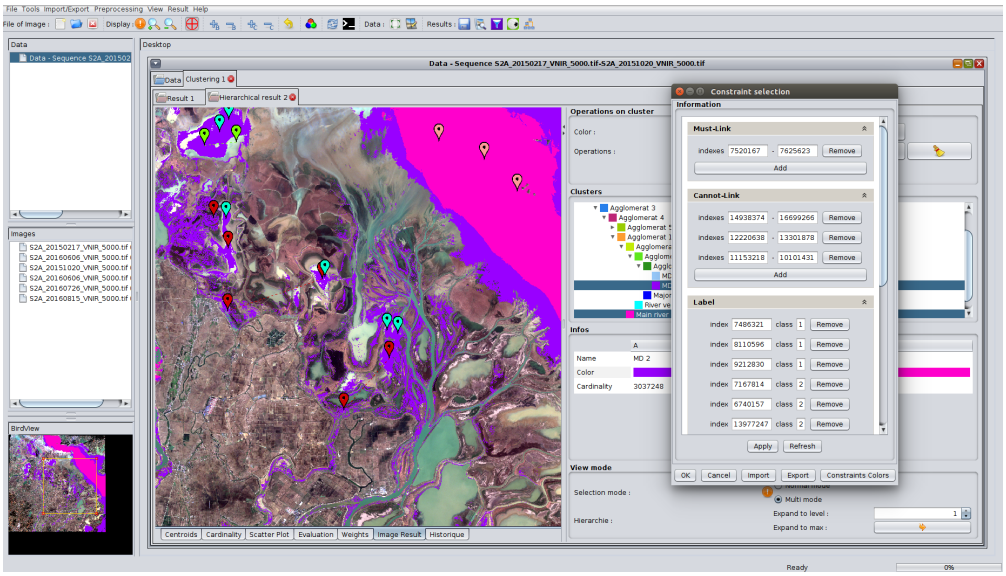

FIG. 3 – MULTICUBE *: Une interface pour le clustering sous contraintes*

Les derniers développements de SAMARAH liés à la prise en compte incrémentale de contraintes (Section 3.2) ont été intégrés. L'expert peut donner de nouvelles contraintes directement sur les images initiales ou résultats via un panel dédié de MULTICUBE (Fig 3).

3 Clustering collaboratif sous contraintes incrémental

3.1 SAMARAH : une approche collaborative multistratégie

Les résultats obtenus par un classifieur non supervisé sont très dépendants des choix initiaux faits par l'utilisateur (algorithme et similarité utilisés, nombre de clusters attendus, etc.). Ces choix sont donc à effectuer avec précaution afin de limiter leur influence. Cependant, il est difficile (voire impossible) d'identifier une recette idéale et générique pour cela. En l'absence d'une telle recette, une solution tentante est de ne pas avoir à choisir et d'utiliser plusieurs algorithmes de clustering avec plusieurs paramétrages et, s'inspirant du succès des méthodes d'ensembles en apprentissage supervisé, de proposer des méthodes de combinaison de classifieurs basée sur la collaboration (Cornuejols et al., 2017). Bien qu'efficaces dans de nombreux domaines, la plupart de ces méthodes ne permettent pas de tirer profit de la variété, de la complémentarité et de la profusion des méthodes de clustering existantes. De nombreux travaux se sont intéressés au développement de métaclassifieurs capables de faire collaborer des classifieurs variés présentant des stratégies de clustering pouvant être différentes. La méthode SAMARAH présente une telle architecture originale et générique de collaboration entre des classifieurs, chacun avec sa propre stratégie interne, afin qu'ils améliorent mutuellement leurs résultats jusqu'à obtenir des résultats "similaires" de qualité. Ces résultats peuvent alors être aisément unifiés. Le processus mis en œuvre se déroule en trois étapes principales (Fig. 4) :

– 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

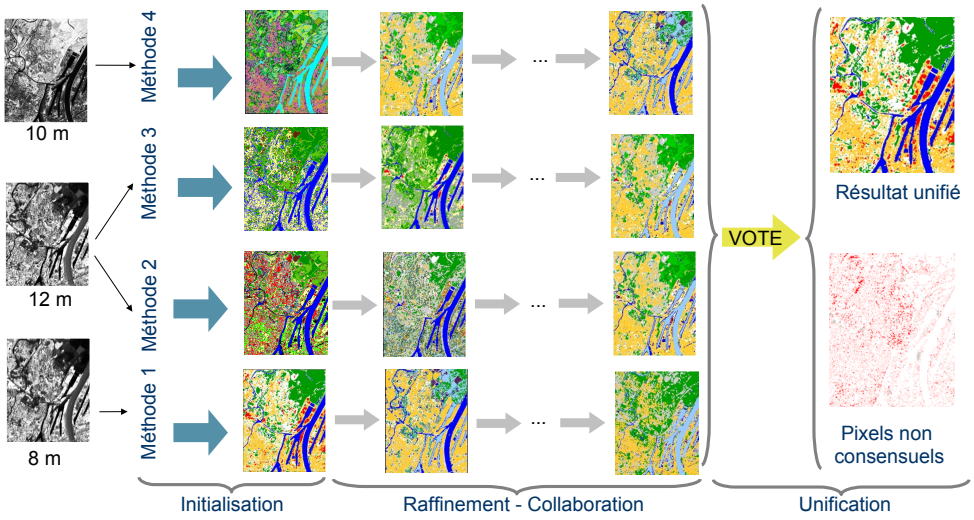

FIG. 4 – SAMARAH *: Classification collaborative multistratégie*

La seule contrainte pour qu'une méthode de clustering générique puisse être intégrée sous forme d'un classifieur collaborant est que celui-ci puisse effectuer trois types d'opérations (scission, fusion et suppression de clusters) sur le résultat de la méthode et ce itérativement. Toute méthode sur laquelle ces trois opérations peuvent être définies est intégrable aisément dans FODOMUST. Ceci est le cas pour une grande majorité des méthodes existantes : Kmeans, Kohonen, Cobweb, EM... Le choix de l'opération à appliquer par chaque classifieur est lui choisi par le métaclassifieur par rapport aux performances du groupe.

3.2 Clustering collaboratif sous contraintes incrémental

La capacité des méthodes de clustering (Petitjean et al., 2012; Aghabozorgi et al., 2015; Khiali et al., 2019) à extraire des regroupements de pixels de qualité est maintenant prouvée. Cependant, le processus de clustering est par définition une approche non supervisée, c'est-à-dire qu'il se base uniquement sur les données et n'utilise pas de connaissances. Or, sans aucune supervision, les algorithmes de clustering peuvent aboutir à des solutions non pertinentes pour l'expert. De plus en plus de recherches proposent d'intégrer les connaissances de l'expert humain sous forme de contraintes opérables et ce dans un formalisme indépendant du domaine. Trois types de contraintes sont communément considérées dans ce cadre :

— Le premier définit des contraintes entre objets (contraintes de comparaison) principalement de ressemblance/dissemblance, comme par exemple des relations « must-link » / « cannot-link » ou plus complexes mettant en jeu plusieurs objets ;
— Le second utilise des objets étiquetés (contraintes d'étiquetage) correspondant directement à des connaissances du domaine ;
— Le troisième propose des contraintes de qualité (nombre, taille, densité, ...) sur des clusters eux-mêmes (contraintes de structure).

Ces contraintes a priori plus aisées à définir, peuvent alors être utilisées pour guider le processus de clustering collaboratif afin de produire des clusters plus proches des « intuitions » de l'expert. Ainsi, la méthode collaborative SAMARAH permet la prise en compte de contraintes dans le processus collaboratif de clustering (Lampert et al., 2019) soit au niveau global du métaclassifieur soit au niveau local des classifieurs de base en les intégrant directement dans la fonction de qualité de chaque clustering. Dans ce cas, les contraintes sont réparties entre les classifieurs : chacun a ainsi une vision partielle des contraintes. Cela a un double objectif de créer de la diversité entre les classifieurs et d'autoriser la présence de classifieurs qui n'intégrent pas de mécanisme de gestion des contraintes. Cependant, donner les contraintes pertinentes (objets à labéliser, contrainte de comparaisons à appliquer...) a priori, c'est-à-dire ayant un impact positif sur le résultat final, est souvent très difficile pour l'expert. Il est donc indispensable de l'aider à travers un processus incrémental l'autorisant à donner ces contraintes « à la volée » en fonction de l'avancement du processus et de la qualité du résultat courant. Or, pour définir des nouvelles contraintes, l'expert utilise quasi-exclusivement une visualisation de la scène et quelques critères statistiques. Ainsi, une extension à l'interface MULTICUBE permet d'injecter des nouvelles contraintes directement sur les résultats produits (Fig 3).

4 Conclusion

La version actuelle de FODOMUST est maintenant à un niveau de finition permettant sa diffusion à l'ensemble de la communauté de la fouille de données temporelles. Cet article a mis l'accent sur le cas de l'imagerie satellittaire bien que la plateforme soit générique. Les expériences montrent la pertinence et le bénéfice à utiliser un processus collaboratif sous contraintes incrémental pour la classification de séries temporelles d'images. Cependant, avec l'augmentation du volume des données et du nombre de classes d'évolution potentielles, la mise en évidence et la formalisation de telles contraintes dans le cadre de l'analyse temporelle apparaît comme plus difficile que prévu et potentiellement chronophage. Ainsi, les expériences ont montré que l'expert n'avait pas de moyen de savoir si les contraintes qu'il proposait étaient cohérentes entre elles et pertinentes a priori. De fait, sélectionner les nouvelles informations est d'après nous, un verrou scientifique important d'autant plus qu'il est indispensable d'optimiser l'utilisation de ces nouvelles informations venant de l'expert, car si celui-ci ne voit pas d'amélioration rapide de la solution suite à son aide, il perdra rapidement confiance dans le système. Mais, paradoxalement, les perturbations potentielles de la solution courante devront être limitées afin de ne pas désorienter l'expert. Notre objectif est donc de proposer, valider et implanter dans FODOMUST une méthode innovante de clustering collaboratif sous contraintes interactif. Il s'agit de permettre à l'expert d'ajouter voire supprimer des contraintes « à la volée ». Pour cela il sera aidé par des conseils ou des propositions de nouvelles contraintes émis par la méthode elle-même.

Références

Aghabozorgi, S., A. Shirkhorshidi, et T. Wah (2015). Time-series clustering—a decade review. *Information Systems 53*, 16–38.

Basu, S., I. Davidson, et K. Wagstaff (2008). *Constrained Clustering : Advances in Algorithms, Theory, and Applications* (1 ed.). Chapman & Hall/CRC.

Cornuejols, A., C. Wemmert, P. Gançarski, et Y. Bennani (2017). Collaborative clustering : Why, when, what and how. *Information Fusion 39*, 81–95.

Dao, T.-B.-H., K.-C. Duong, et C. Vrain (2017). Constrained clustering by constraint programming. *Artificial Intelligence 244*, 70–94.

Flamary, R., M. Fauvel, M. D. Mura, et S. Valero (2018). Analysis of multi-temporal classification techniques for forecasting image times series. *IEEE Geosci Remote S 12*(5), 953–957.

Forestier, G., P. Gançarski, et C. Wemmert (2010). Collaborative clustering with background knowledge. *Data Knowl Eng 69*(2), 211–228.

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

Khiali, L., M. Ndiath, S. Alleaume, D. Ienco, K. Ose, et M. Teisseire (2019). Detection of spatio-temporal evolutions on multi-annual satellite image time series : A clustering based approach. *Int J Appl Earth Obs Geoinf 74*, 103–119.

Lampert, T., B. Lafabregue, T.-B.-H. Dao, N. Serrette, G. Forestier, B. Crémilleux, C. Vrain, et P. Gançarski (2018). Constrained distance based clustering for time-series : A comparative and experimental study. *Data Min Knowl Discov 32*(6), 1663–1707.

Lampert, T., B. Lafabregue, T.-B.-H. Dao, N. Serrette, C. Vrain, et P. Gançarski (2019). Constrained distance based clustering for satellite image time-series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Petitjean, F., J. Inglada, et P. Gançarski (2012). Satellite image time series analysis under time warping. *IEEE Transactions on Geoscience and Remote Sensing 50*(8).

Petitjean, F., A. Ketterlin, et P. Gançarski (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognit 44*(3), 678–693.

Summary

The FODOMUST platform is a concrete implementation of methods, libraries and interfaces dedicated to the clustering of complex data developed within ICube. It integrates a multi-source version of the multi-strategy collaborative clustering method SAMARAH. Its main originality is that it offers the user methods and interfaces allowing clustering under incremental constraints of symbolic or numerical time data. The interface MULTICUBE dedicated to time series analysis also provides access to a set of segmentation algorithms either specific to ICube or using the OTB distributed by CNES. (See https:icube-sdc.unistra.fr/fr/index.php/Plateformes).