UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Summary of the PhD thesis

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

Title : Discovering classification rules using an evolutionary approach : application to remote sensing images

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

Thesis advisor : Jerzy KORCZAK, Professor
LSIIT, Illkirch Campus
03 90 24 45 80
jjk@dpt-info.u-strasbg.fr

Thesis co-advisor : Massimo MENENTI, Research dir. (CNRS)
LSIIT, Illkirch Campus
03 90 24 45 07
mmenenti@termxjy.u-strasbg.fr

1 Context

This thesis is funded by the European project TIDE (Tidal Inlets Dynamics and Environment) aiming to develop and validate complete dynamic models of marshy systems incorporating at the same time ecological and physical processes. The work carried out in this thesis concentrates on the methodology of construction of a classifier system intended for the geographers and thematicians of the field, making it possible to discover new expertises, presented in the form of classification rules (classifiers), starting from a base of large, disturbed and complex images. We propose a new approach of remote sensing images mining based on an evolutionary model which makes it possible to automatically discover classifiers starting from the classified samples.

2 Summary of the thesis

2.1 Introduction

In supervised image classification, the discovery of precise and exact classes is one of the essential objectives of the geographers. To answer this request is relatively difficult for a classification algorithm. Indeed, we often have many information sources (satellite images, high altitude simulations, aerostatic and laser measurements, ground truthing, spectrometric surveys, non formal expert information) of various nature and often not very coherent one with another, from which the thematic classes studied by the expert must be extracted. In many cases, it can prove to be necessary to consider - in addition to the usual problem of classification - the idea that the same pixel can belong to several classes (fuzzy approach). The geographers considers that this kind of algorithms are the only approaches which make possible a correct modeling of the spectral reality of the ground, because each pixel is in fact a mixture of the values of spectral reflectance for several different kind of ground, whose various abundances characterize the final shape of the observed spectrum.

The main information source are high resolution and hyperspectral remote sensing images containing voluminous - up to 200 MB by image - and complex data : they contain many noisy spectral channels (sometimes 50% for a total from 100 to 200 channels), and sometimes erroneous information. In addition, the remote sensing experts are not satisfied at the same time with the acquirement and the analysis of only one data source. Some of the multiple reasons are for instance the exploitation of the multi-field physical characteristics of the instruments (for instance, a laser sensor can be used to measure, in addition to the position of the samples, their heights with a definition of 15 cm, their chemical compositions as well as the speed and the direction of the flight), the profitability (several kind of airborne sensors used together make it possible to reduce the costs but also to be sure of their synchronization and the coherence of the obtained results) and finally the need of orthorectification and validation zones. Thus, one of the objectives of the collaboration initiated by the TIDE project is to work on a method of discovering empirical knowledge starting from multi-source remote sensing images, such as SPOT, LANDSAT, CASI, DAIS or QuickBird images.

2.2 The evolutionary approach

To face the complexity and the size of the data, it is known in the literature that the evolutionary algorithms present an ideal approach for this kind of problem. That which was retained is based on a classifier system coupled to a genetic algorithm. The classifier systems (*Learning Classifier Systems* or LCS) make it possible to discover populations of classification rules simple, readable and generalizing (by the presence of *joker* characters or confidence intervals modeling many similar cases in one single rule). This rule base makes it possible to apply the acquired knowledge to another part of the image and even to a new image while guaranteeing a correct true positive and negative rate for the classification of new samples. The classifier systems also make it possible to chain the rules between them, i.e. the result of the first activated rule will determine the activation of the following one, which makes it possible to model the acquired knowledge in the form of smaller units or to construct more sophisticated behaviors

(*building blocks*). The coupling with a genetic algorithm allows a discovery of solutions in an evolutionary way. That facilitates the absorption of the treatment complexity of such data and especially circumvents the inaccuracies or the missing data of the training base by the creation of suitable and specific rules when new cases arise, using a dedicated genetic operator (*covering operator*). That enables the LCS to modify its own behavior according to the results obtained during the classification or through interactive handling of an expert. A experimentation platform was developed to fill this need.

2.3 Work presented in the thesis

The work of this thesis, devoted to discovering classification rules, can be presented in four parts.

1. The first part explores the influence of the classifiers representation about the recognition capacity and the recognition quality of the various ground classes. Two algorithms were created or adapted : **ICU** and **XCS-R**.
2. In the second part, we study several postprocessing of the rule base produced by **XCS-R** in order to improve or to simplify its contents.
3. In the third part, we propose some modifications of existing representations or we use new paradigms to deal with fuzzy classification problems.
4. Lastly, a given number of quality measurements were developed to judge the robustness of these algorithms and validation protocols were proposed to compare the various results between them. All these algorithms were validated on real remote sensing data.

2.4 Discovering rules with ICU and XCS-R

The first step of our work was the development of a classification software, named **ICU** (available on the following web site : http://lsiit.u-strasbg.fr/afd/logiciels/icu), starting from a couple of images (raw, expert) and discovering classification rules. This software accepts the modification of some parameters during the learning, and has encouraging results. This algorithm uses a pool of classifiers for each thematic class to learn and the representation of each classifier is adapted to formalism of the remote sensing data, based on conjunctions of disjunctions of constraints. The program deals with some significant preprocessing (index calculation, histogram equalization, ...), as well as learning (extraction of the rules) and classification (application of rule base on new images). This prototype was validated by data processing specialists, geographers (*Image and Ville* laboratory in Strasbourg and researchers from the Padova university, Italy), as well as students of International Space University (Strasbourg).

The second step was the improvement of a classifier system named **XCS-R** starting from a system initially conceived by the University of Illinois (Urbana-Champaign). In **XCS-R**, the pool is made up by classifiers corresponding to all the classes and are evolved to discover a complete mapping of the search space. After the convergence of the algorithm, the totality of the pool is preserved. That makes it possible to use it for a nonstandard interrogation (*soft* classification : an object is classified not in one but in N classes).

2.5 Refinement of the XCS-R rules

XCS, on which our algorithm is based, was discovered by S. W. Wilson in 1995 and is often quoted as reference in the literature. A reproach usually evoked for **XCS** relates to the size of the population of produced rules, sometimes between 2000 to 6000 for problems supposed to require some much less. Consequently, the second part of this work concern the post-treatment of the rule population obtained by **XCS-R** to refine it, improve the classification quality and reduce the number of rules. Several approaches were tested. First is based on the genetic algorithms and consisted in creating individuals representing a subpopulation of the initial rule base. Each individual represents a new base obtained by removing, adding, combining or modifying the classifiers of the initial base. Each subpopulation is evaluated separately and this evaluation is used for the genetic evaluation of the individuals. The second approach is based on

the re-use of some rules as principal predicates to form the nodes of a decision tree, discovered with an adaptation of the inductive algorithm **C4.5**. In this approach, the decision tree allows an automatic simplification (in terms of the number of rules) of the rule base as well as a hierarchy of these rules, which tends to give to the tree a more significant generalization capacity.

2.6 Study of representations adapted for the problems of fuzzy classifications

The representation of **ICU** rules was improved to take account of fuzzy expertises. The expert can input, for each pixel, the various compositions in pure classes (genuinly values in percentages), or give intervals to express these percentages. This kind of information makes it possible to define the compositions by extension or comprehension (for instance, such pixel contains from 0 to 0% of class A and from 20 to 50% of class B and from 0 to 100% of class C) and this is used as training facts. The algorithm was named **ICUX** (for *unmix*). By way of completeness, another algorithm was developed, **GramGen**, which is based on grammar-based genetic programming. **GramGen** makes it possible to produce functions using a whole set of operators and constrained by a BNF grammar for representing the best solution for a given problem. The grammar is used to reduce the search space. They are robust evolutionary algorithms able to handle noisy data, but producing often complex solutions (relatively deep trees). However, their design authorizes them to manage the *unmixing* intrinsically. For instance, a typical index used in remote sensing context consists in associating the percentage of vegetation of an area to a ratio of spectral reflectances, according to a formula which can be deduced simply by **GramGen**. A study of the quality and the relevance of the solutions were proposed within this framework. The studied algorithms (**ICUX** and **GramGen**) were compared with statistical algorithms known to be robust in this field (neural networks and Support Vector Machine Regression).

2.7 Quality measurements and validations

In order to produce quality measurements, a validation platform (**VPlat**) was developed. This platform serves two goals : to produce independent quality measurements for each algorithm and to allow the comparison of the algorithms between them by the use of the same validation protocols during the various case studies. The platform was used to extract detailed statistics about the training, true positive and negative rates, matrices of confusion, κ-index or the influence of the training cases sampling on the performance. The developed validation protocols are the holding-out, the cross-validation, the boot-strapping, the jack-knifing, and the study of the influence of the main parameters using ROC curves (*Receiver Operating Characteristic*). This made it possible to test several interesting properties of the classifiers concerning their effectiveness and their resistance to the noise. We also propose a method for the direct comparison of various classifications obtained by different algorithms on the same image. This algorithm is based on the *consensuality* of the results for each pixel (*voting system*) and made it possible to obtain localized information about problematic pixels or problematic classes and propose an improvement of the classification quality by the fusion of existing classifications (*consensuality map*). Many validations are proposed on real-world images, even noisy, using these quality measurements.

2.8 Contributions

The contributions proposed by this thesis can be gathered in two categories : that concerning the study of the models and the quality of the existing classifier systems and that concerning the study of alternative representations for the classifiers. In the first case, quality measurements were developed and tested on images of various kinds (resolutions, sizes, values of pixels and binary, integer or continuous expertises) to evaluate the capacity of generalization of existing classifier systems or that developed in the thesis, in conjunction with their robustness concerning the noise. In the second case, several representations of classifiers were compared to study their legibility, expressivity, complexity or effectiveness to settle the various constraints requested by the expert. This research led to the design and the development of several functional softwares allowing the classification of remote sensing images by classifier systems. New

perspectives can then be proposed concerning the optimization of used algorithms in order to improve, for instance, the computing time or to test more sophisticated representations. The use of complex rules being able to manage temporal data or which take account of contextual information can be also regarded as realizable perspectives.

3 Publications

3.1 Book chapter

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 (*to be published*).

3.2 International conferences with proceedings and review committee

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 National conference with proceedings and review committee

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 Research report

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.