Extension de l'étiquetage géographique des pixels d'une image par fouille de données

Adrien Gressin*, Nicole Vincent**, Clément Mallet*, Nicolas Paparoditis*

*IGN/SR, MATIS, Université Paris-Est, Saint-Mandé; (firstname.lastname@ign.fr)
** LIPADE - SIP, Université Paris-Descartes, Paris; (nicole.vincent@mi.parisdescartes.fr)

Résumé. Les techniques de classification modernes permettent d'étiqueter les zones non couvertes des bases de données cartographiques, mais souffrent d'un manque de robustesse important. Dans cet article, nous proposons une méthode robuste d'extension d'étiquetage sur l'emprise d'une image satellite, par analyse hiérarchique des données existantes. Notre approche est fondée sur une sélection d'attributs par thème de la base de données, une sélection des pixels d'apprentissage et des classifications par objet de chaque thème. La décision finale d'étiquetage est prise après fusion des classifications par thème. Notre méthode est appliquée avec succès et comparée à plusieurs méthodes de classification, couplant données d'occupation du sol et imagerie spatiale très haute résolution.

1 Introduction

Dans la plupart des pays développés, des bases de données géographiques (BD) sont initiées, même si certaines régions ne sont couvertes que partiellement. En particulier, des BD d'occupation du sol (OCS) à grande échelle sont en cours de réalisation, par agrégation de données existantes et leur assemblage ne permet pas une description complète du territoire. D'autre part, des images satellite de résolution sub-métrique couvrent avec une grande précision géométrique de larges territoires et peuvent donc aider à compléter ces BD d'OCS. Les méthodes de classification supervisée sont largement utilisées pour résoudre ce genre de problème de télédétection (Mountrakis et al., 2011). Cependant, elles sont souvent peu robustes, présentent un fort taux de confusion, sont limitées à certaines thématiques. Enfin, elles nécessitent une sélection manuelle des classes et des zones d'apprentissage afin de traiter le cas de classes composées de différentes apparences. Nous proposons ainsi ici une méthode d'inspection hiérarchique d'une BD existante, permettant d'apprendre indifféremment la (ou les) apparence(s) de chaque thème qui la constitue. Ces informations sont ensuite fusionnées à différents niveaux afin d'obtenir des résultats plus robustes.

2 Méthodologie

La structure hiérarchique des BD géographiques permet de faire ressortir trois niveaux possibles d'inspection : (1) le niveau objet, (2) le niveau thème, (3) le niveau BD. La BD initiale est projetée sur l'image, formant une carte qui lui est superposable et où les pixels sont

étiquetés par un thème : un pixel de cette grille ne peut avoir qu'une seule étiquette. Cependant, il peut aussi n'appartenir à aucun thème ("*non étiqueté*") et leur étiquetage est l'objet de cette étude. Notre méthode est fondée sur le principe d'une inspection hiérarchique, de manière ascendante. D'abord, à chaque objet de la BD est associée une classification de toute l'image. Puis, les classifications sont fusionnées au niveau du thème et la décision finale d'étiquetage est prise pour chaque pixel de l'image (Pal, 2008). Le premier niveau d'inspection permet d'apprendre l'apparence de chacun des objets. Il est composé de deux étapes : (1) une sélection, d'un sous-ensemble de pixels *intérieurs* et d'un *extérieur* qui permet de discriminer le mieux l'objet du reste de l'image et (2) une classification des pixels de toute l'image en deux classes (*intérieur* / *extérieur*). La sélection des sous-ensembles de pixels est fondée sur la maximisation du rappel d'une classification à deux classes des pixels de l'objet. L'étape de classification permet d'obtenir pour chaque objet de chaque thème de la BD, une carte d'appartenance au thème de l'objet. L'étape de fusion par thème permet de prendre en compte les différentes apparences d'un thème et de ne privilégier aucune apparence. À cette étape, toutes les classifications calculées au niveau de l'objet sont fusionnées par thème, afin d'obtenir une seule carte d'appartenance. On la considère comme la probabilité de chaque pixel de l'image d'appartenir au thème courant. Enfin, une décision d'étiquetage est prise pour chaque pixel de l'image en intégrant les résultats de toutes les cartes d'appartenance de chaque thème. Une nouvelle classification est obtenue, permettant d'associer à chaque pixel une étiquette de la BD initiale. Cette classification est accompagnée d'une mesure de confiance dans la classification, dérivant des mesures d'appartenance par thème, qui permet de définir des zones d'incertitudes pouvant être étudiées à posteriori par un opérateur.

3 Expérimentation

Une vérité terrain de qualité est difficile à obtenir sur des bases de données couplées à des images satellites, ainsi nous avons décidé de générer une image à partir d'une image réelle (image Pléiades à 50 cm de résolution), pour laquelle on connaît *a priori* le thème d'appartenance (par rapport à une nomenclature donnée). L'*image reconstituée* est composée d'échantillons d'une image Pléiades réelle (Fig. 1a), étiquetés en 9 thèmes eux-mêmes comprenant 10 objets (Fig. 1b). Les objets (100 pixels de côté) sont répartis sur une grille régulière. La diagonale de l'image correspond à une zone non étiquetée dans la *BD initiale*, elle est composée de pixels pouvant appartenir à différents thèmes présents ou non dans la BD. Les 9 thèmes sélectionnés sont présentés dans la figure 1c. Cela correspond à une nomenclature très détaillée (thèmes aux apparences très proches) et à notre connaissance, aucun article n'a jusqu'à présent tenté d'aborder un tel discernement. Afin de comparer notre méthode à l'existant, et d'étudier l'influence du nombre d'objets par thème de la BD, trois méthodes ont été appliquées sur différentes *BD initiales* (de 1 à 9 objets par thème) : (a) un SVM multi-classes, (b) les RF (Breiman, 2001) et (c) notre méthode. Les résultats des classifications pour la BD contenant 9 objets par thèmes sont visibles sur les figure 1d,e & f. La courbe de l'évolution du score de bonne classification (pourcentage de pixels bien classés) en fonction du nombre d'objets présents dans la BD simulée est présentée dans la figure 2. Les résultats de notre méthode montrent qu'avec seulement 10% des objets dans la base, un taux de bonne classification supérieur à 70% des pixels est obtenu. Ce score atteint les 80% dès que l'apprentissage est réalisé sur 30% des objets. Les trois classifications sont visuellement proches. Cependant, on peut noter que cer-

tains thèmes sont confondus dans le cas des RF, alors qu'ils sont bien dissociés avec les SVM. Notre méthode donne des résultats moins bruités que les RF, mais plus confus que les SVM. L'évolution du score de bonne classification en fonction du nombre d'objets dans la BD initiale, pour chacune des trois méthodes comparées, est montrée sur la figure 2. Les résultats de notre méthode sont de meilleure qualité que les RF, quel que soit le nombre d'objets dans la BD initiale. Les SVM obtiennent des résultats en moyenne légèrement supérieurs, mais sont moins stables que notre méthode (e.g. score de bonne classification inférieur pour trois objets), en particulier quand le nombre d'objets est faible. Cette instabilité peut être expliquée par le choix des pixels d'apprentissage par tirage aléatoire. Notre méthode de sélection de l'ensemble d'apprentissage et de fusion des classifications par objet permet ainsi de réduire cette instabilité et donc d'améliorer la robustesse du processus, ce qui est particulièrement intéressant dans le contexte des BD d'occupation du sol, qui peuvent contenir des thèmes peu représentés et donc constitués de peu d'objets.

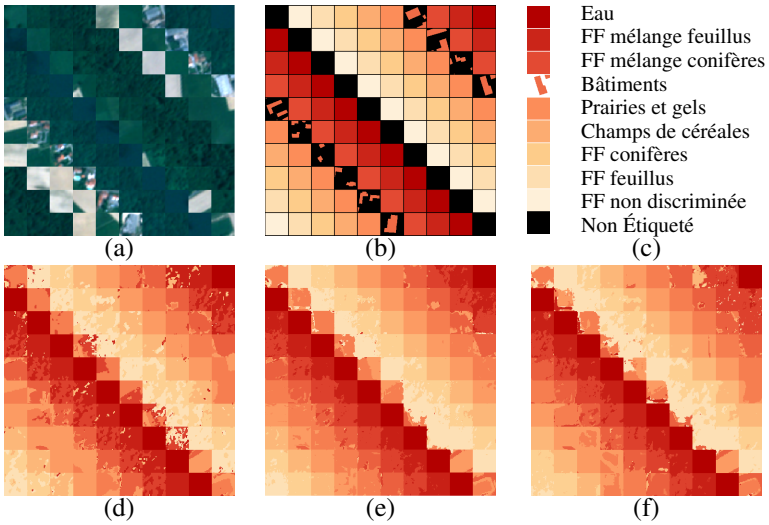

FIG. 1 – *(a) L'image reconstituée : textures provenant d'une image satellite réelle, (b) la vérité terrain et (c) la légende des thèmes composant la BD (FF = forêts fermée). Les résultats de la classification par : RF (d), SVM multi-classes (e) et notre méthode (f).*

4 Conclusion et perspectives

Dans cet article, nous avons présenté une méthode tirant profit d'une BD existante, pour apprendre les différentes apparences de chaque thème qui la compose. Ce processus a été appliqué avec succès à un jeu de données complexe simulé et comparé favorablement à plusieurs méthodes standards de classification supervisée. D'autre part, ce procédé a été développé dans un cadre plus général de détection de changements et de mise à jour de bases de données (Gressin et al., 2013), prouvant ainsi sa polyvalence. Nous envisageons maintenant d'améliorer

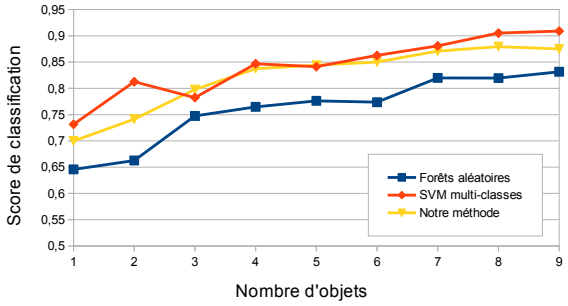

FIG. 2 – *Évolution du score de bonne classification en fonction du nombre d'objets présents dans la BD simulée, pour les RF ■, les SVM multi-classes ♦ et notre méthode ▼.*

notre méthode, en introduisant une étape de sélection d'attributs (Chouaib et al., 2012) et en optimisant la fusion des classifications par thème, et de tester notre méthode sur des données différentes, autant du point de vue image, que du point de vue base de données, afin d'étudier l'indépendance de notre méthode aux données en entrée.

Références

Breiman, L. (2001). Random forests. *Machine learning*, 1–35.

Chouaib, H., F. Cloppet, S.-A. Tabbone, et N. Vincent (2012). Combinaison de classificateurs simples pour une sélection rapide de caractéristiques. In *EGC*, pp. 459–471.

Gressin, A., N. Vincent, C. Mallet, et N. Paparoditis (2013). Semantic approach in image change detection. In *ACIVS, Poznan, Pologne*.

Mountrakis, G., J. Im, et C. Ogole (2011). Support vector machines in remote sensing : A review. *ISPRS Journal of Photogrammetry and Remote Sensing 66*(3), 247–259.

Pal, M. (2008). Ensemble of support vector machines for land cover classification. *International Journal of Remote Sensing 29*(10), 3043–3049.

Summary

Modern classification methods allow to classify unlabelled areas of databases with geospatial image processing techniques, but suffer from a serious lack of robustness and versatility. In this paper, we propose a robust method that alleviates such problems and subsequently allows to extend database labels on unknown areas using a very high resolution satellite image. Our method is based on a hierarchical inspection of the database. First, feature selection is performed among a large attribute set for each theme of the database. Then, a carefully tuned training pixel selection and a classification are carried for each object of each theme. Finally, the labelling decision is taken after merging all the classifications by theme. Our method is successfully applied to a simulated database and favorably compared to different standard methods.