Kernel-based statistical learning for individual tree recognition in forest inventory

Zaremba, M.B.[1] and Gougeon, F.A.[2]

[1] *Professor, Department of Computer Science and Engineering,*
Université du Québec en Outaouais
101 St-Jean-Bosco, Gatineau, QC J8Y 3G5
E-mail: zaremba@uqo.ca
[2] *Senior Scientist, Natural Resources Canada, Pacific Forestry Centre*
506 West Burnside Rd., Victoria, BC V8Z 1M5
Email: fgougeon@nrcan.gc.ca

Abstract

The general trend toward precision and sustainable forestry calls for a transition from mapping relatively homogeneous forest stands and manually interpreting their content to the use of semi-automated, computer-assisted analysis of high-resolution multi-spectral images realized on the individual tree crown (ITC) basis. Automation of the image analysis tasks can be achieved through the application of the theory of machine learning. This paper presents the kernel-based methodological approach to classification problems, and discusses its application to processing of high-resolution satellite imagery and LIDAR data in the area of precision forest management.

Introduction

Implementation of computer-assisted analysis of high-resolution multi-spectral images realized on an individual tree crown basis (ITC) [1] is a prerequisite for automated generation of precise forest management inventories. Efficient and precise analysis at the level of the individual crown has been made possible by recent developments in high-resolution remote sensing. The high spatial and radiometric resolution of such satellites as QuickBird, OrbView3 or Ikonos facilitates visual interpretation. Temporal resolution of image databases can be largely increased, due to the known revisit time and pointing capabilities of the satellite platform, which facilitates large-scale change-detection and monitoring of selected areas in order to keep forestry databases up-to-date.

Achieving a semi-automated ITC-based interpretation of high spatial resolution digital data for forestry requires that several components work together. The key components are:
- automatic isolation of visible individual tree crowns,
- classification of individual tree crowns,
- regrouping into stands or strata,

The delineation process uses a rule-based approach to systematically follow boundaries from the inside of a specific crown (or cluster) to produce more distinct crowns. A near infrared band is typically used because of its sensitivity to illumination variations and its

good response to vegetative material. That process contains numerous, context-sensitive crown separation criteria. After the acquisition of all the tree species signatures, classification is done on an individual crown basis. The degree to which the programs will confuse one species with another is a major consideration in determining accuracy. Different classification algorithms, such as those using maximum likelihood function or neural networks can be used to assign a species (class) to the unknown ITC.

Recently, a methodological approach based on kernel methods [2] has been used for classification and regression – the approach is systematic, reproducible and properly motivated by statistical learning theory. The approach consists in mapping the input vectors into a high-dimensional feature space through an *a priori* chosen, usually nonlinear, mapping that generates kernel representations. Different kernel functions can be used. The kernel-based approach has gained popularity because of such attractive features as its greater ability to generalize due to the minimization of an upper bound on the expected risk. From the statistical point of view, the empirical success of kernel-based learning can be attributed to the fact that for appropriately chosen tuning parameters, the optimal classification rule can be implemented asymptotically in a very efficient manner. Another advantage of using a kernel function is that the number of tunable parameters no longer depends on the number of attributes used in the input space; mapping to high-dimensional feature space does not increase the number of these parameters.

This paper discusses issues related to multi-class classification of tree crowns using the Support Vector Machines (SVM) technology, a class of kernel methods. After presenting the ITC and SVM methods, some results of the research on the detection and classification of individual trees in a boreal forest in Alberta (north-east of Edmonton) are provided.

Individual Tree Crown (ITC) Approach

For the typical medium density Canadian forest, most dominant and co-dominant tree crowns are visible on high spatial resolution (< 100cm/pixel) satellite imagery and can often be separated based on the shade found between them. After having masked the non-forested areas, the ITC program follows the valleys of shade that exist between the much brighter tree crowns in the spectral topography of a panchromatic image. The delineation process is then improved and completed using a set of rules aimed at the remaining tree clusters. Typically, depending on the spatial resolution of the image, 65- 80% of the main visible tree crowns can be separated by these two processes. An infra-red pansharpened image of the boreal forest used as an example illustrating the method is shown in Fig. 1a. The results of the application of the ITC tree delineation procedure are shown in Fig 1b. In addition to the crown polygons, the image in Fig. 1b also shows the illuminated portions of the tree crown zones.

Following a supervised classification protocol, spectral signatures are created for all the tree species. The classification process then compares one-by-one all of the ITC's signatures with the species (classes) signatures and using a maximum likelihood decision rule assigns them a class (within a specified confidence interval). Finally, the ITCs are

regrouped into forest stands based on species composition, crown closure and stem density.

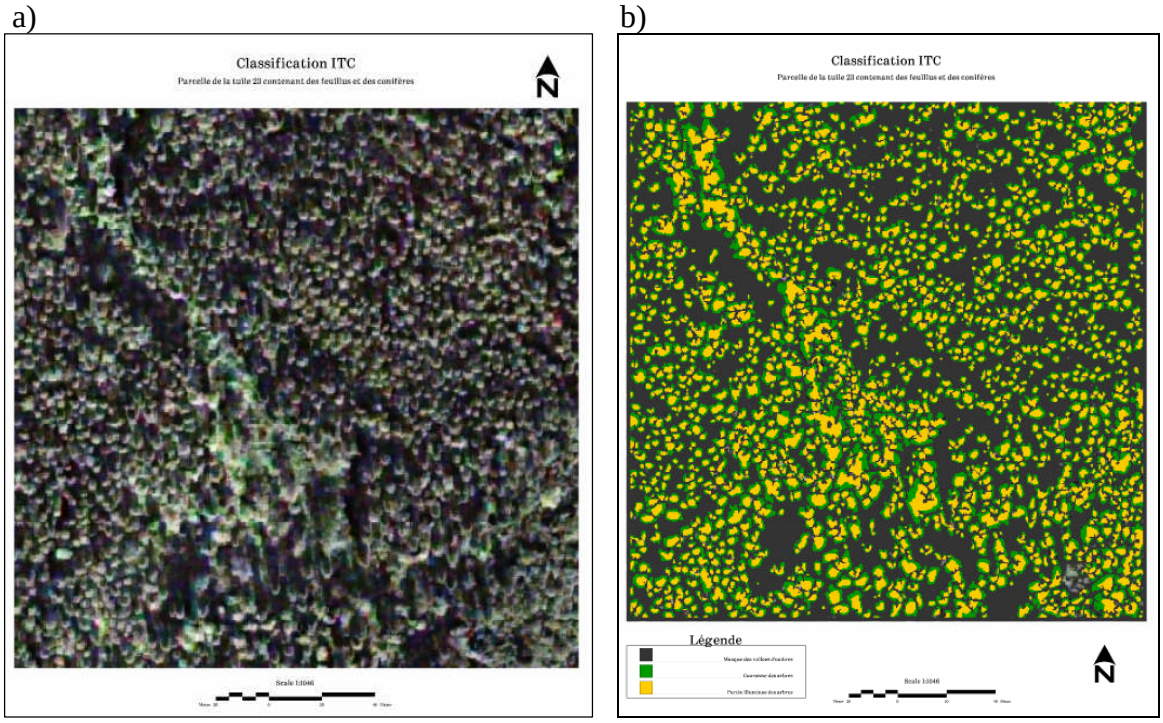


Figure 1. Pansharpened QuickBird image of the Alberta boreal forest (a) and the results of the ITC tree crown delineation procedure (b).

Support Vector Machines

A method that bounds the generalization error and manages the model complexity, by applying the concept of Structural Risk Minimization, was proposed by Vapnik [3]. The method, known as the Support Vector Machine (SVM) kernel-based method, consists in mapping the input vectors ***x*** into a high-dimensional feature space *Z* through an a priori chosen, usually nonlinear, mapping. Projection of the data into a high-dimensional feature space is performed using a preprocessing strategy that generates kernel representations. In general, a kernel is a function *K*, such that for all **x**, **x'** 0 *X*,

$$K(\mathbf{x}, \mathbf{x'}) = +\phi(\mathbf{x})\mathbb{A}\ \phi(\mathbf{x'}), \qquad (1)$$

where ϕ is a mapping from *X* to a feature space *F*. Different kernel functions can be used. For example:

- polynomial kernel: $K(\mathbf{x}, \mathbf{x'}) = (+\mathbf{x}, \mathbf{x'}), +1)^{d}$
- Radial Basis Function (RBF) kernel: $K(\mathbf{x}, \mathbf{x'}) = \exp(-2\mathbf{x} - \mathbf{x'}2/2\sigma^{2})$
- multi-layer perceptron: $K(\mathbf{x}, \mathbf{x'}) = \tanh(l+\mathbf{x}, \mathbf{x'}), + m)$

Selection of the best kernel for a particular problem is a question that arises in many applications. In the context of the SVM theory, a means of comparing different kernels is the evaluation of the upper bound of the Vapnik-Chervonenkis dimension [3]. In practice, statistical methods, such as bootstrapping and cross-validation are used for kernel selection.

After the selection of the kernel function, an optimal discrimination hyperplane is calculated in *Z*. The discrimination hyperplane is a function defined in the feature space

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{2}$$

where $\mathbf{w}$ is the function coefficient vector, b is a constant, and $\langle \cdot, \cdot \rangle$ denotes the dot product in the feature space. The SVM algorithm places the hyperplane such that the margin between two classes is maximized, which improves the generalization performance, while the classification error is minimized. The minimization of the classification error is achieved by introducing the constraint of the cost function in the following optimization problem [1]:

> *Given a linearly separable training sample,* $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_s, y_s))$, *find the hyperplane* $(\mathbf{w}, b)$ *that minimizes* $\langle \mathbf{w} \cdot \mathbf{w} \rangle_{w,b}$ *subject to* $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$; $i = 1, \ldots, s$.

The constraint is regarded as a canonical representation of the separating hyperplane, and it requires that the margin between the two classes is $2/2\|\mathbf{w}\|^2$. Consequently, in order to achieve good generalization, we should minimize $2\|\mathbf{w}\|^2/2$ subject to constraint (3). Hence, the minimization problem can be described as:

> *Minimize* $\frac{1}{2}\|\mathbf{w}\|^2$ *subject to* $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1 \geq 0$

and solved by applying the Lagrangian multipliers method. The primary Lagrangian function takes the form:

$$L(\mathbf{w}, b, \alpha) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{s} \alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1] \tag{3}$$

where $\alpha_i \geq 0$ are the Lagrangian multipliers. Since the cost function, i.e., $\frac{1}{2}\|\mathbf{w}\|^2$, is convex, function (3) is defined in terms of a convex quadratic programming problem.

Maximizing the corresponding dual is resolved to solving the gradients of (3) with respect to $\mathbf{w}$ and b:

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{s} y_i \alpha_i \mathbf{x}_i = 0 \tag{4}$$

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} = \sum_{i=1}^{s} y_i \alpha_i = 0 \tag{5}$$

and resubstituting $\mathbf{w} = \sum_{i=1}^{s} y_i \alpha_i \mathbf{x}_i$ and $0 = \sum_{i=1}^{s} y_i \alpha_i$ into the primary function (3). Thus, the optimization problem is defined in terms of the minimization of function

$$L(\mathbf{w}, b, \alpha) = \sum_{i=1}^{s} \sum_{j=1}^{s} \alpha_i \alpha_j y_i y_j \, \mathbf{x}_i, \mathbf{x}_j, - \sum_{i=1}^{s} \alpha_i$$

$$\text{s. t.} \quad \sum_{i=1}^{s} y_i \alpha_i = 0 \qquad (6)$$

For nonlinear cases, the dot product in (6) can be replaced by a kernel *K*. The solution of the dual objective function is sparse. Most of the Lagrangian multipliers are equal to zero. The sample examples that correspond to the Lagrangian multipliers which are not null are called *support vectors.*

The results of SVM classification are shown in Figure 2b. They can be compared with the results of the classification using the maximum likelihood approach (Fig. 2a). Higher sensitivity to less frequently occurring tree species and better classification of shaded areas can be observed.

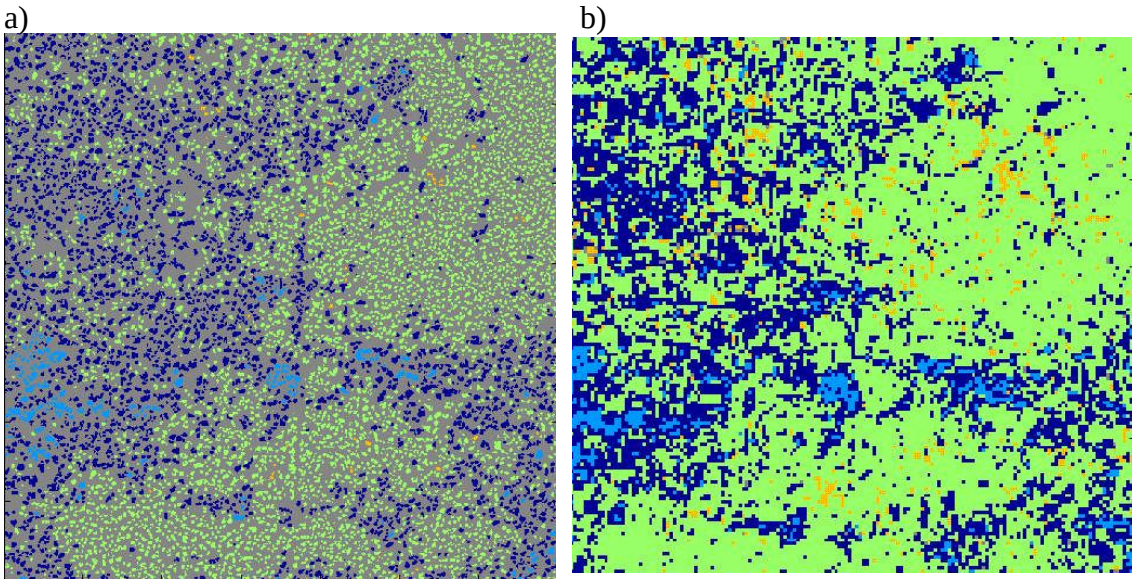


Figure 2. Classification results: maximum likelihood method (a), and the SVM method (b).

Conclusions

In the SVM method, the importance is placed on the samples located on the boundaries of the distribution zones. This allows for a better discrimination between classes closely located in the feature space, i.e., crowns of similar spectral characteristics. Since the classification solution of the SVM method depends on a small number of data points (the support vectors), judicious solution of the training samples can significantly reduce their

number. The issue of dealing with a small volume of training data and optimal selection of the training patterns is important from the standpoint of practical applications. Another advantage of using a kernel function is that mapping to high-dimensional feature space does not increase the number of tunable parameters. This provides a solution to the curse of dimensionality problem.

Acknowledgements

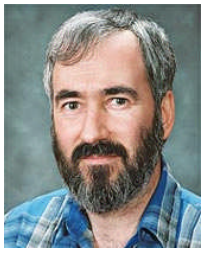
The authors wish to acknowledge the support of the NSERC Collaborative Research and Development program.

References

[1] Gougeon, F.A. and D.G. Leckie, Forest information extraction from high spatial resolution images using an individual tree crown approach. *NRC CFS Pacific Forestry Centre Report BC-X-396*, 2003.
[2] Schölkopf B. and A.J. Smola, *Learning with Kernels*. MIT Press, 2002.
[3] Vapnik, V.N., *Statistical learning theory*. J. Wiley, New York, 1998.

About the Authors

Marek B. Zaremba *is currently a Professor of Computer Engineering at the Gatineau campus of the Université du Québec. He has also been a Visiting Scholar at the Canada Centre for Remote Sensing (CCRS) in Ottawa, a Visiting Professor at the Vienna University of Technology in Austria, and at Université de Nancy I in France. He received the M.Sc. and Ph.D. degrees in control systems, both from the Warsaw University of Technology. His primary research interests are in the application of computational intelligence paradigms in remote sensing and image processing, adaptive systems, and learning control. He has authored or coauthored about 150 books and technical papers. Dr. Zaremba has served in different capacities in about 40 international conferences, symposia, and workshops, chairing three conferences.*



François A. Gougeon *is a Senior Scientist at the Pacific Forestry Centre, Natural Resources Canada, in Victoria, BC. He obtained his Ph.D. degree from the University of Waterloo in 1988, M.A.Sc. and B.A.Sc. from the University of Ottawa in 1976 and 1980 respectively. Dr. Gougeon's current research involves the development of new techniques, methods and processes to delineate, identify and regroup the individual tree crowns seen in high spatial resolution (30-100 cm/pixel) multispectral images from airborne or satellite sensors. An important long-term goal is the semi-automatic production of precise, accurate and timely forest inventories.*