ISPRS Journal of Photogrammetry and Remote Sensing 87 (2014) 241–248

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Review Article

An innovative support vector machine based method for contextual image classification

Rogério Galante Negri *, Luciano Vieira Dutra [1], Sidnei João Siqueira Sant'Anna [2]

Instituto Nacional de Pesquisas Espaciais (INPE), Divisão de Processamento de Imagens (DPI), 12227-010 São José dos Campos, SP, Brazil

ARTICLE INFO

Article history:
Received 11 March 2013
Received in revised form 6 November 2013
Accepted 8 November 2013

Keywords:
Image classification
Contextual information
Support vector machine

ABSTRACT

Several remote sensing studies have adopted the Support Vector Machine (SVM) method for image classification. Although the original formulation of the SVM method does not incorporate contextual information, there are different proposals to incorporate this type of information into it. Usually, these proposals modify the SVM training phase or make an integration of SVM classifications using stochastic models. This study presents a new perspective on the development of contextual SVMs. The main concept of this proposed method is to use the contextual information to displace the separation hyperplane, initially defined by the traditional SVM. This displaced hyperplane could cause a change of the class initially assigned to the pixel. To evaluate the classification effectiveness of the proposed method a case study is presented comparing the results with the standard SVM and the SVM post-processed by the mode (majority) filter. An ALOS/PALSAR image, PLR mode, acquired over an Amazon area was used in the experiment. Considering the inner area of test sites, the accuracy results obtained by the proposed method is better than SVM and similar to SVM post-processed by the mode filter. The proposed method, however, produces better results than mode post-processed SVM when considering the classification near the edges between regions. One drawback of the method is the computational cost of the proposed method is significantly greater than the compared methods.

1. Introduction

The process of classifying remote sensing images consists of extracting information from image pixels for automatic identification of targets among different thematic classes (Khedam et al., 2003). This process is commonly conducted using pointwise classifiers. With the emergence of imaging sensors that are able to acquire images with high spatial and spectral resolution, historically used approaches are no longer satisfactory in certain cases (Zortea et al., 2007; Besbes et al., 2009). An alternative approach for such cases is contextual classifier techniques. In this type of classifier, the information is expressed in terms of the spatial relationships between a pixel and its neighbors. This information is incorporated in the classification process (Gurney and Townsend, 1983).

According to Mather (2004), the easiest method to perform contextual classification consists of applying a mode filter to the labels of a pointwise classification result. Another method to include contextual information in the classification process is to apply stochastic models based on Markov Random Fields (MRF) theory. This method became popular after Besag's work (Besag, 1986), in which the Iterated Conditional Modes (ICM) algorithm was presented.

Introduced by Vladmir Vapnik, SVM is a Pattern Recognition method that has surpassed many other methods in a wide variety of applications (Cristianini and Shawe-Taylor, 2000). However, the original formulation of the SVM method does not incorporate contextual information in the classification of digital images. Different extensions of the SVM method with contextual architecture classification can be found in the literature. In Bovolo and Bruzzone (2005), the decision rule in the SVM method is transformed into a probability distribution and is then integrated in a MRF architecture, the solution of which is obtained using Simulated Annealing (SA) (Geman and Geman, 1984), Maximization of the Posterior Marginals (MPM) (Marroquin et al., 1987), and ICM algorithms (Besag, 1986). In Bruzzone and Persello (2009) and Bovolo et al. (2006), modifications to the original formulation of the SVM method are proposed to include contextual information in the classification process. It is noteworthy that the application of the proposals from Bovolo and Bruzzone (2005) and Bruzzone and Persello (2009) to multiclass classifications are tied to specific multiclass strategies.

* Corresponding author. Tel.: +55 12 3208 6469.
E-mail addresses: rogerio@dpi.inpe.br (R.G. Negri), dutra@dpi.inpe.br (L.V. Dutra), sidnei@dpi.inpe.br (S.J.S. Sant'Anna).
[1] Tel.: +55 12 3208 6514.
[2] Tel.: +55 12 3208 6509.

0924-2716/$ - see front matter © 2013 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.11.004

This study presents a new version of the SVM method for contextual image classification. Unlike the proposals previously mentioned, neither of the modifications of the original formulation of the SVM method, the use of stochastic techniques, or the use of filtering techniques are adopted. Instead, a local adjustment of the separating hyperplane based on contextual information is adopted.

This article is organized as follows: in Section 2, fundamental ideas regarding image classification and SVM are presented. The proposed contextual SVM method is described in Section 3. In Section 4, the proposed contextual image classification method is applied on a case study. Conclusions are presented in Section 5.

2. Image classification

2.1. General concepts

A classifier is represented by a function $F : \mathcal{X} \to \mathcal{Y}$ that associates elements from the attribute space $\mathcal{X}$ to a class of $\{\omega_1, \omega_2, \dots, \omega_c\}$, where $c \in \mathbb{N}^*$, from a given label (indicator) of classes in $\mathcal{Y} = \{1, 2, \dots, c\}$. Under these conditions, for $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y},\ y = F(\mathbf{x})$ indicates that $\mathbf{x}$ belongs to class ω_y. Note that vector variables are represented in bold.

Image classification consists of applying F to the patterns (the attribute vector in a pixel s) that make up an image $\mathcal{I}$, which is defined over a grid (support) $\mathcal{S} \subset \mathbb{N}^2$, for which the classification result can be denoted by $F(\mathcal{I})$. With regard to the image that undergoes the classification process, $\mathcal{I}(s) = \mathbf{x}$ denotes that the pixel $s \in \mathcal{S}$ of $\mathcal{I}$ has attributes represented by $\mathbf{x} \in \mathcal{X}$. Additionally, the positions occupied by the neighboring pixels of s are given by the set $\mathcal{V}_\rho(s) = \{t \in \mathcal{S} : 0 \leqslant md(s,t) \leqslant \rho\}$, where ρ is represented by *the radius of the neighborhood influence* and $md(\cdot,\cdot)$ is a distance measure. In this study, we use the *maximum distance*.[3] Under these conditions, is defined as context of s the set of patterns and the respective classifications associated to the pixels contained in $\mathcal{V}_\rho(s)$.

The estimate of F is conducted according to the learning paradigm of the classification method. The classifiers that use supervised learning produce estimations of F from information that is extracted from a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \dots, m\}$ with $m \in \mathbb{N}^*$, where y_i is the indicator of class ω_{y_i}, and $\mathbf{x}_i$ is a attribute vector.

2.2. SVM

SVM is a recently developed method of Pattern Recognition with supervised learning. This method has recently received significant attention from the scientific community because of its characteristics, which include independence of the statistical distribution models, simple architecture algorithms, moderated computational complexity, excellent generalization ability and robustness to Hughes' phenomenon (Zortea et al., 2007; Bruzzone and Persello, 2009). The classification process using SVM consists of distinguishing patterns using hyperplanes for which the separating margin is maximal. A separating hyperplane is equal to the geometric place where the value of the function $f_{SVM} : \mathcal{X} \to \mathbb{R}$ becomes null:

$$f_{SVM}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \tag{1}$$

where $\mathbf{w}$ is the orthogonal vector to the separating hyperplane $f_{SVM}(\mathbf{x}) = 0$ and b is a scalar such that $|b|/\|\mathbf{w}\|$ represents the distance from the hyperplane to the origin of the attribute space. The notations $|\cdot|, \|\cdot\|$ and $\langle\cdot,\cdot\rangle$ represent the absolute value, vector norm, and inner vector product, respectively. The parameters $\mathbf{w}$ and b are obtained by solving the following optimization problem in its dual formulation using the patterns in a training set $\mathcal{D}$ (Theodoridis and Koutroumbas, 2008):

$$\begin{aligned} &\max_{\eta} \sum_{i=1}^{m} \eta_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \eta_i \eta_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ &\text{subject to} : \begin{cases} 0 \leqslant \eta_i \leqslant C, & i = 1, \dots, m \\ \sum_{i=1}^{m} \eta_i y_i = 0, \end{cases} \end{aligned} \tag{2}$$

where η_i are Lagrange multipliers, $\mathcal{Y} = \{-1, +1\}$ is the set of class indicators, and the parameter C is introduced to treat non-separable classes and acts as a misclassification penalty during the training stage. When the value of C is increased, fewer incorrect classifications are permitted. The inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ in (2) can be replaced by symmetric functions (i.e., $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$), which meet Mercer's conditions (Schölkopf and Smola, 2002). These functions are called kernel functions. These functions implicitly compute the inner product between patterns mapped onto a higher-dimensional space for which the separation can be conducted linearly. Examples of kernel functions include the following:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1\right)^q$
- Radial Basis: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$,

where $q \in \mathbb{N}$ and $\sigma \in \mathbb{R}$.

Let SV be the set of support vectors, which are determined through optimization of (2), such that $SV = \{\mathbf{x}_i : \eta_i \neq 0; i = 1, 2, \dots, l \leqslant m\}$. The parameter $\mathbf{w}$ is computed as $\sum_{\mathbf{x}_i \in SV} \eta_i y_i \mathbf{x}_i$, whereas b equals $1 - \langle \mathbf{w}, \mathbf{x}_i \rangle$ for any $\mathbf{x}_i \in SV$ such that $f_{SVM}(\mathbf{x}_i) = 1$. After the determination of the parameters of $f_{SVM}(\mathbf{x})$, the decision rule in the SVM method is given by $F_{SVM} : \mathcal{X} \to \mathcal{Y}$ such that $F_{SVM}(\mathbf{x}) = sgn(f_{SVM}(\mathbf{x}))$, where $sgn(\cdot)$ is the sign function. Under these conditions, $\mathbf{x}_i$ is associated with class ω_1 if $F_{SVM}(\mathbf{x}) = +1$ or class ω_2 otherwise.

According to the definitions given, it can be noted that the separating hyperplane is defined by the set of support vectors. The support vectors occupy positions in the attribute space located on the boundaries or on the separating margin. Training vectors are also considered support vectors when they are mistakenly classified in the training process. Considering that the separating margin is a transition zone of the *decision rule*, it is concluded that the hyperplane is determined from training patterns that have low classification reliability.

Thus, it is reasonable to assume that the *reliability* of a pattern classification $\mathbf{x}_i$ corresponds to the value of $|f_{SVM}(\mathbf{x}_i)|$, which equals the distance between $\mathbf{x}_i$ and the separating hyperplane $f_{SVM}(\mathbf{x}) = 0$. Thus after classification, patterns far from the hyperplane will have greater classification *reliability*. Fig. 1 shows the relationship

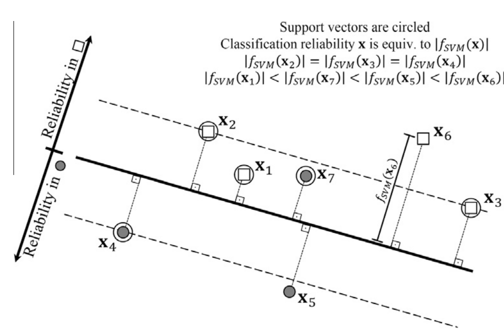

Fig. 1. The relationship between the classified vector distance and its classification reliability.

[3] Let $s, t \in \mathbb{N}^2$ such that $s = (s_1, s_2)$ and $t = (t_1, t_2)$ The measure $md(s,t) = \max\{|s_1 - t_1|, |s_2 - t_2|\}$ is called *maximum distance*.