Apprentissage de connaissances structurelles à partir de cartes et classification multi-classes : Application à la mise a jour de cartes d'occupation du sol

Meriam Bayoudh, Emmanuel Roux, Richard Nock, Gilles Richard

Apprentissage de connaissances structurelles à partir de cartes et classification multi-classes: Application à la mise à jour de cartes d'occupation du sol

Meriam Bayoudh[1,4]
Emmanuel Roux[1]
Richard Nock[2,4]
Gilles Richard[3]

[1] ESPACE-DEV, UMR228 IRD/UM2/UR/UAG, Institut de Recherche pour le Développement
[2] Centre d'Etude et de Recherche en Economie, Gestion, Modélisation et Informatique Appliquée (CEREGMIA)
[3] Institut de Recherche en Informatique de Toulouse (IRIT - UMR 5505) - Université Paul Sabatier
[4] Université des Antilles et de la Guyane
meriam.bayoudh@ird.fr
emmanuel.roux@ird.fr
Richard.Nock@martinique.univ-ag.fr
richard@irit.fr

Résumé :
Le nombre de satellites et de capteurs pour la télédétection dédiés à l'observation de la Terre ne cesse d'augmenter, permettant ainsi d'avoir une masse de données importante en particulier en matière d'images. Parallèlement, un effort permanent vise, d'une part, à améliorer l'accès à ces données et, d'autre part, à développer d'avantages d'outils pour les manipuler. De tels efforts sont particulièrement utiles dans des contextes socio-environnementaux très dynamiques spatialement et temporellement, pour lesquels il est nécessaire de suivre et de prédire les événements environnementaux et sociétaux. En revanche, en présence d'un tel flux de données, nous avons besoin de méthodes automatiques d'interprétation d'images. Une solution envisageable pour répondre à ce besoin est de bénéficier des atouts de l'intelligence artificielle pour l'obtention de cartes d'occupation du sol issues d'une classification des régions des images. Afin de contribuer à l'automatisation de la classification, nous proposons une méthode d'induction de règles interprétables par des non-experts et mettant en évidence, explicitement, des connaissances structurelles. Cette méthode s'appuie sur la programmation logique inductive (PLI) et en particulier sur le système inductif "Aleph". L'application de la méthode de classification *Multi-class Rule Set Intersection* (MRSI) permet ensuite de classifier tout nouvel objet au regard des ses caractéristiques intrinsèques et de celles des objets environnants. Nous avons appliqué notre méthodologie à l'étude de la dynamique du littoral de la Guyane Française. Suite à ce travail, nous avons induit 136 règles de classification pour 38 classes d'occupation du sol. Ces règles sont intelligibles et simples à interpréter de par l'utilisation de la logique du premier ordre. Les performances du système ont été évaluées par la validation croisée. En moyenne, la précision, la spécificité et la sensibilité sont, respectivement, égales à 84,62%, 99,57% et 77,22%. Ces résultats quantitatifs montrent une bonne performance de la méthodologie pour la mise à jour automatique de cartes d'occupation du sol et/ou l'assistance aux opérateurs utilisant l'analyse d'image orientée-objet.

Mots-clés : Programmation Logique Inductive (PLI), Apprentissage, Télédétection, Système d'Information Géographique (SIG), Cartes d'occupation du sol.

1 Introduction

La disponibilité et l'usage des données de télédétection augmentent continuellement, notamment pour la recherche et l'aide à la décision dans le cadre des politiques

publiques. Cette augmentation de la masse de données provient du nombre grandissant de satellites et de capteurs dédiés à l'observation de la terre et des assouplissement des politiques de distribution de données, les pays et / ou les organisations qui distribuent les données de télédétection gratuitement devenant de plus en plus nombreux. Ce flux d'informations crée, alors, des nouveaux défis pour les ingénieurs et les chercheurs qui s'intéressent au traitement et l'interprétation des données. Par conséquence, des efforts sont nécessaires afin de concevoir de nouvelles approches permettant une mise à jour automatique des cartes d'occupation du sol dans l'optique, notamment, d'extraire de l'information utile pour une meilleur prise de décision.
Dans ce contexte, certains travaux ont vu le jour qui visent à formaliser, représenter et exploiter les connaissances expertes pour une classification et interprétation automatiques des images. Les études récentes s'orientent en particulier vers l'utilisation des ontologies. Par exemple, dans Hudelot *et al.* (2008), une ontologie des relations spatiales permettant de guider l'interprétation des images est proposée. Cette ontologie est, alors, enrichie par une représentation flou des concepts. Deux autres travaux (Durand *et al.* (2007); Andres *et al.* (2012)) proposent des approches automatiques de classification d'images pour l'interprétation de celles-ci, au travers d'ontologies.

Une approche complémentaire à celles basées sur la formalisation des connaissances expertes est l'extraction de connaissances à partir de données. La grande majorité des méthodes permettant la classification supervisée des images satellites ne considèrent que l'information associée aux pixels à l'intérieur de régions de l'image appartenant à la même classe afin d'apprendre les signatures des classes. Les aspects structurels sont, essentiellement, pris en compte en calculant des indices de texture à l'intérieur de ces mêmes régions. A nos connaissances, il n'existe aucun outil qui permet de trouver des règles de classification générales et efficaces et qui permettent d'exprimer des connaissances structurelles de haut niveau sémantique.
Dans la littérature scientifique, rares sont les études qui ont été proposées pour l'apprentissage de connaissances structurelles à partir des cartes existantes. Dans Malerba *et al.* (2003), les auteurs proposent une approche pour aider à l'interprétation des cartes topographiques. Leur système, appelé *INductive GEographic iNformation System* (INGENS) intègre, à la fois, des outils d'apprentissage et des fonctionnalités d'un Système d'Information Géographique (SIG). Ce système permet l'extraction des caractéristiques et des concepts pertinents à partir d'une base de données spatiale en utilisant des propriétés classiques d'un SIG. Le système inductif intégré permet quant à lui de trouver des règles pour la reconnaissance de contextes géographiques complexes définis par la présence d'objets géographiques élémentaires et leur organisation spatiale.

Dans Vaz *et al.* (2007), les auteurs utilisent un système d'induction logique appelé *APRIL* (Fonseca *et al.* (2006)) pour apprendre des règles de classification à partir, d'une part, d'une carte détaillée fournie par des botanistes et, d'autre part, de cartes issues du projet *Corine Land Cover* (CLC) de la même zone. Ces règles sont destinées à désagréger automatiquement les informations fournies par CLC, jugées trop génériques dans le cadre applicatif donné. L'apprentissage inductif à partir des caractéristiques structurelles de l'occupation du sol et des informations sur les feux de forêt passés a également été utilisé pour la prédiction de feux de forêt qui dépendent, notamment, des caractéristiques du paysage (Vaz *et al.* (2010)).

Enfin, afin de mieux prendre en compte la dimension spatiale dans l'étude de la contamination des crustacés dans la lagune du bassin de Thau, Chelghoum *et al.* (2006)

utilisent la PLI à partir de données géographiques.

Dans ce contexte, notre travail consiste à une mise en oeuvre d'une méthode d'apprentissage de connaissances structurelles et symboliques à partir de cartes d'occupation de sol et de différentes couches d'informations géographiques complémentaires. Cette étude s'appuie sur les travaux préliminaires présentés par Bayoudh *et al.* (2012). Elle présente des avancées méthodologiques et applicatives significatives, notamment en ce qui concerne la classification multi-classes et les résultats associés.

Nous avons choisi la Programmation Logique Inductive (PLI) (Muggleton (1991)) pour la mise en oeuvre de la méthode d'apprentissage, de par la clarté du langage utilisé et l'intelligibilité des règles induites. Ensuite, nous avons appliqué l'approche récemment proposée par Abudawood & Flach (2011) et appelée *Multi-class Rule Set Intersection* (MRSI) afin d'attribuer, dans un contexte multi-classes, un nouvel objet à une unique classe. Enfin, nous avons appliqué la méthodologie proposée à la mise à jour de cartes d'occupation du sol du littoral Guyanais.

Notre article est organisé comme suit : la section 2 présente la méthodologie générale, avec l'introduction de la PLI, la présentation des méthodes d'extraction et de codage de l'information géographique, la description de la méthode de classification multi-classes et des procédures d'évaluation. La section 3 décrit l'application de la méthodologie à la mise à jour des cartes d'occupation de sol du littoral Guyanais. Dans la section 4, nous présentons les résultats que nous discutons dans la section 5.

2 Méthodologie générale

2.1 Programmation Logique Inductive : généralités

La Programmation Logique Inductive (PLI) (Muggleton (1991)) combine, à la fois, les notions d'apprentissage et de la programmation logique. Cette technique permet d'apprendre une théorie générale H à partir d'une base de connaissances B et une base d'exemples E dans un formalisme à base de clauses logiques.
Malgré la simplicité de son principe, la PLI permet de modéliser des problèmes complexes, ce qui explique son utilisation dans plusieurs domaines. Elle a, ainsi, été utilisée en chimie (Blockeel *et al.* (2004)), en biologie, en physique, en médecine (Luu *et al.* (2012); Fromont *et al.* (2005)), en écologie et en bio-informatique (Santos *et al.* (2012); Lavrac & Dzeroski (1994); Srinivasan *et al.* (1996)). Elle a, également, été utilisée pour modéliser la prise de décision au jeu d'échec (Goodacre (1996)) et pour étudier la qualité des eaux de la rivière (Cordier (2005)). En revanche, rare sont les applications qui prennent en compte l'information géographique (Malerba *et al.* (2003); Vaz *et al.* (2007, 2010); Chelghoum *et al.* (2006)).
Plus formellement, le principe de la PLI est le suivant (Lavrac & Dzeroski (1994)) :
Etant donnés :

- Une base de connaissances B exprimée en logique du premier ordre décrivant un ensemble de connaissances et de contraintes ;
- Un ensemble d'exemples E, divisé en deux sous-ensembles, E^+ et E^- correspondant, respectivement, aux ensembles des exemples positifs et négatifs.
- Un langage de description L.

Trouver :
Une théorie H exprimée en logique du premier ordre en utilisant le langage de description L qui doit couvrir les exemples positifs E^+ et ne pas couvrir les exemples négatifs

E^-.
Parmi les systèmes inductifs existants, nous avons opté pour *Aleph* (Srinivasan (2007)), un système gratuit développé en Prolog, utilisant la méthode de recherche "top-down" et basé sur l'implication inverse (Muggleton (1995)).

2.2 Extraction et codage de l'information géographique

A partir d'une carte d'occupation du sol, chaque entité géographique élémentaire, appartenant à une seule classe d'occupation du sol, est par la suite appelé *"objet"*. Ce dernier constitue l'entité élémentaire sur laquelle porte le raisonnement. Ces objets permettent, ainsi, de définir les exemples d'apprentissage et de test.
Un ensemble de prédicats permettent de caractériser, pour chaque objet, les propriétés intrinsèques (classe d'occupation du sol, surface, dimension fractale, compacité, périmètre, latitude, longitude) et les relations avec les autres objets (adjacence, inclusion, positions relatives suivant les dimensions de latitude et de longitude) (*cf.* Tableau 1).

La PLI étant adaptée aux informations symboliques, les variables quantitatives ont été discrétisées et l'information recodée de la manière suivante : pour toute variable quantitative V, les 10ème, 20ème, ..., 90ème percentiles de la distribution empirique de V, notés p_k ($k \in [1, 9]$), sont calculés. Puis, pour tout p_k, nous définissons deux prédicats permettant de recoder l'information en indiquant si une valeur X de V est soit inférieure ou égale soit supérieure à p_k. Par exemple, l'aire X d'un objet O est recodée, pour le percentile k, de la manière suivante :

`aire_symb(`O`,`I_k`) :- aire_num(`$O1$`,X),` $X \leq p_k$
ou
`aire_symb(`O`,`S_k`) :- aire_num(`$O1$`,X),` $X > p_k$

Les valeurs absolues de latitude et de longitude sont quant à elles recodées en indiquant les positions relatives des objets deux à deux (*cf.* Tableau 1).

Prédicat	**Signification**
`object(`O`)`	Déclaration de l'objet O
`classe(`O`,label_classe)`	L'objet O appartient à la classe d'occupation de sol `label_classe`
`adjacent(`$O1$`,`$O2$`)`	Les objets $O1$ et $O2$ sont adjacents
`inclus(`$O1$`,`$O2$`)`	L'objet $O2$ est inclus dans l'objet $O1$
`contient(`O`,`E`)`	L'objet O contient l'entité E, (*e.g.* $E \in \{$*bâti*, *rivière*, $route, ...\}$)
`aire_num(`O`,`X`)`$\star$	O a, respectivement, une aire (m^2), une compacité, une dimension fractale et un périmètre (m) de X, avec $X \in \Re$
`compacité_num(`O`,`X`)`$\star$	
`dim_fract_num(`O`,`X`)`$\star$	
`périmètre_num(`O`,`X`)`$\star$	
`aire_symb(`O`,`I_k^{aire} `ou` S_k^{aire}`)`	O a une aire, une compacité, une dimension fractale et un périmètre appartenant, pour chaque $k^{\text{ème}}$ percentile de chaque variable, à l'intervalle I_k^V ou S_k^V
`compacité_symb(`O`,`I_k^{comp} `ou` S_k^{comp}`)`	
`dim_fact_symb(`O`,`I_k^{df} `ou` S_k^{df}`)`	
`périmètre_symb(`O`,`I_k^{per} `ou` S_k^{per}`)`	
`lat(`O`,`X`)`$\star$	O a une latitude de X ($X \in \Re$)
`long(`O`,`X`)`$\star$	O a une longitude de X ($X \in \Re$)
`north(`$O1$`,`$O2$`) :- lat(`$O1$`,A),lat(`$O2$`,B),A>B.`	L'objet $O1$ est au nord de l'objet $O2$
`south(`$O1$`,`$O2$`) :- lat(`$O1$`,A),lat(`$O2$`,B),A`$\leq$`B.`	L'objet $O1$ est au sud de l'objet $O2$
`east(`$O1$`,`$O2$`) :- long(`$O1$`,A),long(`$O2$`,B),A>B.`	L'objet $O1$ est à l'est de l'objet $O2$
`west(`$O1$`,`$O2$`) :- long(`$O1$`,A),long(`$O2$`,B),A`$\leq$`B.`	L'objet $O1$ est à l'ouest de l'objet $O2$

TABLE 1 – Liste complète des prédicats utilisés pour la caractérisation des objets. Un astérisque indique que le prédicat n'est pas utilisé dans les prémisses des règles

2.3 Induction des règles : approche *one-vs-rest*

Les phases d'extraction et de codage de l'information terminées, les règles de classification sont induites en utilisant le système inductif Aleph. Dans le cas où les objets sont répartis dans plus de deux classes, chaque objet appartenant à une et une seule classe (cadre multi-classes), la PLI est habituellement appliquée au travers de l'approche *one-vs-rest* (Abudawood & Flach (2011)). Cette méthode consiste à générer autant de classifieurs que de classes, en constituant, pour chaque classe c, les ensembles d'exemples positifs et négatifs comme suit :

$E^+ = \{O/\texttt{classe}(O, c)\}$
et
$E^- = \{O/\texttt{classe}(O, \overline{c})\}$

2.4 Classification dans le cadre multi-classes

L'approche *one-vs-rest* génère autant de classifieurs que de classes. Le problème est par conséquent d'attribuer une et une seule classe à un nouvel objet si les différents classifieurs sont utilisés de façon indépendante. Dans Abudawood & Flach (2011), les

auteurs proposent différentes solutions, dont la méthode *Multi-class Rule Set Intersection* (MRSI), la plus performante dans le cadre multi-classes et définie comme suit : i) les théories induites pour chaque classe sont rassemblées en un seul ensemble de règles ; ii) les ensembles C_i des exemples couverts par chaque règle r_i sont stockés ; iii) une règle par défaut est établie, qui concerne la classe majoritaire des exemples non couverts ; iv) pour un nouvel exemple, l'intersection I des ensembles d'exemples couverts par les règles déclenchées est calculée ($I = \cap C_i / r_i$ déclenchée), et enfin, v) la classe prédite est la classe majoritairement présente dans l'ensemble I.

2.5 Evaluation quantitative

L'évaluation des résultats est réalisée au travers des valeurs de précision, sensibilité et spécificité de classification, par validation croisée stratifiées avec 10 sous-ensembles d'apprentissage et de test associés (*stratified ten-fold cross validation*) : pour chaque sous-ensemble d'apprentissage, les règles sont induites puis appliquées à l'ensemble de test correspondant, permettant d'obtenir une table de contingence multi-classes (*cf.* Figure 1). La précision globale est calculée comme suit (Abudawood & Flach (2011)) :

$$\textit{Précision globale} = \sum_{1}^{n} \frac{VP^{(i)}}{E}, \tag{1}$$

avec n correspondant au nombre des classes, $VP^{(i)}$ au nombre de vrais positifs pour la classe i et E au nombre total des exemples de test.

		Classes prédites							
		C_1	...	C_{i-1}	C_i	C_{i+1}	...	C_n	Total
Classes Actuelles	C_1	$VN_1^{(i)}$	...	...	$FP_1^{(i)}$	...	...	...	E_1
	...	...	...	...	...	...	...	...	...
	...	...	...	$VN_{i-1}^{(i)}$	$FP_{i-1}^{(i)}$	...	...	...	E_{i-1}
	C_i	$FN_1^{(i)}$	...	$FN_{i-1}^{(i)}$	$VP^{(i)}$	$FN_{i+1}^{(i)}$	...	$FN_n^{(i)}$	E_i
	...	...	...	...	$FP_{i+1}^{(i)}$	$VN_{i+1}^{(i)}$	...	...	E_{i+1}
	...	...	...	...	...	...	...	...	...
	C_n	...	...	...	$FP_n^{(i)}$	...	...	$VN_n^{(i)}$	E_n
	Total	$\hat{E}_1$	...	$\hat{E}_{i-1}$	$\hat{E}_i$	$\hat{E}_{i+1}$	...	$\hat{E}_n$	E

FIGURE 1 – Table de contingence obtenue pour un sous-ensemble de test. Les notations (VP : Vrai Positif ; VN : Vrai Negatif ; FP : Faux Positif ; FN : Faux Negatif) sont celles associées à la classe i

Pour une classe i donnée, la formule utilisée pour calculer la sensibilité est la suivante :

$$Sensibilite^{(i)} = \frac{VP^{(i)}}{VP^{(i)} + \sum_{j=1, j\neq i}^{n} FN_j^{(i)}} = \frac{VP^{(i)}}{E_i} \tag{2}$$

avec $VP^{(i)}$ le nombre de vrais positifs pour la classe i et $FN_j^{(i)}$ le nombre de faux négatifs pour la classe i, associés d'une façon incorrecte à la classe j.

La spécificité est calculée en utilisant la formule suivante :

$$Specificite^{(i)} = \frac{\sum_{j=1, j\neq i}^{n} VN_j^{(i)}}{\sum_{j=1, j\neq i}^{n} VN_j^{(i)} + \sum_{j=1, j\neq i}^{n} FP_j^{(i)}} \tag{3}$$

$VN_j^{(i)}$ correspondant au nombre de vrais négatifs pour la classe i, attribués correctement à la classe j, $FP_j^{(i)}$ correspond au nombre de faux positifs de la classe i qui appartiennent effectivement à la classe j.

Nous avons calculé, finalement, l'indice de kappa de Cohen (Cohen (1960)) pour chaque sous-ensemble de test. Cet indice correspond à une mesure statistique de la concordance entre deux classifications (dans notre cas, entre la classification effective ou réelle et la classification prédite par la méthode MRSI). Le coefficient Kappa se calcule en appliquant la formule suivante :

$$kappa = \frac{P(A) - P(H)}{1 - P(H)} \tag{4}$$

Avec $P(A)$ correspond à la proportion d'accords entre les deux classifications et $P(H)$ à la proportion de cas où, selon la théorie des probabilités, on peut s'attendre à un accord simplement dû au hasard. La valeur de cette mesure varie entre -1 et 1 (Anthony & Garrett (2005)) : le cas idéal (kappa=1) correspond à une concordance parfaite ; une valeur nulle indique des résultats de classification identique à ceux obtenus par hasard ; enfin une valeur égale à -1 correspond à un désaccord total entre les deux classifications.

3 Application de la méthodologie pour la mise à jour de carte d'occupation du sol

Cette partie présente l'application des concepts et méthodes définis précédemment à une situation réelle. Le territoire Guyanais est soumis à des dynamiques anthropiques et naturelles intenses (Edward *et al.* (2010)) : phénomènes d'érosion et d'accrétion cycliques des côtes dûe au transport, par les courants marins, des sédiments provenant du fleuve Amazone ; extension des zones urbaines et péri-urbaines ; des zones agricoles, etc. Dans ce contexte, il est nécessaire de développer des méthodes automatiques de suivi de l'occupation du sol du territoire Guyanais, en exploitant notamment les données de télédétection (photographies aéroportées et images satellitaires). Si la mise à jour des contours des objets géographiques est envisageable par des opérateurs et ne demande pas un niveau d'expertise très important, la mise à jour de l'appartenance de ces objets aux classes d'occupation du sol est, malgré les efforts de formalisation et de normalisation des procédures, autrement plus complexe et subjective et demande une connaissance fine des différents types d'occupation du sol, tant dans le domaine de l'image que sur le terrain. En appliquant la méthode d'apprentissage présentée précédemment à la mise à jour des classes d'occupation du sol, nous espérons contribuer à définir une méthode automatique, efficace, objective et reproductible pour le suivi de l'occupation du sol Guyanais.

3.1 Description des données

Nous disposons d'une série de 3 cartes d'occupation de sol correspondant aux années 2001, 2005 et 2008 et présentant 39 classes. La nomenclature de la classification est basée sur la nomenclature Européenne *CORINE Land Cover* (CLC) adaptée au contexte Amazonien par l'ajout de 15 classes, dont 9 classes correspondant à différents types de forêt.

Les cartes ont été produites et nous ont été fournies par l'Office National des Forêts (ONF). Les cartes correspondant aux années 2001 et 2005 ont été réalisées par photo-interprétation des photographies aériennes d'une résolution spatiale de 50 cm et issues de la BD-Ortho® (produit de l'Institut Géographique National, IGN). La carte d'occupation de sol de l'année 2008 a été obtenue grâce à la mise à jour des cartes antérieures en utilisant des images du satellite Spot 5 à une résolution spatiale de 2.5 mètres. Ces images ont été fournies grâce au projet *SEAS-Guyane* [1].

Deux couches d'information géographiques complémentaires ont été ajoutées (*cf.* Figure 2) :

1. Le réseau routier, fourni par la BD-Cartho® qui est une base de données cartographiques de référence de l'IGN ;
2. Le réseau hydrographique fourni par la BD-Carthage® (produit du ministère chargé de l'environnement et de l'IGN) qui constitue le référentiel hydrographique français et qui a été produite en 2009 pour le territoire Guyanais, par la Direction de l'Environnement, de l'Aménagement et du Logement (DEAL) de la Guyane et l'Office National de l'Eau et des Milieux Aquatiques (ONEMA).

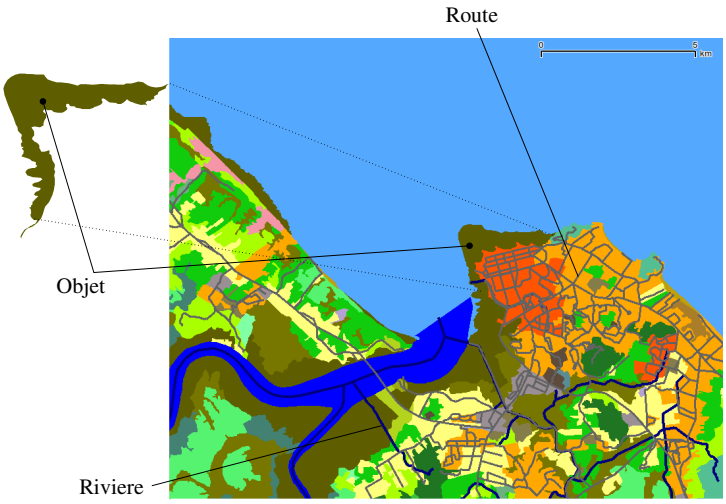


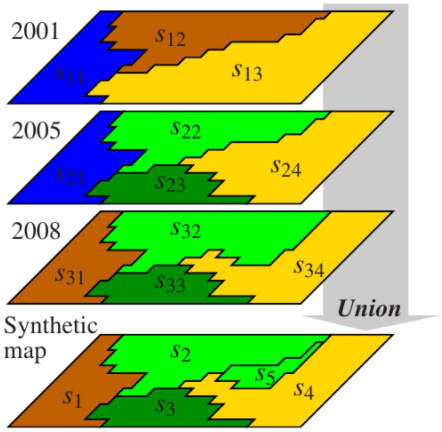
FIGURE 2 – Représentation d'une partie des cartes d'occupation du sol et des réseaux routier et hydrographiques. Sources : Office National des Forêts (ONF), Institut Géographique National (IGN) ; ministère français chargé de l'environnement ; Direction de l'Environnement, de l'Aménagement et du Logement (DEAL) de la Guyane ; Office National de l'Eau et des Milieux Aquatiques (ONEMA). *cf.* §3.1 pour plus de détails.

1. https ://www.seas-guyane.org

3.2 Extraction et codage de l'information géographique appliqués aux données du littoral Guyanais

3.2.1 Prétraitements, définition des objets et codage de l'information

Tout d'abord, nous avons complété la carte d'occupation de sol initiale en ajoutant 3 classes : *Ocean*, *River* et *Unknown*. Les deux premières classes ont été ajoutées car elles contribuent significativement à la structuration de l'environnement du territoire Guyanais. Nous avons défini explicitement la classe *Unknown* afin de prendre en compte les informations manquantes sur les cartes des années 2001 et 2005.
Nous avons, ensuite, produit une carte synthétique en fusionnant les informations des 3 cartes par le biais de l'opérateur "union" des SIG. L'entité géographique élémentaire de la carte résultante définie l'*objet* sur lequel portera l'apprentissage et le raisonnement de classification. Chaque objet appartient ainsi à une seule classe à une date donnée, comme le montre la figure 3.



Synthetic map attribute table

Object	Class (2001)	Class (2005)	Class (2008)[1]
s_1	blue	blue	brown
s_2	brown	green1	green1
s_3	orange	green2	green2
s_4	orange	orange	orange
s_5	orange	orange	green1

[1] Classes the membership of which is to be learned

FIGURE 3 – Illustration de la conception de la carte synthétique combinant les 3 cartes initiales

Compte tenues des caractéristiques diachroniques des données, nous avons défini 3 prédicats pour indiquer la classe d'un objet en fonction de l'année : `classeA(`O`,label_classe)` indique la classe d'occupation du sol de l'objet O pour l'année A.
Le prédicat cible (*i.e.* le concept devant être appris) est la classe d'occupation du sol à laquelle appartiennent les objets de la carte synthétique en 2008. Nous n'avons pas considéré les classes *Ocean* , *River* et *Unknown* pour l'apprentissage. Enfin, nous avons éliminé la classe "Rizière", cette dernière étant sous-représentée sur le territoire (deux objets seulement appartenant à cette classe). Ainsi, 38 classes d'occupation du sol sont considérées dans notre application, chacune devant faire l'objet d'un apprentissage compte tenu de l'approche *one-vs-rest* utilisée (*cf.* liste des classes dans le tableau 2).
En outre, étant données les deux couches d'informations complémentaires choisies, le prédicat `contient(`O`,`E`)` se rapporte aux entités *rivière* et *route* ($E \in \{$*rivière*, *route*$\}$).
L'extraction des objets et de leurs propriétés a été effectuée avec le système d'information géographique gratuit et libre GRASS (Team (1999 2012)).

3.2.2 Paramétrage du système inductif Aleph

Lors du processus d'induction, les clauses candidates sont déclarées admissibles si elles présentent une précision supérieure ou égale à la valeur 0,7 considérée comme un bon compromis afin d'assurer à la fois la généricité et la précision des règles finales. Cette précision est définie par *précision* $= p/(p+n)$ avec p et n les nombres d'exemples, respectivement positifs et négatifs, couverts par la clause. Elle diffère par conséquent de la précision globale définie au §2.5 et qui évalue la précision de classification globale à l'issue de l'apprentissage.
De plus, nous avons limité à 5 le nombre de littéraux dans les prémisses des règles, assurant ainsi l'induction de règles facilement intelligibles et interprétables par les utilisateurs finaux.

4 Résultats

4.1 Caractéristiques des règles induites

Suite au processus d'induction, nous avons obtenu 136 règles de classification pour 38 classes d'occupation du sol. Ces règles peuvent couvrir entre 2 et 692 exemples positifs et de 0 à 212 exemples négatifs.
Ci-dessous, nous citons quelques règles induites. Entre crochets sont indiqués le nombre d'exemples positifs (*Pos.*) et le nombre d'exemples négatifs (*Neg.*) couverts par la règle, ainsi que le nombre total d'exemples positifs (*Total pos.*) concernant le prédicat cible considéré (*i.e.* le nombre total d'objets de l'ensemble d'apprentissage appartenant à la classe considérée).

```
(1) [Pos. = 472 Neg. = 88 Total pos. = 552]
  classe08(A,Habitat multidisciplinaire):-
  aire_symb(A,<=165566.67), adjacent(A,B),
  classe05(B,Habitat multidisciplinaire).
(2) [Pos. = 575 Neg. = 212 Total pos. = 814]
  classe08(A, Systèmes culturaux et parcellaires complexes
  (abattis)):- adjacent(A,B),
  classe05(B,systèmes culturaux et parcellaires complexes
  (abattis)),
  aire_symb(A, <=165566.67).
(3) [Pos. = 3 Neg. = 1 Total pos. = 166]
  classe08(A,Tissu urbain discontinu):- adjacent(A,B),
  classe01(B,Tissu urbain continu),compacité
  _symb(B,<=1.31).
```

La première règle signifie qu'un objet A appartient à la classe *Habitat multidisciplinaire* en 2008 si : **i)** l'objet A a une aire inférieure ou égale à 165566.67 m^2, **ii)** est adjacent à un objet B qui appartenait à la même classe 3 ans auparavant.

4.2 Performance de prédiction

Le tableau 2 résume les résultats de la sensibilité obtenus pour chaque classe à l'issue de la validation croisée présentée au §2.5. La spécificité est quant à elle égale à 100% pour toutes les classes, exceptée pour la classe *Forêt et végétation arbustive en mutation* qui présente une spécificité de 83,1%.

TABLE 2 – Résultats de la sensibilité pour chaque classe d'occupation du sol du littoral Guyanais

Intervalles de sensibilité (nombre de classes concernées et pourcentage par rapport au nombre total de classes)	**Classes d'occupation du sol et sensibilité associée, en %**
$\leq$50% (5 classes, soit 13,1%)	Plages, dunes, sables et vase (5,0) ; Décharges (25,0) ; Chantiers (30,1) ; Territoires principalement occupés par l'agriculture avec présence de végétation (abattis itinérant) (41,1) ; Forêts basses sur sable blanc (41,7)
$>$50% et $\leq$80% (20 classes, soit 31,6%)	Réseaux routiers et réseaux de communication et espaces associés (56,9) ; Forêts dégradées de terre ferme (60,3) ; Extraction de matériaux (63,5) ; Zones industrielles ou commerciales (65,0) ; Prairies (67,9) ; Terres arables hors périmètres d'irrigation (70,0) ; Forêts littorales sur rochers (70,0) ; Espaces verts artificialisés non agricoles (75,0) ; Forêts hautes (76,4) ; Forêts de la plaine côtière ancienne (79,9) ; Zones portuaires (80,0) ; Forêts inondables ou marécageuses dégradées (80,0)
$>$ 80% (21 classes, soit 35,3%)	Vergers et petits fruits (87,1) ; Plantations forestières (81,7) ; Systèmes culturaux et parcellaires complexes (abattis) (81,9) ; Forêts sur cordons sableux (82,0) ; Pisciculture et autres bassins (85,0) ; Tissu urbain discontinu (87,9) ; Marais maritimes (88,9) ; Forêts inondées ou marécageuses (91,7) ; Savanes inondables ou inondées (92,0) ; Marais intérieurs et marécages boisés (92,6) ; Tissu urbain continu (93,0) ; Mangroves (93,0) ; Savanes sèches (93,9) ; Habitat pluridisciplinaire (94,4) ; Forêts basses (98,0) ; Bâti isolé (95,3) ; Aéroports (100,0) ; Roches nues, savane roche (100,0) ; Forêts et végétation arbustive en mutation (100,0) ; Marécages ripicoles (100,0) ; Plans d'eau (100,0)

La précision varie quant à elle entre 82.4% et 87.34% en fonction des sous-ensembles de test, avec une moyenne sur les 10 sous-ensembles égale à 84.62%.
Enfin, la valeur de Kappa varie entre 0.69 à 0.77 avec une moyenne égale à 0.7.

5 Discussion

Le nombre des règles induites est relativement élevé. En revanche, il est variable d'une classe à l'autre : 20 règles ont été obtenues pour la classe *Forêt et végétation arbustive en mutation* alors qu'une unique règle a été induite pour la classe *Forêt inondées ou marécageuses*.

Du point de vue qualitatif, les règles induites semblent en accord avec les connaissances de l'environnement concernant la zone étudiée. Ces règles sont de plus intelligibles et d'interprétation facile, même par un non-expert en apprentissage auto-

matique. En revanche, il existe des règles très spécifiques couvrant un nombre très faible d'exemples positifs (seulement 2 ou 3 exemples) par rapport au nombre total d'exemples positifs pour la classe considérée (*cf.* règle (3) dans les exemples du §4.1). Les prédicats *south*, *north* and *west* sont absents des règles induites. Cela montre que ces prédicats ne sont pas pertinents pour la discrimination des classes, montrant ainsi que la caractérisation des objets n'est pas nécessairement pertinente. Il serait possible de mieux caractériser les objets en exploitant plus, et de manière systématique, les connaissances expertes. En particulier, les ontologies de domaine, qui bénéficie d'effort de recherche de plus en plus importants, pourraient guider le processus d'apprentissage en spécifiant les prédicats et les contraintes à utiliser.
Malgré la longueur maximale des prémisses initialement fixée à 5 littéraux, cette longueur est au plus égale à 3 dans les résultats obtenus. Cela peut être expliqué par la limitation du nombre de noeuds à explorer (par défaut 5000) dans l'espace de recherche afin de trouver une clause "acceptable". Pour certaines classes, le processus de recherche est par, conséquent, stoppé avant que le système ait exploré la totalité de l'espace de recherche.
En observant les valeurs de sensibilité, nous remarquons que les classes pour lesquelles cette valeur est élevée subissent peu (voir aucun) changement dans le temps. Autrement dit, un objet connu pour appartenir à une telle classe dans le passé, a de très fortes chances d'appartenir à cette même classe dans le présent et l'avenir. Ces objets correspondent, en particulier, aux zones anthropisées associées aux classes *Aéroport* et *Bâti isolé*, ou à des types d'occupation du sol naturels mais stables dans le temps et ne pouvant être exploités, de par des contraintes naturelles et/ou légales les concernant. Il s'agit par exemple des classes *Roches nues, Savane roche*, *Marais maritimes*, *Plan d'eau*.

Contrairement aux classes précédemment décrites, certaines classes présentent des sensibilités faibles. Les objets correspondant à ces classes d'occupation du sol subissent pour certains des changements continuels relativement rapides. Il s'agit, en particulier, des objets associés à la classe *Vase ou sable* (Edward *et al.* (2010)), mais également aux classes *Chantiers* et *Territoires principalement occupés par l'agriculture avec la présence de la végétation non cultivées*, qui constitue une classe complexe correspondant notamment à l'agriculture itinérante sur brûlis consistant à cultiver une surface puis à laisser la végétation naturelle se régénérer. Pour ces classes, il semble donc l'information dont nous disposions soit insuffisante en terme d'antériorité et de résolution temporelle.
Outre les résultats prometteurs de sensibilité et de spécificité, nous avons obtenu des valeurs élevées pour les précisions globales et pour l'indice de concordance Kappa. Selon la table d'interprétation de kappa proposée par (Richard & Koch (1977)), ces résultats correspondent à un "accord fort" entre les classes prédites et les classes effectives.
Enfin, d'un point de vue méthodologique, la programmation logique inductive traite des données symboliques. La prise en compte de l'information numérique en PLI constitue un champ de recherche à part entière. Dans notre cas, le codage proposé au §2.2 semble réaliser un bon compromis, lors de la phase d'apprentissage, entre perte d'information et capacité de généralisation. En particulier, en comparaison avec une discrétisation classique en plusieurs classes de valeurs, il permet de produire des résultats plus généraux et facilite la découverte d'intervalles de valeurs discriminants.

6 Conclusion

Nous avons proposé, dans ce papier, une méthode d'induction de règles de classification intégrant explicitement des connaissances structurelles, dans l'optique de concevoir automatiquement des cartes d'occupation du sol. Nous avons opté pour la programmation logique inductive et une méthode de classification adaptée au cadre multi-classes. Nous avons appliqué la méthode proposée à la mise à jour de cartes d'occupation du sol du littoral Guyanais.
Les résultats obtenus montrent que les règles de classification induites sont facilement intelligibles et interprétables, et qu'elles permettent effectivement de mettre en évidence des connaissances structurelles. L'évaluation des performances par le biais de la validation croisée donne des résultats prometteurs permettant d'envisager la mise à jour automatiquement des classes d'occupation du sol ou, tout au moins, l'aide aux opérateurs utilisant l'analyse d'images orientée objets. Cette méthode d'analyse d'images permet d'intégrer, dans le processus de classification des objets, des connaissances concernant les relations spatiales entre objets. A notre connaissance, les logiciels actuels utilisant cette "approche objet" ne fournissent aucune aide à l'utilisateur pour spécifier des règles de classification à la fois performantes et générales (et par conséquent reproductibles). Dans nos futurs travaux, nous envisageons d'exploiter les ontologies de domaines afin de guider l'apprentissage, en alimentant et enrichissant la base de connaissances. En retour, les règles induites pourraient enrichir les ontologies par de nouvelles connaissances inconnues ou implicites non exprimées par les experts.

Remerciement

Ce travail s'inscrit dans le cadre du projet CARTAM-SAT (CARtographie du Territoire AMazonien : des Satellites aux AcTeurs), financé sous une convention FEDER (Fonds Européen de Développement Régional) pour la région de La Guyane, ainsi que le projet GEOSUD EQUIPEX.

Références

ABUDAWOOD T. & FLACH P. (2011). Learning multi-class theories in ilp. In *Proceedings of the 20th international conference on Inductive logic programming*, ILP'10, p. 6–13, Berlin, Heidelberg : Springer-Verlag.

ANDRES S., ARVOR D. & PIERKOT C. (2012). Towards an ontological approach for classifying remote sensing images. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, p. 825–832.

ANTHONY J. & GARRETT J. (2005). Understanding Interobserver Agreement : The Kappa Statistic. *Family Medicine*, **37**, 360–363.

BAYOUDH M., ROUX E., NOCK R. & RICHARD G. (2012). Automatic learning of structural knowledge from geographic information for updating land cover maps. In *Symposium of the Latin American Society for Remote Sensing and Spatial Information Systems (SELPER)*, p. 1–10 : SELPER.

BLOCKEEL H., DZEROSKI S., KOMPARE B., KRAMER S., PFAHRINGER B. & LAER W. (2004). Experiments in predicting biodegradability. *Applied Artificial Intelligence*, **18(2)**, 157–181.

CHELGHOUM N., ZEITOUNI K., LAUGIER T., FIANDRINO A. & LOUBERSAC L. (2006). Fouille de donnees spatiales - approche basee sur la programmation logique inductive. p. 529–540.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

CORDIER M. (2005). Sacadeau : A decision-aid system to improve stream-water quality. *ERCIM News*, **61**, 37–38.

DURAND N., DERIVAUX S., FORESTIER G., WEMMERT C., GANCARSKI P., BOUSSAID O. & PUISSANT A. (2007). Ontology-based object recognition for remote sensing image interpretation. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 01*, ICTAI '07, p. 472–479, Washington, DC, USA : IEEE Computer Society.

EDWARD J., GARDEL A., GRATIOT N., PROISY C., ALLISON M., DOLIQUE F. & FROMARD F. (2010). The amazon-influenced muddy coast of south america : A review of mud-bank-shoreline interactions. *Earth-Science Reviews*, **103**(3-4), 99–121.

FONSECA N., SILVA F. & CAMACHO R. (2006). April - an inductive logic programming system. In *JELIA*, p. 481–484.

FROMONT E., CORDIER M. & QUINIOU R. (2005). Extraction de connaissances provenant de données multisources pour la caractérisation d'arythmies cardiaques. In O. BOUSSAID, P. GANÇARSKI, F. MASSEGLIA, B. TROUSSE, G. VENTURINI & D. ZIGHED, Eds., *Fouille de données complexes*, volume RNTI-E-4 of *Revue des Nouvelles Technologies de lĪnformation*, p. 25–45. Cepaduès.

GOODACRE J. (1996). *Inductive Learning of Chess Rules Using Progol*. Oxford University.

HUDELOT C., ATIF J. & BLOCH I. (2008). Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, **159**(15), 1929–1951.

LAVRAC N. & DZEROSKI S. (1994). *Inductive Logic Programming : Techniques and Applications*. Ellis Horwood.

LUU T., RUSU A., WALTER V., LINARD B., POIDEVIN L., RIPP R., MOULINIERAND JEAN MULLER L., RAFFELSBERGER W., WICKER N., LECOMPTE O., THOMPSON J., POCH O. & NGUYEN H. (2012). Kd4v : Comprehensible knowledge discovery system for missense variant. *Nucleic Acids Research*, **40**(1), W71–W75.

MALERBA D., ESPOSITO F., LANZA A., LISI F. & APPICE A. (2003). Empowering a gis with inductive learning capabilities : the case of ingens. *Computers, Environment and Urban Systems*, **27**(3), 265 – 281.

MUGGLETON S. (1991). Inductive logic programming. *New Generation Computing*, **8**, 295–318.

MUGGLETON S. (1995). Inverse entailment and progol. *New Generation Computing*, **13**, 245–286.

RICHARD L. & KOCH G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), pp. 159–174.

SANTOS J., NASSIF H., PAGE D., MUGGLETON S. & STERNBERG M. (2012). Automated identification of protein-ligand interaction features using inductive logic programming : A hexose binding case study. *BMC Bioinformatics*, **13**(1), 162.

SRINIVASAN A. (2007). The aleph manual. `http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html`.

SRINIVASAN A., MUGGLETON S., STERNBERG M. & KING R. (1996). Theories for mutagenicity : A study in first-order and feature-based induction. *Artificial Intelligence*, **85**, 277–299.

TEAM G. D. (19992012). Welcome to grass gis. `http://grass.fbk.eu/`.

VAZ D., FERREIRA M. & LOPES R. (2007). Spatial-yap : a logic-based geographic information system. In *Proceedings of the 23rd international conference on Logic programming*, ICLP'07, p. 195–208, Berlin, Heidelberg : Springer-Verlag.

VAZ D., SANTOS COSTA V. & FERREIRA M. (2010). Fire ! firing inductive rules from economic geography for fire risk detection. In *ILP*, p. 238–252.