Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril 2009, INPE, p. 7063-7070.

Land Use Classification Using Optimum-Path Forest

Rodrigo José Pisani[1]
João Paulo Papa[2]
Célia Regina Lopes Zimback[1]
Alexandre Xavier Falcão[2]
Ana Paula Barbosa[1]

[1]State University of São Paulo - FCA/UNESP
PO Box 237 - 18610307 - Botucatu - SP, Brazil
{pisani,czimback,apbarbosa}@fca.unesp.br

[2]University of Campinas - IC/ UNICAMP
PO Box 515 - 12245-970 - Campinas - SP, Brazil
{jpaulo,afalcao}@ic.unicamp.br

Abstract. It was introduced in this paper the Optimum-Path Forest for land use classification aiming a better environmental management, using images obtained from CBERS 2B CCD satellite covering the area of the Rio das Pedras watershed, Itatinga City, São Paulo State, Brazil. We also compared the Optimum-Path Forest algorithm with the well known supervised classifiers: Artificial Neural Networks using Multilayer Perceptrons, Bayesian Classifier and Support Vector Machines. The Optimum-Path Forest and Support Vector Machines classifiers had similar results and outperformed the other ones, but the first was much faster than the last one. As far we know, we are the first that used the Support Vector Machines classifier in this context of research field. Also were presented some qualitative results, in which the Optimum-Path Forest and Maximum Likelihood classifiers were compared against each other for visual purposes.

Keywords: watershed handling, land use supervised classification, optimum-path forest, image foresting transform, planejamento de microbacias, classificação supervisionada de uso da terra, floresta de caminhos ótimos, transformada imagem floresta

1. Introduction

The big grow of the mankind needs has been coughing several modifications in the natural environments, weakening the systems that keep the life form in the earth. The Industrial Revolution started a new mankind way of life, in which the nature is seen as a source of the human needs. In these systems that only economics values are led in account, the people do not take care about the importance of the nature as a crucial component for the life neither the sustainable exploitation.

Based on this assumption, several studies have been directed for a better environmental management, with special attention to the watershed handling (Rocha, 1991; Benetti and Bidone, 2001; Goes, 1994). The natural resources management needs to take account the natural units, such that the hydric balance, soil, vegetation and land use (Rocha, 1991), and any program that aims some kind of environmental controlling, such as pollution handling, for instance, must start from watershed territorial planning, since the water quality strongly depends on the vegetation and land modifications. The environmental planning term is wide and gains different definitions in agreement to the research field. For instance, the land use can be a method of support to the technician-scientific decisions, administrative politics and can define rational norms of performance and ordinance of the space with efficiency (Goes, 1994). Quantifying the land use, one can better control the amount of human intervention (farms, protected areas deforestation and water pollution) and its environmental impacts.

The remote sensing techniques have been a useful tool for the watersheds monitoring, allowing the fast identification of deforesting areas and illegal land use, for instance

(Daianese, 2001). Beside this, there exists a lot off of available information provided by the satellites, been a hard task for only human interpretation. Based on this, the agrarian sciences researches have been used automatic classification systems, trying to allow a fast data processing. In the remote sensing research field, to automatic classify means to join points of one image into groups that share similar properties (Rosa, 1992), labeling them into classes. The methods can be divided according to the available information: (i) supervised classification and (ii) unsupervised classification. The first one is used in situations in which you have fully information about the data and the second when you do not have it (Jain, 2000). It is also concerned that supervised classification algorithms can have a better performance than the unsupervised ones, due to the available information about the data behavior.

There exists a several variety of supervised classifiers in the literature, such that the well known Artificial Neural Networks using Multilayer Perceptrons (ANN-MLP) (Haykin, 1994), Bayesian Classifiers (BC) (Duda, 2000), Support Vector Machines (Vapnik, 1995) and the Optimum-Path Forest classifier (OPF) (Papa, 2008). The last one has been demonstrated to overcome the ANN-MLP and BC classifiers and to be similar to SVM, but much faster, which is a very important feature in the sense that we have a great number of available information provided by satellites. In this context, was proposed here the environmental management by the land use classification using the OPF classifier. Notice that is the first time that the OPF classifier is used in this research field, as well a complete study of the performance comparison of several supervised classifiers is done for these purposes. The Sections 2 presents the OPF classifier theory. The Sections 3 and 4 show the Experimental Results and Conclusions, respectively.

2. Optimum-Path Forest

Let Z_1 and Z_2 be training and test sets with $|Z_1|$ and $|Z_2|$ samples of a given dataset. Were used samples as pixels from images in this paper. Let $\lambda(s)$ be the function that assigns the correct label i, $i = 1,2,\ldots,c$, of class i to any sample $s \in Z_1 \cup Z_2$, $S \subset Z_1$ be a set of prototypes from all classes, and v be an algorithm which extracts n features (color, shape, texture properties) from any sample $s \in Z_1 \cup Z_2$ and returns a vector $\vec{v}(s)$. The distance $d(s,t)$ between two samples, s and t, is the one between their feature vectors $\vec{v}(s)$ and $\vec{v}(t)$. One can use any distance function suitable for the extracted features. The most common is the Euclidean norm $\|\vec{v}(t) - \vec{v}(s)\|$, but some image features require special distance algorithms (Rubner, 1998; Wang, 1990). A pair (v,d) then describes how the samples of a dataset are distributed in the feature space.

The problem consists of projecting a classifier which can predict the correct label $\lambda(s)$ of any sample $s \in Z_2$. It was proposed a classifier which creates a discrete optimal partition of the feature space such that any sample $s \in Z_2$ can be classified according to this partition. This partition is an optimum-path forest (OPF) computed on Z_1 by the image foresting transform (IFT) algorithm (Falcão, 2004).

Let (Z_1, A) be a complete graph whose nodes are the training samples and any pair of samples defines an arc in $A = Z_1 \times Z_1$ (Figure 1a). The arcs do not need to be stored and so the graph does not need to be explicitly represented. A path is a sequence of distinct samples $\pi_t = \langle s_1, s_2, \ldots, t \rangle$ with terminus at a sample t. A path is said trivial if $\pi_t = \langle t \rangle$. We assign to each path π_t a cost $f(\pi_t)$ given by a connectivity function f. A path π_t is said optimum if

$f(\pi_t) \leq f(\tau_t)$ for any other path τ_t. We also denote by $\pi_s \cdot \langle s,t \rangle$ the concatenation of a path π_s and an arc (s,t).

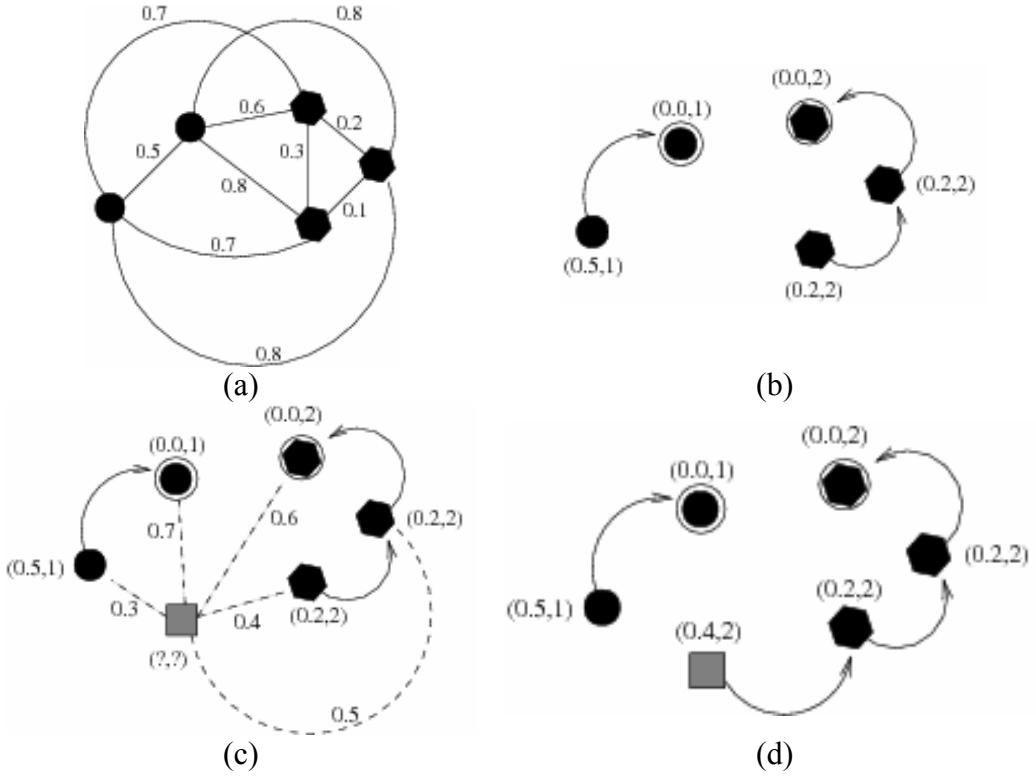

Figure 1: (a) Complete weighted graph for a simple training set. (b) Resulting optimum-path forest for $f_{\max}$ and two given prototypes (circled nodes). The entries (x, y) over the nodes are, respectively, the cost and the label of the samples. The directed arcs indicate the predecessor nodes in the optimum path. (c) Test sample (gray square) and its connections (dashed lines) with the training nodes. (d) The optimum path from the most strongly connected prototype, its label 2, and classification cost 0.4 are assigned to the test sample. The test sample is classified in the class hexagon, although its nearest training sample is from the class circle.

The OPF algorithm may be used with any smooth connectivity function which can group samples with similar properties (Falcão, 2004). A function f is smooth in (Z_1, A) when for any sample $t \in Z_1$, there exists an optimum path π_t which either is trivial or has the form $\pi_s \cdot \langle s,t \rangle$, where

1. $f(\pi_s) \leq f(\pi_t)$,
2. π_s is optimum,
3. for any optimum path τ_s, $f(\tau_s \cdot \langle s,t \rangle) = f(\pi_t)$.

We will address the connectivity function $f_{\max}$.

$$f_{\max}(\langle s \rangle) = \begin{cases} 0 & if\ s \in S, \\ +\infty & otherwise \end{cases}$$

$$f_{\max}(\pi_s \cdot \langle s,t \rangle) = \max\{f_{\max}(\pi_s), d(s,t)\} \quad (1)$$

such that $f_{\max}(\pi_s \cdot \langle s,t \rangle)$ computes the maximum distance between adjacent samples along the path $\pi_s \cdot \langle s,t \rangle$.

The OPF algorithm minimizes $f_{\max}$, by storing the minimum costs in a map C,

$$C(t) = \min_{\forall \pi_t \in (Z_1, A)} \{f_{\max}(\pi_t)\}, \quad (2)$$

and assigning one optimum path $P^*(t)$ from S to every sample $t \in Z_1$. Its result is an optimum-path forest P (a function with no cycles which assigns to each $t \in Z_1 \setminus S$ its predecessor $P(t)$ in $P^*(t)$ or a marker *nil* when $t \in S$, as shown in Figure 1b). The root $R(t) \in S$ of $P^*(t)$ can be obtained from $P(t)$ by following the predecessors backwards along the path, but its label is propagated during the algorithm by setting $L(t) \leftarrow \lambda(R(t))$.

2.1 Training

It was said that S^* is an optimum set of prototypes when the OPF algorithm minimizes the classification errors in Z_1. S^* can be found by exploiting the theoretical relation between minimum-spanning tree (MST) (Cormen et al., 1990) and optimum-path tree for $f_{\max}$ (Allène et al., 2007; Miranda et al., 2008). The training essentially consists of finding S^* and an OPF classifier rooted at S^*.

By computing an MST in the complete graph (Z_1, A), was obtained a connected acyclic graph whose nodes are all samples of Z_1 and the arcs are undirected and weighted by the distances d between adjacent samples (Figure 3a). The spanning tree is optimum in the sense that the sum of its arc weights is minimum as compared to any other spanning tree in the complete graph. In the MST, every pair of samples is connected by a single path which is optimum according to $f_{\max}$. That is, the minimum-spanning tree contains one optimum-path tree for any selected root node.

The optimum prototypes are the closest elements of the MST with different labels in Z_1. By removing the arcs between different classes, their adjacent samples become prototypes in S^* and the OPF algorithm can compute an optimum-path forest in Z_1 (Figure 1b). Note that, a given class may be represented by multiple prototypes (i.e., optimum-path trees) and there must exist at least one prototype per class.

It is not difficult to see that the optimum paths between classes tend to pass through the same removed arcs of the minimum-spanning tree. The choice of prototypes as described above aims to block these passages, reducing the chances of samples in any given class be reached by optimum paths from prototypes of other classes.

2.2 Classification

For any sample $t \in Z_2$, we consider all arcs connecting t with samples $s \in Z_1$, as though t were part of the training graph (Figure 1c). Considering all possible paths from S^* to t, we find the optimum path $P^*(t)$ from S^* and label t with the class $\lambda(R(t))$ of its most strongly connected prototype $R(t) \in S^*$ (Figure 1b). This path can be identified incrementally, by evaluating the optimum cost $C(t)$ as

$$C(t) = \min\{\max\{C(s), d(s,t)\}\}, \forall s \in Z_1. \quad (3)$$

Let the node $s^* \in Z_1$ be the one that satisfies Equation 3 (i.e., the predecessor $P(t)$ in the optimum path $P^*(t)$). Given that $L(s^*) = \lambda(R(t))$, the classification simply assigns $L(s^*)$ as the class of t (Figure 1d). An error occurs when $L(s^*) \neq \lambda(t)$.

3. Experimental Results

As aforementioned, the present work aims the land use classification using the OPF classifier. We use one image obtained from CBERS 2B CCD satellite covering the area of the Rio das Pedras watershed, located in Itatinga city, São Paulo State, Brazil. The watershed has 5172 ha and it is located between 732000 – 740000 and 7440000 – 7460000 UTM coordinate (Figure 2). We also applied the ANN-MLP, BC and SVM classifiers in this wok, allowing a comparison among them.

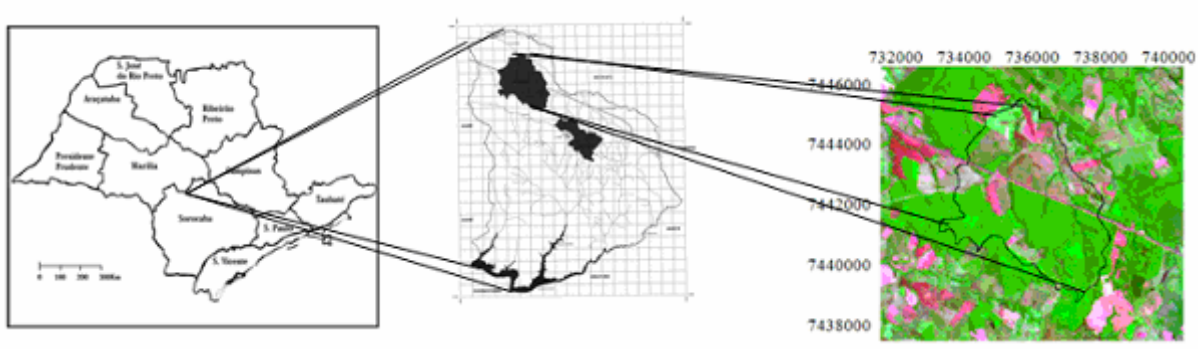

Figure 2: Rio das Pedras watershed location.

Were used as samples for the supervised classification a three dimensional feature vector $\vec{v}(s) = \{s_1, s_2, s_3\}$, in which s_1, s_2 and s_3 are, respectively, the gray values for each image pixel from bands 2, 3 and 4, respectively.

The experiments were conducted into two phases: a quantitative evaluation and a qualitative one. As a ground truth of the whole image is an unviable task, a technician was requested to label some parts of the image into five classes: (i) reforesting, (ii) pasture, (iii) culture, (iv) native bushes and (v) displayed soil. The labeled pixels were used as a training set for the aforementioned classifiers. In the quantitative phase, we divided the training set into two disjoint subsets: a new training set and a test set. The classifiers were trained and tested in these sets, allowing a comparison about their accuracies. With respect to the qualitative evaluation, we compared the OPF results against the ones obtained by the Maximum Likelihood classifier implemented in the IDRISI Andes software, from Clark Labs©.

For ANN-MLP algorithm, we used the FANN package (Nissen, 2003), which implements the following artificial neural network architecture 3:8:5, in which 3, 8 and 5 are the number of neurons in the input, hidden and output layers, respectively. For ANN-MLP training, we used the backpropagation algorithm. For SVM classifier, we used the LibSVM package (Chang and Lin, 2001), which implements some optimization procedures using the RBF kernel. With respect to the OPF classifier, we used the LibOPF (Papa, 2008b), which is a C library for the design of the optimum-path forest classifiers and finally, for BC classifier, we used our own implementation.

3.1 Quantitative Evaluation

The training set was divided into two new sets, with 50% of the whole size each. The ANN-MLP, SVM, BC and OPF classifiers were trained and tested in these sets. The experiments were executed 10 times with different randomly generated training and test sets to obtain the mean accuracy, standard deviation and mean execution time in seconds. Tables 1 and 2 display, respectively, the accuracy and execution time results.

Classifier	Mean accuracy	Standard deviation
ANN-MLP	75.97%	0.057
BC	87.45%	0.037
SVM	95.22%	0.005
OPF	94.56%	0.007

Table 1 – Mean accuracy and standard deviation for the applied supervised classifiers.

Classifier	Mean execution time
ANN-MLP	44.5575
BC	0.75449
SVM	34.9650
OPF	0.53747

Table 2 – Mean execution time in seconds.

Recall that the OPF and SVM had similar results and outperformed the ANN-MLP and BC classifiers. Even so the SVM classifier outperformed the OPF by 0.66%, the last was approximately 65 times faster than SVM, which is a very important feature with respect to the large amount of available satellite data to be classified.

3.2 Qualitative Evaluation

We present here the classification results using the OPF classifier and the Maximum Likelihood classifier implemented in the IDRISI Andes software for visual purposes. Figures 3, 4 and 5 display, respectively, the acquired, OPF classified and the Idrisi classified images.

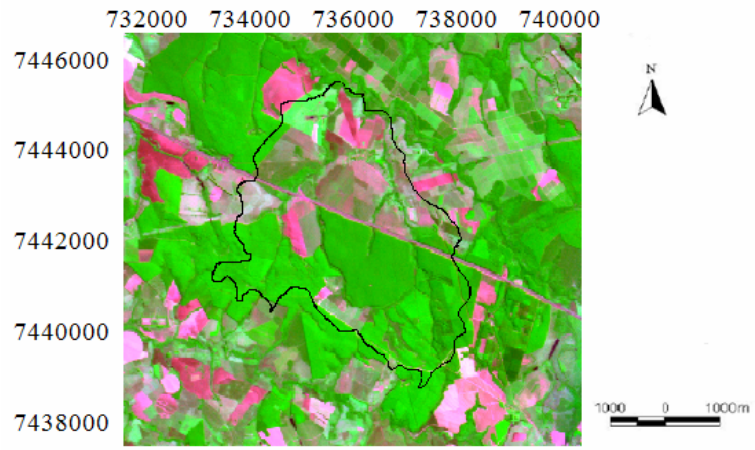

Figure 3: Acquired image from CBERS 2B CCD covering the area of Rio das Pedras watershed (bounded region), Itatinga city, São Paulo State, Brazil. In the x and y-axis one can see the UTM coordinates.

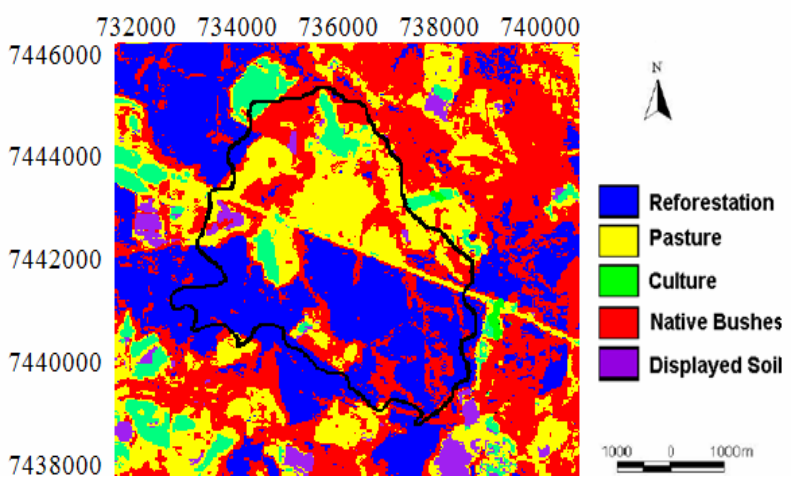

Figure 4: Image classified by the OPF algorithm.

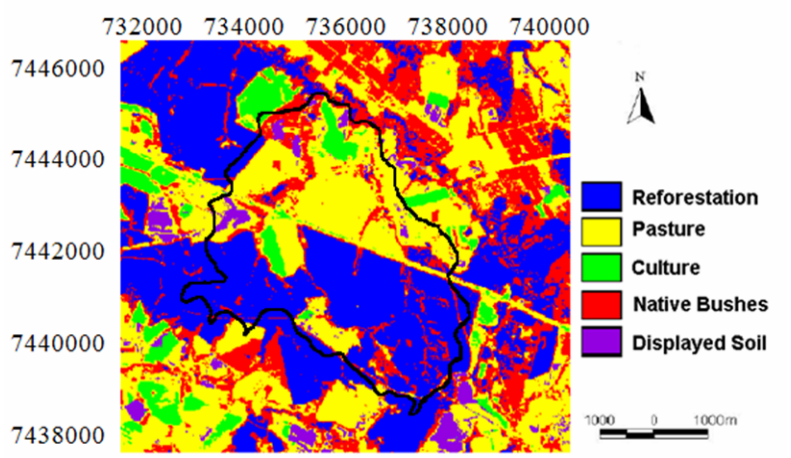

Figure 5: Image classified by the Maximum Likelihood algorithm implemented in Idrisi.

In the middle right location of the images we can distinguish some interesting results provided by OPF, which identified more red regions (native bushes) than the Maximum Likelihood algorithm. There no exists reforestation regions in these locations, but the Idrisi algorithm classified them as belonging to the blue class (reforestation). With respect to the other classes, both algorithms had similar classification results.

4. Conclusions

It was introduced here the Optimum-Path Forest classifier in the agrarian sciences research field, by recognizing the land use for environmental management purposes. We also compared the OPF classifier against ANN-MLP, BC and SVM. As far we know, we are also the first that applied the SVM classifier in this context.

The results demonstrated that OPF and SVM had similar results and outperformed the other ones used, but the first was approximately 65 times faster than SVM, which is an important characteristic nowadays, in which we have a large amount of available satellite data to be processed. We also presented some qualitative results for visual purposes, in which the OPF classifier was compared with the Maximum Likelihood algorithm and outperformed it. We intend to extend this work for land use classification with more classes, such that urbanized areas, marginal bushes, roads and different cultures.

References

Allène C., Audibert J. Y., Couprie M., Cousty J. and Keriven R.. Some Links between Min-cuts, Optimal Spanning Forests and Watersheds. **Mathematical Morphology and its Applications to Image and Signal Processing**, MCT/INPE , pages 253--264, 2007.

Bennetti, A; Bidonne F. O Meio Ambiente e os Recursos Hídricos. In: Tucci, C. M. R. (Org.). **Hidrologia: Ciência e Aplicação.** Porto Alegre: Editora da Universidade do Rio Grande do Sul, UFGRS, ABHR, 2001.

Chang, C. C. and Lin, C. J. LIBSVM: A Library for Support Vector Machines. Software available at url http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2001.

Cormen T., Leiserson C. and Rivest R. Introduction to Algorithms. **MIT**, 1990.

Daianese, R.C. **Sensoriamento Remoto e Geoprocssamento aplicado ao estudo temporal do uso da terra e na comparação entre classificação não - supervisionada e análies visual**. 2001. 186 p. Dissertação (Mestrado em Agronomia). Universidade Estadual Paulista – Faculdade de Ciências Agronômicas, Botucatu. 2001 **.**

Duda., R.O., Hart, P.E. and D.G. Stork. Pattern Classification. Wiley-Interscience, 2 edition, 2000.

Falcão A.X., Stolfi J. and Lotufo R.A. The Image Foresting Transform: Theory, Algorithms, and Applications. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 26(1):19--29, 2004.

Haykin, S. Neural networks: a comprehensive foundation. Prentice Hall, 1994.

Jain, A. K.; Duin R. P. W. and Mao J. Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence.** Vol. 22, No. 1, pp. 4-37, 2000.

Miranda P.A.V., Falcão A.X., Rocha A. and Bergo F.P.G. Object delineation by κ-connected components. **EURASIP Journal on Advances in Signal Processing**, 2008.

Nissen S. Implementation of a Fast Artificial Neural Network Library (FANN). **Department of Computer Science University of Copenhagen (DIKU)**. Software available *at http://leenissen.dk/fann/*. 2003.

Papa J.P., Falcão A.X., Suzuki C.T.N. and Mascarenhas N.D.A. A Discrete Approach for Supervised Pattern Recognition. **12th International Workshop on Combinatorial Image Analysis**, pages 136--147, 2008. LNCS Springer Berlin/Heidelberg.

Papa J.P., Suzuki C.T.N. and Falcão A.X. LibOPF}: A library for the design of optimum-path forest classifiers. Software version 1.0 available at `http://www.ic.unicamp.br/~afalcao/LibOPF`, 2008.

Rubner T., Tomasi C. and Guibas, L. J. A Metric for Distributions with Applications to Image Databases. **Proceedings of the 6th International Conference on Computer Vision**, pages 59, 1998.

Rocha, J.S.M. da. **Manual de Manejo Integrado de Bacias Hidrográficas**. 2.d Santa Maria - RS: UFSM, 1991. 181 p.

Rosa, R. **Introdução ao Sensoriamento Remoto**. Uberlândia - MG: Editora da Universidade Federal de UberlândiaBlucher LTDA, 2003. 228 p.

Vapnik, V. The Nature of Statistical Learning Theory. **NY Springer**. 1995.

Wang Y.P. and Pavlidis T.. Optimal Correspondence of String Subsequences. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 12(11):1080--1087, 1990.