Fouille d'images géoréférencées avec RapidMiner

Thomas Guyet[1,3], Hervé Nicolas[2,3], Boussad Ghedamsi[1,3], Élise Athané[2,3]

1. AGROCAMPUS-OUEST/IRISA, Laboratoire d'Informatique
65 rue de Saint-Brieuc, F-35042 Rennes, France
thomas.guyet@agrocampus-ouest.fr

2. AGROCAMPUS-OUEST/INRA, UMR1069 Sol Agro et hydrosystème Spatialisation
65 rue de Saint-Brieuc, F-35042 Rennes, France
herve.nicolas@agrocampus-ouest.fr

3. Université Européenne de Bretagne

RÉSUMÉ.
Cet article présente l'intégration de données spatiales dans le logiciel de fouille de données RapidMiner. RapidMiner est un outil pour concevoir interactivement des chaînes de traitements, orientées vers la fouille de données. Dans le cadre de ce travail, nous avons implémenté une extension de RapidMiner pour traiter des données spatiales. On bénéficie ainsi d'une large gamme de méthodes de fouille de données déjà à disposition pour répondre à une question d'analyse sur des données spatiales. L'article illustre le fonctionnement de l'extension par une comparaison de méthodes de classification sur un problème de détection de plante invasive (la Jussie) à partir d'une image aérienne hyperspectrale.

ABSTRACT.
This article presents the integration of spatial data in the data mining software RapidMiner. RapidMiner is a data mining tool to design data processing chains. In this work, we implemented an extension of RapidMiner to process spatial data. It benefits from the wide range of data mining methods already available and it enables spatial data users to answer to their analysis questions. The paper illustrates the extension by a comparison of classification methods on a problem of detecting invasive plant (Jussie) from an hyperspectral image.

MOTS-CLÉS : Fouille de données, images de télédétection, chaîne de traitement, évaluation

KEYWORDS: Data mining, remote sensing images, processing chain, evaluation

1. Introduction

L'augmentation continuelle de la quantité de données spatialisées, la démocratisation de leur accès et de leurs usages fait croître l'intérêt de ces données pour de nombreuses applications allant de la prédiction en agronomie au géo-marketing. Des outils d'analyse de données sont nécessaire aider à l'exploitation de cette masse de données. Les outils d'analyse de données intègrent à la fois des méthodes de traitement et prétraitements des données, des méthodes d'extraction d'information et des outils de visualisation. Les méthodes de fouille de données désignent quant à elle, l'ensemble des méthodes algorithmiques et statistiques qui résolvent des tâches élémentaires d'extraction d'information, p. ex. classification, clustering, regression (voir (Miller et Han, 2009) pour une revue des questions actuelles en fouille de données spatiale). Par la suite, on s'intéresse principalement à des tâches d'analyse d'images géoréférencées.

Les outils d'analyse de données prenant en compte l'information spatiale peuvent être organisés en deux grandes classes : d'une part, les outils dédiés à l'analyse et à la visualisation de données spatiales (tels que les Systèmes d'Information Géographiques, SIG, ou les outils d'analyse d'images géoréférencées) et, d'autre part, les outils de fouille de données.

La première catégorie offre l'avantage d'être bien adaptées aux données géographiques, nécessitant des traitements efficaces en gestion de la mémoire et du temps de calcul, p. ex. ENVI ou IDRISI. Ces outils intègrent des méthodes dédiés aux tâches courantes d'analyse d'images tels que des méthodes d'apprentissage supervisé. Deux problèmes se posent aux utilisateurs, d'une part, le choix du paramétrage de ces algorithmes. La forte dépendance des paramètres aux données rend cette question du paramétrage difficile. D'autre part, les outils intégrés ne permettent pas aux utilisateurs d'explorer d'autres méthodes d'apprentissage existantes. L'Orfeo Toolbox (Inglada et Christophe, 2009) offre des possibilités étendues pour le traitement de l'information spatiales. Mais cette boîte à outils n'intègre que très peu de méthodes d'analyse de données.

La seconde catégorie offre une large gamme d'algorithmes d'analyse de données ainsi que des fonctionnalités pour les comparer et pour choisir le paramétrage des algorithmes qui convienne le mieux aux données. Cependant, ces outils sont souvent difficiles à utiliser pour les non-informaticiens, des logiciels comme R, Weka (Hall *et al.*, 2009) ou Sci-Learn (Pedregosa *et al.*, 2011) nécessitent de la programmation pour mener des analyses de données. De plus, ils n'intègrent que rarement des traitements explicites sur la dimension spatiale des données géographiques.

Il apparaît nécessaire de proposer des outils facilitant l'expérimentation de méthodes de fouille de données potentiellement complexes sur des données géographiques pour des utilisateurs cherchant à analyser ces données. On ne cherche

pas à reconstruire un outil capable de traiter directement les images, mais plutôt d'aider au prototypage rapide de la méthode d'analyse. Un tel outil doit donner la possibilité d'expérimenter différentes méthodes d'analyse (choix des algorithmes et des paramètres) et de comparer leurs performances sur des jeux de test. La mise en production du traitement des données devant être menée dans un second temps par l'implémentation d'un programme dédié ou depuis les plate-forme dédiées.

L'approche proposée dans cet article est l'intégration de fonctionnalités liées aux données spatiales dans un outil de fouille de données disposant de la capacité de mise en œuvre de chaînes complètes d'analyse de données. On cherche ainsi à palier les difficultés rencontrées par l'une ou l'autre des classes d'outils existants : (1) on facilite l'accès à une large gamme d'algorithmes de fouille de données sur des données spatiales, (2) on rend possible l'analyse automatique des performances de la méthode d'analyse pour faciliter les choix des paramètres.

Notre choix s'est porté sur l'utilisation de l'outil de fouille de données nommé RapidMiner. D'autres outils tels que Orange (Curk *et al.*, 2005) ou KNIME (Berthold *et al.*, 2007) offrent des fonctionnalités similaires, le choix de RapidMiner a été fait, d'une part, pour sa visibilité internationale et, d'autre part, par l'importance du support au développement d'extensions du logiciel.

À notre connaissance, aucune extension pour l'utilisation de données geospatiales n'a été proposée pour aucune de ces trois plate-forme. (Burget *et al.*) ont proposés l'extension IMMI de RapidMiner pour intégrer des fonctionnalités de fouille d'images visuelles, mais ces fonctionnalités ne permettent pas d'intégrer les aspects spatiaux des images géoréférencées.

2. Vers une extension de RapidMiner pour des données spatiales

2.1 Présentation du logiciel RapidMiner

RapidMiner (Mierswa *et al.*, 2006), anciennement appelé YALE (Yet Another Learning Environment) est un environnent d'apprentissage, de fouille de données, d'analyse prédictive et d'analyse de données. Actuellement, il est très largement utilisé[1] dans le domaine de la recherche, de l'éducation et en milieu d'entreprise. C'est un logiciel puissant pour mettre en place rapidement une chaîne complète de traitement de données, de la saisie des données à la visualisation des résultats d'analyse.

Une chaîne de traitements de RapidMiner est constituée d'un ensemble de « blocs » reliés entre eux pour représenter la succession des traitements. Chaque bloc correspond à un traitement : les blocs de chargement de données depuis un

[1]RapidMiner a été classé de 2010 à 2012 comme l'outil interactif de fouille de données le plus utilisé au niveau mondial (voir enquête de KDNuggets, http://www.kdnuggets.com)

fichier, les blocs de transformation des données, les blocs implémentant des algorithmes de fouille de données. En sortie de la chaîne de traitements, les données peuvent être visualisées ou exportées. La Figure 1 illustre l'interface du logiciel dont la zone centrale contient une chaîne de traitements simple.

Pour étendre les possibilités de traitement, des méta-blocs définissent des modalités de traitements spécifiques à certaines analyses de données. Par exemple, le méta-bloc *XValidation* facilite la réalisation de validations croisées tandis que le bloc *Optimize Parameters* explore certains paramètres d'une chaîne de traitements pour identifier la meilleure combinaison de paramètres relativement à une évaluation.

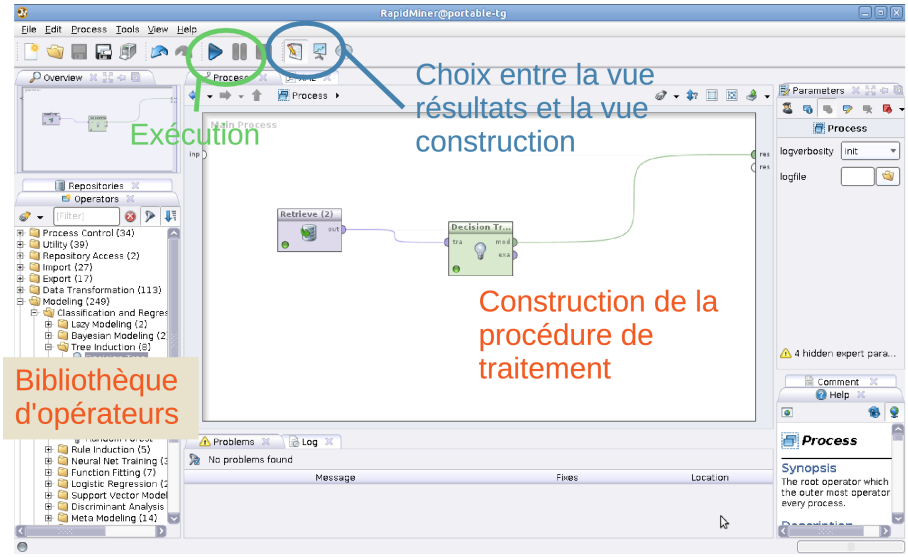

Figure 1. Illustration de l'interface de RapidMiner. Dans la zone centrale, la chaîne de traitement contient deux blocs. Les panneaux gauche et droite sont respectivement dédiés à l'ajout de bloc et au paramétrage du bloc sélectionné.

Pour que les blocs fonctionnent correctement, il est nécessaire de leur fournir en entrée des données correctement structurées. Les blocs implémentant des méthodes de fouille de données requièrent des données organisées de manière tabulaire, c'est-à-dire de données constituées d'enregistrements (une ligne dans un tableau) décrits par différents attributs (les colonnes du tableau).

RapidMiner est un logiciel conçu de manière modulaire, il offre la possibilité de programmer ses propres blocs. Une extension RapidMiner est un ensemble de blocs qui peuvent être utilisés dans l'interface de RapidMiner. Parmi les extensions plus plus populaires, on peut citer l'intégration de Weka ou de R.

2.2 Objectifs d'une extension pour les données spatiales

Les objectifs d'une extension RapidMiner pour les données spatiales est de pouvoir mettre en place les différentes méthodes d'extraction d'information à partir de données spatiales telles que listées par (Miller et Han, 2009) en s'appuyant les méthodes de fouille de données (non-spatiales) déjà existantes.

RapidMiner étant un logiciel de fouille de données ne supportant par le format des données géographiques, le défi principal est de rendre ces données accessibles depuis ce logiciel en implémentant une extension pour les données spatiales. Cette extension inclut des blocs de lecture et d'écriture de données. Pour garantir la communication entre nos blocs et les blocs préexistants, il a été nécessaire de respecter le principe d'utilisation des données tabulaires de RapidMiner.

Dans le cadre de cette première version de l'extension, nous nous sommes focalisé sur des tâches de classification d'images de télédétection (images satellite ou aériennes), de prédiction d'observations spatialisées dans le temps à partir de séquences d'images géoréférencées et d'évaluation des performances des deux tâches précédentes.

3. Extension GeoDM - Geospatial Data Mining

Dans cette section, on présente l'extension « Geospatial Data Mining » en introduisant le principe de transformation des données géoréférencées aux données tabulaires, puis on présente les différents modules de l'extension pour en décrire les fonctionnalités principales.

3.1. Principe de transformation des données géoréférencées aux données tabulaires

Le principe de transformation des données géoréférencées est l'utilisation d'un échantillonnage spatial, c'est-à-dire d'un ensemble de points définis par leurs coordonnées dans l'espace (coordonnées et système de coordonnées associé).

Chaque point d'un échantillon correspond à un enregistrement dans le format tabulaire. Les couches de données utilisées servent à définir les attributs qui caractérisent chacun des points de l'échantillon. La Figure 2 illustre la construction d'un jeu de données à partir d'un échantillon. Pour chaque point de l'échantillon, la valeur des attributs de chaque couche est récupérée à la position de ce point. Lorsque le point se trouve en dehors d'une couche, il est possible de donner une valeur par défaut à l'attribut ou de le définir comme « inconnu ».

L'intérêt majeur de cette méthode d'échantillonnage est d'utiliser des couches avec des représentations hétérogènes : couches vecteur et couches raster, couches

raster avec des résolutions spatiales différentes, couches avec des géoréférencements différents. Cette approche n'est néanmoins pas totalement générique. Elle est limitée par l'utilisation de couches de données au format vectorielle qui, d'une part, ne contiennent que des polygones et que, d'autre part, ces polygones ne se superposent pas. Dans le cas de superposition de polygones, le nombre d'attributs à prendre en compte pour un point qui se situe dans la zone de recouvrement ne serait pas le même que pour les autres points. Ces limitations restent néanmoins raisonnables par rapport aux usages prévus de l'extension GeoDM.

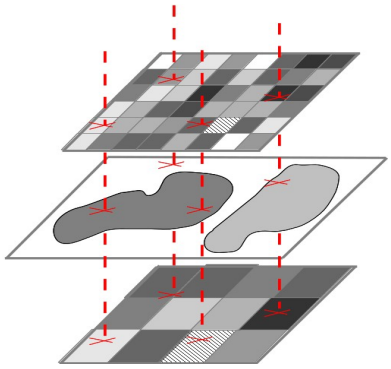

Figure 2. Illustration de la méthode d'échantillonnage. L'échantillon comprend ici 4 points visualisés en rouges (tirets-verticaux pointillés). Pour chaque point, on récupère les attributs pour les trois couches (deux raster et une vecteur).

Plusieurs types d'échantillonnages sont prévus pour répondre aux besoins d'analyses spatiales des données :

– l'échantillonnage selon un *transect* sert à l'analyse de l'influence d'un phénomène selon une dimension de l'espace : influence d'un objet sur une variable bio-physique, étude de la transition d'un état à un autre (ex. de l'urbain au rural) (Buckland, 2001).

– l'échantillonnage aléatoire assure un sous-échantillonnage d'une zone rectangulaire de manière à répéter des apprentissages avec différentes données,

– l'échantillonnage régulier correspondant, par exemple, aux centres des pixels d'une image raster géoréférencée,

– l'échantillonnage spécifique que peut définir un utilisateur par une couche de points.

3.2 Présentation générale

L'extension est organisée en trois modules :

– le module « I/O » contient des blocs destinés à ouvrir ou enregistrer des données géoréférencées. Actuellement, les données peuvent être chargées sous la forme d'images vectorielles (format Shapefile) et d'images raster (format GeoTiff). Ce module s'appuie sur les fonctionnalités de la GeoTools[2] et prend pleinement en charge les transformations géométriques entre différentes sources de données. L'enregistrement et la visualisation de données géoréférencées est possible uniquement en format raster (GeoTiff).

– le module « Sampling » contient les outils pour la génération et la manipulation d'échantillons (voir section précédente).

– le module « Data Transformation » contient des outils de transformation des données au format spécifique à l'extension. Il introduit des outils pour la gestion des attributs de coordonnées. En particulier, les attributs de coordonnées doivent être facilement masqués avant d'effectuer une tâche de fouille de données.

4. Exemples d'utilisation : comparaison de méthodes de classification supervisée sur des images hyperspectrales

Nous utilisons ici notre extension GeoDM dans le but d'identifier une bonne méthode de classification supervisée pour la classification d'images hyperspectrales[3]. De telles méthodes sont proposées dans des outils tels que ENVI, mais il peut être difficile de savoir laquelle est la mieux adaptée et quels paramètres des algorithmes d'apprentissage fonctionnent le mieux. On présente deux expérimentations permettant de répondre ces questions : une expérimentation comparant plusieurs algorithmes de classification supervisée, puis la comparaison de résultats obtenus avec différents paramétrages des SVM.

Notre objectif n'est pas de discuter les méthodes de classification elles-mêmes mais plutôt de mettre en évidence que l'utilisation de GeoDM facilite le choix des paramètres et la validation d'une telle méthode.

4.1. Problématique applicative et données

La jussie est une plante aquatique invasive que les gestionnaires de rivière surveillent particulièrement pour éviter son pullulement. Ce travail s'inscrit dans une étude sur la faisabilité d'une approche de la détection de la jussie par télédétection. Le site d'étude se trouve au confluent de l'Oust et de l'Aff, à Glénac (Morbihan).

Pour cette expérimentation, on dispose d'une image aérienne hyperspectrale comprenant 160 bandes spectrales ainsi qu'un fichier *shapefile* contenant une vérité terrain obtenue par des observations ponctuelles partielles. Quatre classes sont

[2]GeoTools : www.geotools.org

[3]Une image hyperspectrale est une image contenant un grand nombre de bandes. Chaque bande correspondant à la réponse du sol à une certaine longueur d'onde.

distinguées : *eau*, *jussie*, *nénuphar* et *autre végétation*. Le but de la classification de l'image est d'identifier en tout point de l'image la présence ou non de jussie.

Pour limiter les temps de calcul, seules les 15 premières bandes de l'image hyperspectrale sont utilisées. On utilise également une image contenant les 15 premières bandes de la transformation MNF, Minimum Noise Fraction (Boardman et Kruse, 1994). De manière similaire à une analyse en composantes principales, la transformation MNF est une transformation linéaire des données qui consiste à séparer le signal du bruit. Les premières bandes de l'image transformée contiennent ainsi les données les plus informatives de l'image originale.

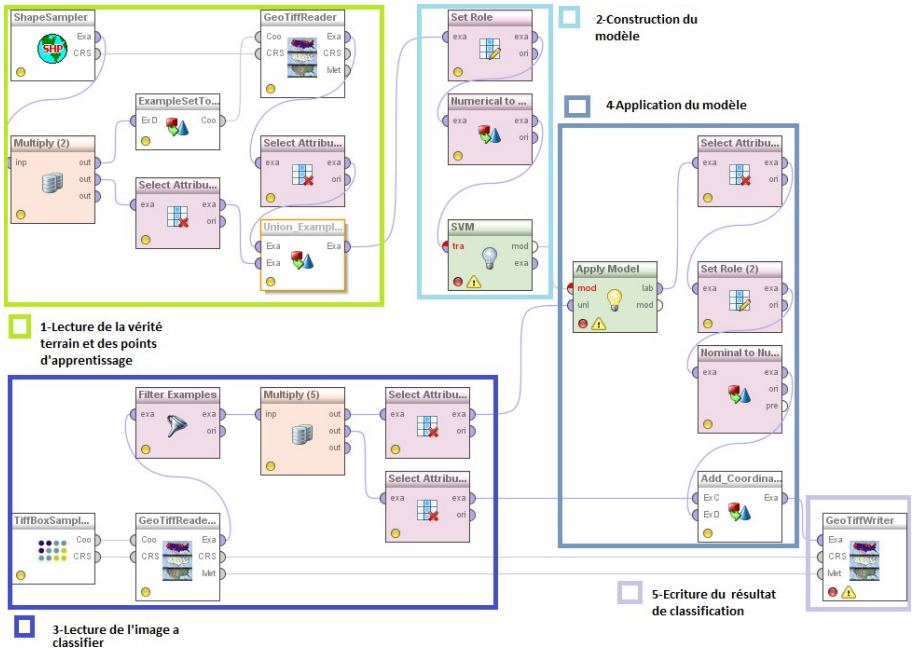

Figure 3. Illustration de la chaîne de traitements pour la classification supervisée d'une image géoréférencée. Les blocs blancs sont des blocs spécifiques à l'extension GeoDM. Les autres blocs sont des blocs prédéfinis dans RapidMiner.

4.2. Utilisation de l'extension GeoDM pour la classification supervisée d'une image hyperspectrale

On décrit le processus de la Figure 3 pour la classification supervisée de l'image hyperspectrale. Ce processus illustre une large gamme des fonctionnalités de l'extension GeoDM. La chaîne de traitements comporte cinq parties distinctes :

– les parties 1 et 2 effectuent l'apprentissage du classifieur. La partie 1 charge le jeu d'apprentissage depuis un fichier *shapefile* contenant la vérité terrain, et d'un

fichier GeoTiff contenant les informations spectrales. Le fichier de vérité terrain sert à définir les points d'échantillonnage. La partie 2 construit le classifieur. L'illustration montre l'usage d'un classifieur SVM, mais il existe de nombreux autres algorithmes déjà implémentés dans RapidMiner.

– la partie 3 et 4 classent une image à partir du classifieur construit précédemment. La partie 3 charge une image raster à classifier (sans vérité terrain ici), puis la partie 4 classe les pixels de l'image.

– finalement, la partie 5 enregistre l'image construite par classification dans un nouveau fichier GeoTiff.

Plus que le calcul d'une image résultant d'une classification, l'usage de RapidMiner facilite l'automatisation des comparaisons entre plusieurs modèles et entre plusieurs paramètres. Pour cela, on utilise des fonctionnalités d'automatisation disponibles dans RapidMiner telles que le bloc de validation-croisée ou le bloc de test des combinaisons de paramètres.

Pour illustrer les types de résultats recherchés nous avons construit deux chaînes de traitements dont les résultats sont présentés dans les sections suivantes. Pour des raisons de place, on ne présente pas ici les détails des chaînes de traitements mises en place. Plus d'information sur les chaînes de traitements qui peuvent être construites sont disponibles sur le site de l'extension[4].

4.3. Une chaîne de traitements pour la comparaison de différents algorithmes

La première expérimentation nous a conduit à comparer différents algorithmes de classification (*SVM*, *DecisionTree*, *NeuralNetworks* à trois couches, *k-NN*). Les évaluations sont effectuées à l'aide d'une validation croisées (10 validations) selon le critère kappa (κ). L'indice kappa est une valeur entre 0 et 1. Plus κ est proche de 1, meilleures sont les performances de classification.

Tableau 1. Indice kappa (κ) pour les classifications pour différents algorithmes.

	Image originale	**Image MNF**
SVM	0.151 +/- 0.013	0.778 +/- 0.014
Neural Network	0.902 +/- 0.012	0.991 +/- 0.003
k-NN	0.854 +/- 0.008	0.996 +/- 0.001
Decision Tree	0.753 +/- 0.025	0.968 +/- 0.006

[4] http://geoimagermp.gforge.inria.fr/

4.4. Une chaîne de traitements pour le paramétrage d'un classifieur SVM

On se place ici dans la situation où l'utilisateur souhaite utiliser un classifieur SVM pour construire ses classes. Le problème qui se pose est d'identifier les paramètres optimaux pour un classifieur *SVM*, c'est-à-dire le jeu de paramètres qui permettra d'obtenir le meilleur kappa. L'utilisateur à classiquement deux paramètres à donner γ et ε spécifiques aux *SVM* avec noyau Gaussien. Dans le cas de l'étude précédente, les paramètres utilisés étaient les paramètres par défaut.

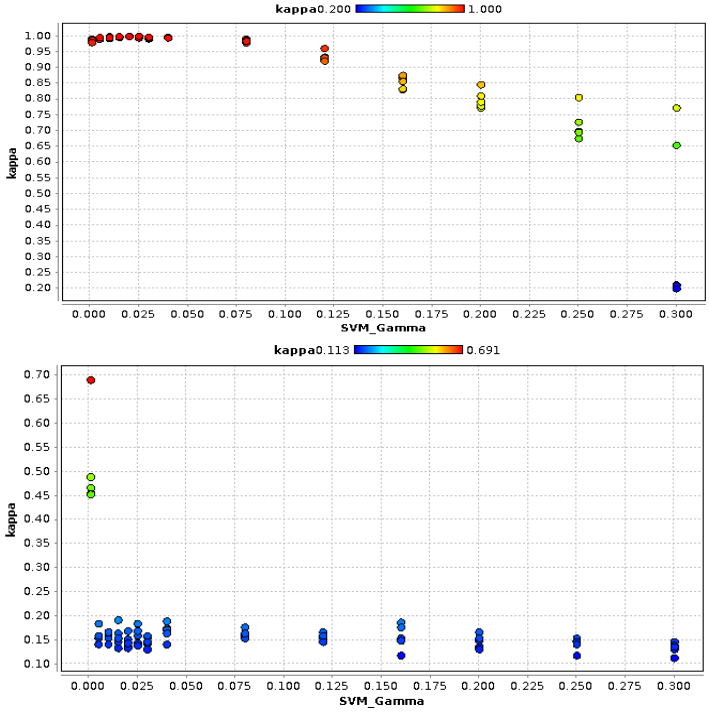

Figure 4. Résultats pour l'image originale (à gauche) et l'image MNF (à droite). Valeurs de κ en fonction de γ.

La méthode consiste à mettre en place une chaîne de traitement qui fait varier les paramètres γ et ε. Pour chaque combinaison de paramètres, une validation croisée évalue les performances du classifieur. Après lancement de cette chaîne de traitement, RapidMiner fournit la Figure 4 pour analyser les résultats pour l'image originale et pour l'image MNF.

On constate d'une part que les résultats obtenus avec l'image originale sont beaucoup moins bons qu'avec l'image MNF et, d'autre part, que les meilleurs

résultats sont obtenus pour γ=0,02 et ε=0,5 (κ=1). Ces valeurs pourront être réutilisées pour paramétrer un classifieur *SVM* dans ENVI par exemple.

Pour une image, l'ensemble des traitements a pu être réalisés en moins d'une minute. On obtient ainsi rapidement une indication importante sur le paramétrage qui guidera l'expert dans sa classification d'image.

L'objectif de ce papier n'étant pas de présenter une amélioration d'une méthode d'analyse d'image, on ne discute pas plus les résultats obtenus par les chaînes de traitements présentées. L'intérêt de cette expérimentation est de montrer que notre extension permet de mettre en place assez simplement une chaîne d'analyse de données à partir de données spatiales répondant à un besoin auquel les outils classiques ne répondent pas.

5. Perspectives et conclusion

Dans cet article, nous nous sommes intéressé à l'utilisation d'un outil de fouille de données, le logiciel RapidMiner, pour construire des chaînes de traitements pour l'analyse spatiale de données. Nous avons proposé une extension du logiciel qui intégre des données géoréférencées dans une chaîne de traitements usuelle. Cette approche a été illustrée pour l'identification d'une méthode optimale de classification d'une image hyperspectrale. Les chaînes de traitements mises en place permettent d'évaluer facilement et rapidement différents algorithmes de classification supervisée et d'identifier les paramètres optimaux d'un classifieur de type *SVM*.

Les applications visées de cette extension sont à la fois le prototypage rapide d'une méthode d'analyse d'images de télédétection, la mise en place de nouvelles méthodes d'analyse de données spatiales et, dans un but pédagogique, la facilitation de l'expérimentation des méthodes de fouille de données déjà existantes sur des données spatiales afin de sensibiliser à ses apports potentiels.

Cette extension a été utilisée dans un cadre pédagogique pour faire découvrir des méthodes de fouilles de données à des étudiants en formation de Master en Geoinformation. Cette usage a montré que l'outil RapidMiner nécessite un temps d'apprentissage. Malgré le caractère intuitif de l'utilisation des boîtes, les chaînes de traitements comportent de nombreux éléments « techniques » pour la transformation des données qui nécessitent une bonne compréhension du fonctionnement du logiciel. Une fois ce temps d'apprentissage passé, GeoDM a permis d'expérimenter différentes méthodes de fouille de données sur des problématiques réelles d'analyse d'images de télédétection. En pratique, les chaînes de traitements mises en place permettent de répondre à la question récurrente de choix du classifieur.

Seule un sous-ensemble limité des méthodes de fouille de données spatiales (Miller et Han, 2009) ont été actuellement intégrées à l'extension GeoDM. La version actuelle de l'extension est orientée vers cette tâche de classification d'images

de télédétection. La modularité de RapidMiner et la généricité de l'approche de spatialisation des données assurent l'intégration future d'un grand nombre de méthodes sous la forme de nouveau blocs ou méta-blocs sans avoir à réimplémenter les outils de base de l'apprentissage et de la fouille de données.

Remerciements :

Les auteurs souhaitent remercier Benjamin Bottner de l'Institut d'Aménagement de la Vilaine pour avoir fournis les images ainsi que Jacques Haury pour son aide à la compréhension des problématiques liées à la Jussie.

Bibliographie

Berthold M., Cebron N., Dill F., Gabriel T., Kötter T., Meinl T., Ohl P., Sieb C. Thiel K., Wiswedel B. (2007) KNIME: The Konstanz Information Miner,Studies in Classification, Data Analysis, and Knowledge Organization, Springer

Boardman J. W., Kruse F. A., (1994) Automated spectral analysis: a geological example using AVIRIS data, north Grapevine Mountains, Nevada, Proceedings of the Tenth Thematic Conference on Geologic Remote Sensing, pp. 407-418.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. and Thomas, L. (2001) Introduction to Distance Sampling: Estimating Abundance of Biological Populations. Oxford University Press. 432pp

Burget R., Karasek J., Smékal Z., Uher, V., Dostal, O. (2010) Rapidminer image processing extension: A platform for collaborative research. In proceedings of the 33rd International Conference on Telecommunication and Signal Processing, p. 114-118.

Curk T., Demšar J., Xu Q., Leban G., Petrovič U., Bratko I., Shaulsky G., Zupan B. (2005) Microarray data mining with visual programming. Bioinformatics, 1;21(3):396-398.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Inglada, J., Christophe E. (2009) The Orfeo Toolbox remote sensing image processing software, Geoscience and Remote Sensing Symposium, Volume 4, pp.733-736.

Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Miller H., Han J. (2009) Geographic Data Mining and Knowledge Discovery, Second Edition (Chapman & Hall/CRC).

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. (2011) Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12:2825-2830.