UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Résumé de la thèse de doctorat

Résumé de la thèse

Discipline : Informatique (42000 18)

Présentée par : Arnaud QUIRIN
LSIIT, Campus d'Illkirch
03 90 24 45 78
quirin@lsiit.u-strasbg.fr

Titre : Découverte de règles de classification par approche évolutive : application aux images de télédétection

Unité de recherche : Laboratoire des Sciences, de l'Image, de l'Informatique et de la Télédétection
UMR 7005 - ULP/CNRS

Directeur de thèse : Jerzy KORCZAK, Professeur
LSIIT, Campus d'Illkirch
03 90 24 45 80
jjk@dpt-info.u-strasbg.fr

Co-directeur de thèse : Massimo MENENTI, DR
LSIIT, Campus d'Illkirch
03 90 24 45 07
mmenenti@termxjy.u-strasbg.fr

1 Contexte

Cette thèse s'inscrit dans le projet européen TIDE (Tidal Inlets Dynamics and Environment) visant à développer et valider des modèles dynamiques complets de systèmes marécageux incorporant à la fois des processus écologiques et physiques. Les travaux effectués dans cette thèse se concentrent sur la méthodologie de construction d'un système de classifieurs destinés aux géographes et thématiciens du domaine, permettant de découvrir de nouvelles expertises, présentées sous forme de règles de classification (classifieurs), à partir d'une base d'images volumineuses, bruitées et complexes. Nous proposons une nouvelle approche de fouille d'images de télédétection basée sur un modèle évolutif qui permet de découvrir automatiquement des classifieurs à partir des instances des classes.

2 Résumé de la thèse

2.1 Introduction

En classification supervisée d'images, la découverte de classes précises et exactes compte parmi les objectifs essentiels que se fixent les thématiciens. De la part d'un algorithme de classification, répondre à cette requête est une mission relativement ardue. En effet, nous disposons souvent de nombreuses sources d'informations (images satellitaires, simulations à haute altitude, mesures aérostatiques, mesures laser, validation terrain, validation spectrométriques, informations expertes informelles) de différentes natures et souvent peu cohérentes entre elles, à partir desquelles les concepts thématiques étudiés par l'expert doivent être extraits. Dans de nombreux cas, il peut s'avérer nécessaire de considérer - en plus du problème habituel de classification - l'idée qu'un même pixel peut appartenir à plusieurs classes (approche floue). Les thématiciens considère que ce type d'algorithmes sont les seuls permettant de s'approcher davantage d'une modélisation correcte de la réalité spectrale du terrain, car chaque pixel n'est en fait qu'une mixture des valeurs de réflectance spectrale de plusieurs types de terrain différents, dont les différentes abondances caractérisent la forme finale du spectre observé.

Les sources principales d'information sont des images de télédétection haute résolution et des images hyperspectrales contenant des données volumineuses - jusqu'à 200 Mo par image - et complexes : elles renferment de nombreux canaux spectraux bruités (parfois jusqu'à un canal sur deux pour un total de 100 à 200 canaux), contenant parfois des informations erronées. D'autre part, les experts en télédétection ne se contentent pas d'acquérir et d'analyser qu'une seule source de données à la fois. L'exploitation des caractéristiques physiques pluridisciplinaires des instruments imageurs (un capteur laser peut par exemple servir à mesurer, outre la position des échantillons, leur élévation avec une définition de 15 cm, leur composition chimique ainsi que la vitesse et la direction du vol), la rentabilité (plusieurs types de capteurs aéroportés permettent de réduire les coûts mais encore de s'assurer de leur synchronisation et de la cohérence des résultats obtenus) et le besoin de zones de recoupement et de validation en sont quelques-unes des multiples raisons. L'un des objectifs de cette collaboration avec le projet TIDE est donc d'élaborer une méthode de découverte empirique de connaissances à partir d'images de télédétection multi-sources, telles que des images SPOT, LANDSAT, CASI, DAIS ou QuickBird.

2.2 L'approche évolutionnaire

Pour faire face à la complexité et la taille des données, il est connu dans la littérature que les algorithmes évolutifs présentent une approche idéale pour ce genre de problème. Celle qui a été retenue est basée sur un système de classifieurs couplé à un algorithme génétique. Les systèmes de classifieurs (*Learning Classifier Systems* ou LCS) permettent de développer des populations de règles de classification simples, lisibles et généralisantes (par la présence de caractères *joker* ou d'intervalles de confiance permettant de modéliser de nombreux cas similaires en une seule règle). Cette base de règles permet d'appliquer la connaissance apprise sur une autre partie de l'image voir une nouvelle image tout en garantissant un taux correct de vrais positifs et vrais négatifs pour la classification de nouveaux exemples. Les systèmes de classifieurs permettent aussi de chaîner les règles entre elles, c'est-à-dire que le résultat de la première

règle activée va déterminer l'activation de la suivante, ce qui permet de modéliser la connaissance apprise ou de découvrir un découpage de cette connaissance sous forme d'unités plus petites (*building blocks*). Le couplage avec un algorithme génétique permet une découverte de solutions de manière évolutive, ce qui permet à l'algorithme d'absorber une grande partie de la complexité de traitement d'une telle masse de données et surtout de combler les imprécisions ou les défaillances de la base d'apprentissage par la création de règles appropriées et spécifiques lorsque de nouveaux cas se présentent, à l'aide d'un opérateur génétique dédié (nommé *covering operator*). Cela lui permet de modifier son propre comportement en fonction des résultats obtenus en cours de classification ou à travers les manipulations interactives d'un expert. Une plate-forme d'expérimentation a été développée pour répondre à cette ambition.

2.3 Travaux présentés dans la thèse

Les travaux de cette thèse consacrés à la découverte des règles se découpent en quatre parties.

1. La première cherche à explorer l'influence de la représentation des classifieurs sur la capacité et la qualité de reconnaissance des différentes classes de terrain. Deux algorithmes ont été créés ou adaptés : **ICU** et **XCS-R**.
2. La deuxième partie a consisté à étudier plusieurs post-traitements de la base de règles produites par **XCS-R** afin d'en améliorer ou d'en simplifier le contenu.
3. La troisième partie a consisté à modifier des représentations existantes ou utiliser de nouveaux paradigmes pour traiter les problèmes de type flous.
4. Enfin, un certain nombre de mesures de qualité ont été développées pour juger de l'efficacité de ces algorithmes et des protocoles de validation ont été proposés pour comparer les différents résultats entre eux. Ces algorithmes ont tous été validés sur des données réelles de télédétection.

2.4 La découverte de règles avec ICU et XCS-R

Une première étape de nos travaux a donc été la mise en place d'un prototype de logiciel de classification, nommé **ICU** (disponible à l'adresse suivante : http://lsiit.u-strasbg.fr/afd/logiciels/icu), fonctionnant à partir d'un couple d'images (brute, experte) qui accepte la modification de certains paramètres en cours d'apprentissage, et qui présente des résultats encourageants. Cet algorithme utilise un pool de classifieurs pour chaque classe de concepts à apprendre et chaque classifieur dispose d'un pouvoir de représentation adapté aux données à apprendre, fondé sur des conjonctions de disjonctions de contraintes. Il prend en charge certains pré-traitements importants (calcul d'indices, ...), ainsi que tout le processus d'apprentissage (extraction des règles) et classification (application d'anciennes règles sur de nouvelles images). Ce prototype a été validé par des informaticiens, des géographes (issus du laboratoire Images et Ville de Strasbourg et de chercheurs de l'université de Padoue, Italie), ainsi que des étudiants de l'International Space University.

La seconde étape a été le développement d'un système de classifieurs nommé **XCS-R** à partir du système **XCS** initialement conçu par l'University of Illinois (Urbana-Champaign). Nous avons mis en place des procédures spécifiques au traitement de ce type particulier de données, alors que **XCS** était autrefois réservé à la résolution de problèmes binaires simples. Dans **XCS-R**, le pool est constitué de classifieurs pour toutes les classes à apprendre et doivent collaborer entre eux pour découvrir une couverture complète de l'espace de recherche. Après convergence de l'algorithme, la totalité du pool est conservée ce qui permet de l'utiliser lors d'une interrogation non standard de classification *soft* (un objet est classé dans non plus une mais N classes, sans que les proportions puissent être déduites).

2.5 Le raffinement des règles de XCS-R

XCS, sur lequel est basé notre algorithme, a été découvert par S. W. Wilson en 1995 et est souvent cité comme référence dans la littérature. Un reproche souvent évoqué pour **XCS** concerne la taille de la population de règles produites, parfois de 2000 à 6000 pour des problèmes supposés en nécessiter beaucoup

moins. La deuxième partie de ces travaux a donc consisté à post-traiter la population de règles obtenue par **XCS-R** pour la raffiner, améliorer la qualité de classification et réduire le nombre de règles. Pour cela, plusieurs approches ont été testées. La première est basée sur les algorithmes génétiques et a consisté à créer des individus représentant des sous-population de la base de règles initiale. Chaque individu représente une nouvelle base obtenue en supprimant, ajoutant, combinant ou modifiant les classifieurs de la base initiale. Chaque sous-population est évaluée à part et correspond au critère d'évaluation des individus génétique. La seconde approche est basée sur la réutilisation des règles de la base comme prédicats principaux pour former les nœuds d'un arbre de décision, découvert avec une adaptation de l'algorithme inductif **C4.5**. Dans cette approche, l'arbre de décision permet une simplification automatique (en nombre de règles) de la base de règles ainsi qu'une hiérarchisation de ces règles, ce qui tend à donner à l'arbre un pouvoir de généralisation plus important.

2.6 L'étude de représentations adaptées pour les problèmes de classification flous

La représentation numérique des règles d'**ICU** a ensuite été améliorée pour tenir compte d'une expertise floue. L'expert peut ainsi saisir pour chaque pixel les différentes compositions en classes pures (valeurs réelles en pourcentages), ou en donner des intervalles. Ce type d'information permet de définir les compositions par extension ou par compréhension (tel pixel contient de 0 à 0% de la classe A et de 0 à 100% de la classe B ou C) et ceci est utilisé comme fait d'apprentissage. L'algorithme a été nommé **ICUX** (pour *unmix*). À titre de complétude, un autre algorithme a été développé, **GramGen**, basé sur la programmation génétique contrainte par grammaire (Grammar-based Genetic Programming ou GGP). **GramGen** permet de produire des fonctions utilisant un ensemble d'opérateurs prédéfinis par un expert et contraint par une grammaire BNF représentant au mieux la solution à un problème donné. Ce sont aussi des algorithmes évolutionnaires donc robustes aux données bruitées, mais produisant souvent des solutions complexes (arbres relativement profonds). Cependant, leur conception les autorise à produire des fonctions qui gèrent intrinsèquement l'*unmixing*. Par exemple, un indice typique utilisé en télédétection consiste à associer le pourcentage de végétation d'une région à un rapport normalisé de réflectances spectrales, selon une formule qui peut être déduite simplement par **GramGen**. Une étude de la qualité et de la pertinence des solutions a été proposée dans ce cadre. Les algorithmes étudiés (**ICUX** et **GramGen**) ont été comparés à des algorithmes statistiques connus pour être robustes dans ce domaine (réseaux de neurones et Support Vector Machine Regression).

2.7 Les mesures de qualité et les validations

Afin de produire des mesures de qualité objectives, une plate-forme de validation (**VPlat**) a été développée. Cette plate-forme sert deux buts : produire des mesures de qualité indépendantes pour chaque algorithme et permettre la comparaison des algorithmes entre eux par l'utilisation de protocoles de validation identiques lors des différentes études de cas. Elle a été utilisé pour extraire des statistiques détaillées sur l'apprentissage, les taux de vrai et faux positifs, les matrices de confusion (κ-index) ou l'influence de la sélection des exemples d'apprentissage. Les protocoles de validation développés sont le holding-out, la cross-validation, le boot-strapping, le jack-kniffing, et l'étude de l'influence des paramètres importants à l'aide de courbes ROC (*Receiver Operating Characteristic*). Ceci a permis de dégager plusieurs propriétés intéressantes des classifieurs concernant leur efficacité et leur résistance au bruit. Dans le cadre de la comparaison directe de différentes classifications obtenues par des algorithmes différents mais sur la même image, un algorithme basé sur la consensualité des résultats pour chaque pixel (*voting system*) a permis d'obtenir une information localisée dans l'espace des pixels ou des classes problématiques et de proposer une amélioration de la qualité de classification par fusion de classifications existantes (*consensuality map*). De nombreuses validations sont proposées sur des images réelles, voir bruitées, en utilisant ces mesures de qualité.

2.8 Les contributions

Les contributions apportées par cette thèse peuvent être regroupées en deux catégories : celle concernant l'étude des modèles et de la qualité des systèmes de classifieurs existants et celle concernant l'étude de représentations alternatives pour les classifieurs. Dans le premier cas, des mesures de qualité ont été développées et testées sur des images de différentes sortes (résolutions, tailles, valeurs de pixels et expertises binaires, entières ou réelles) pour évaluer la capacité de généralisation des systèmes de classifieurs existants ou développés dans la thèse, ainsi que leur robustesse face au bruit. Dans le second cas, plusieurs représentations de classifieurs ont été comparées par rapport à leur lisibilité, expressivité, complexité ou efficacité à répondre aux différentes contraintes posées par l'expert. Ces recherches ont abouti à la conception et au développement de plusieurs logiciels fonctionnels permettant la classification d'images de télédétection par systèmes de classifieurs. Et cela a permis de dégager de nouvelles perspectives concernant l'optimisation des algorithmes utilisés afin d'améliorer, par exemple, le temps de calcul ou passer à des représentations encore plus sophistiquées. L'utilisation de règles complexes pouvant gérer l'aspect temporel de certaines données ou la prise en compte d'informations contextuelles peuvent être aussi considérées comme des perspectives réalisables.

3 Publications

3.1 Chapitre de livre

Quirin A., Korczak J., *Discovering of Classification Rules from Hyperspectral Images*, volume 2936 de la série LNCS (Springer-Verlag), Genetic and Evolutionary Computation in Image Processing and Computer Vision, 2005 (*à paraître*).

3.2 Conférences internationales avec actes et comité de lecture

Quirin A., Korczak J., *Representation of Genetic Individuals for Unmixing Multispectral Data*, [in] 2005 IEEE Congress on Evolutionary Computation (CEC'2005), Edinburg, 2005.

Quirin A., Korczak J., Butz M. V., Goldberg D. E., *Analysis and Evaluation of Learning Classifier Systems applied to Hyperspectral Image Classification*, [in] 5th International Conference on Intelligent Systems Design and Applications (ISDA'2005), Wroclaw, pages 280-285, 2005.

Korczak J., Quirin A., *Evolutionary Approach to Discovery of Classification Rules from Remote Sensing Images*, [in] 5th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (EvoIASP'2003), Essex, 2003.

Korczak J., Quirin A., *Evolutionary Mining for Image Classification Rules*, [in] 6th International Conference on Artificial Evolution (EA'2003), Marseille, 2003.

3.3 Conférences nationales avec actes et comité de lecture

Korczak J., Quirin A., *Découverte de règles de classification par approche évolutive : application aux images de télédétection*, Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2003), Lyon, 2003.

3.4 Rapport de recherche

Quirin A., Korczak J., Butz M. V., Goldberg D. E, *Learning Classifier Systems for Hyperspectral Images Processing*, Research Report 2001/05, Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, CNRS, Université Louis Pasteur, Illkirch, 2004.