A COMPARISON OF MULTI-CLASS SVM STRATEGIES AND KERNEL FUNCTIONS FOR LAND COVER CLASSIFICATION

Willian Vieira de Oliveira[1], Luciano Vieira Dutra[2], and Sidnei João Siqueira Sant'Anna[3]

Brazilian National Institute for Space Research – INPE, São José dos Campos, São Paulo, Brazil
{[1]willian.oliveira; [2]luciano.dutra; [3]sidnei.santanna}@inpe.br

ABSTRACT

Support Vector Machines (SVMs) are powerful machine learning algorithms originally proposed for solving linear and binary problems and, later, extended to perform non-linear and multi-class tasks. In remote sensing applications, SVMs have been widely applied to land cover classification. However, SVMs are highly sensitive to the choice of the kernel function and its parameters. These elements have a direct influence on the classification accuracy. The purpose of this study is to assess the performance of the SVM classifier when combined with distinct kernel functions and multi-class approaches for land cover classification. We carried out experiments using a multispectral image of a highly urbanized area. The experimental results demonstrated the efficiency of the SVM classifier with the radial basis function for land cover classification. In this study, the type of multi-class approach did not present a significant impact on the SVM performance when combined with this kernel function.

Key words **—** SVM. Land cover classification. Kernel functions. Multi-class approaches.

1. INTRODUCTION

Support Vector Machine (SVM) [1] is a nonparametric pattern recognition algorithm that has been widely applied to numerous applications, including the classification of remote sensing data. SVM uses training samples to construct an optimal separating hyperplane (i.e., decision boundary) between two classes, based on a structural risk minimization strategy [2], [3]. This method only requires the samples that lie on the edge of the classes in feature space, defined as support vectors, to obtain the separating hyperplane that maximizes the margin between the classes. Despite its simple architecture, SVM presents high generalization power and low sensitivity to the Hughes phenomenon [3]-[6]. For that reason, SVM methods normally perform well in classification problems that involve high dimensional data, such as hyperspectral images, or tasks with limited training samples.

SVM was originally proposed to deal with linear and binary problems. Diverse modifications were proposed in literature over the last decades in order to apply SVM also to non-linear and multi-class tasks. For non-linear cases, SVM can use kernel functions to project the data into a feature space of higher dimensionality, where the classes are linearly separable. In addition, strategies to overcome the binary restriction and perform multi-class classifications include the one-against-one (OAO) and one-against-all (OAA) strategies. In general, these approaches aim to divide a multi-class problem into a series of binary SVM classifications.

The high sensitivity to the choice of hyper-parameters is one of the major limitations of SVMs [2]. There are several possible combinations of kernel functions and parameters that may be chosen for classification. The selection of an appropriate setup has a direct influence on the classifier's performance. For instance, distinct kernel functions normally produce different separating hyperplanes, which may cause significant variations in the classification accuracy. For that reason, SVM usually requires the use of strategies to establish optimal parameters, such as grid-search based approaches [3]. However, these methods might present a high computational cost, particularly for problems that involve a large number of classes.

In this paper, we investigate the influence of the choice of the kernel function and its parameters on the SVM performance for land cover classification, when combined with distinct multi-class approaches. We applied SVM with distinct classification setups to classify a multispectral image from the Zurich Summer dataset [7]. We focus on the standard OAO and OAA multi-class approaches, and the linear, polynomial, radial basis and sigmoid kernel functions.

2. SUPPORT VECTOR MACHINES

In its simplest version, SVM is a binary classifier that aims to determine an optimal separating hyperplane between two classes (ω_1 and ω_2) based on their geometric distribution in the available feature space. A hyperplane is considered optimal when it separates the data in such a way that its distance from the training data of each class is as large as possible [2]. This hyperplane is defined by the function:

$$f(x) = w \cdot \boldsymbol{x} + b = 0$$

where, $w \cdot \boldsymbol{x}$ is the inner product between the normal to the hyperplane (w) and the sample vector ($\boldsymbol{x}$), and b is the offset that denotes the closest distance to the origin. The distance of the hyperplane from the origin is $b/\|w\|$, where $\|w\|$ is the Euclidean norm of w.

Consider a training data set $S = \{(x_i, y_i): i = 1, \dots, N, y_i \in \{-1, +1\}\}$, where x_i is a sample instance and y_i is its class indicator. The indicator $y_i = +1$ establishes that $x_i \in \omega_1$ and $y_i = -1$ defines that $x_i \in \omega_2$. For the case of linearly separable data, the hyperplane that defines each class can be

generalized by the inequality $y_i(w \cdot x_i + b) - 1 \geq 0$. The parameters w and b are determined by solving the following quadratic optimization problem [2], [6]:

$$\max_y \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$subject\ to \begin{cases} 0 \leq \lambda_i \leq C, i = 1, \ldots, N; \\ \sum_{i=1}^{N} \lambda_i y_i = 0 \end{cases}$$

where, λ are Lagragian multipliers, which have an upper bound of C. The C term is a penalty parameter introduced to handle non-separable data (i.e., soft margin). It controls the trade-off between margin and misclassifications and, therefore, the generalization capabilities of the classifier. Note in the optimization process that there is a Lagrange multiplier λ_i for every training sample. The set of samples with $\lambda_i > 0$ are defined as support vectors.

In addition to the soft margin strategy, the inner product $(x_i \cdot x_j)$ in Equation (2) can also be replaced by kernel functions $K(x_i, x_j)$ in order to classify non-linearly separable classes. The kernel function projects the data into a higher dimensional space [3], where the classes can be separated with a hyperplane. Examples of kernel functions commonly used in remote sensing applications include [2]:

- Polynomial: $K(x_i, x_j) = (\gamma(x_i \cdot x_j) + r)^d$;
- Radial basis function: $K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2)$;
- Sigmoid: $K(x_i, x_j) = tanh(\gamma(x_i \cdot x_j) + r)$;

where, exp and $tanh$ indicate the exponential and hyperbolic tangent functions, respectively. The term d is the degree of the polynomial function, γ is a positive parameter that modifies the flexibility of the hyperplane, and r is an independent term in the kernel functions. These terms are user-defined parameters that have a significant impact on the classifier's performance.

Tasks that involve more than two classes require the use of multi-class approaches, such as the One-Against-All and One-Against-One strategies, which decompose the classification into multiple two-class sub-problems. In a task with K classes, defined by $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$, the OAA strategy produces K binary SVM models, where each SVM analyses a given class ω_k against the remaining $K - 1$ classes. In the OAA strategy, the problem becomes unbalanced. The SVM classifiers perform training with significantly more negative than positive samples [6], which might reduce the prediction efficiency of the class of interest. The OAA classification rule is as follows:

$$assign\ \boldsymbol{x}\ to\ \omega_y \Leftrightarrow y = \arg \max_{k=1,\ldots,K} [w_k \cdot \boldsymbol{x} + b_k]$$

where, w_k and b_k are the w and b parameters computed for the k^{th} SVM classifier.

In contrast, the OAO rule analyses all possible pairwise class combinations, which produces $\frac{K(K-1)}{2}$ binary SVM models. Although OAO requires a larger number of classifiers than OAA, an advantage of this strategy is that the problem remains balanced if the dataset is balanced. A majority voting scheme is often applied to determine the output classification. However, it may result in ambiguous regions. Instead, the classification rule of the OAO approach can be defined as follows:

$$assign\ \boldsymbol{x}\ to\ \omega_y \Leftrightarrow y = \arg \max_{k=1,\ldots,K} \sum_{i=1}^{K} [w_{k,i} \cdot \boldsymbol{x} + b_{k,i}]$$

where, $w_{k,i}$ and $b_{k,i}$ are the w and b parameters computed for the classifier created to distinguish between ω_k and ω_i.

3. MATERIAL AND METHODS

3.1. Datasets

In this paper, we carried out experiments with a multispectral image from the Zurich Summer dataset [7]. The Zurich Summer dataset includes a collection of 20 images taken from a QuickBird scene acquired over Zurich, Switzerland, in August 2002. These images, obtained with a spatial resolution of 0.61 meters, are composed of four spectral channels from the near-infrared (NIR) to the visible (RGB) spectrum. Figure 1 illustrates the true-colour composition of the image analysed in this study, which represents a highly urbanized region.

This image presents six distinct urban classes, including: roads, buildings, trees, grass, bare soil, and water. This image represents building rooftops composed of distinct materials. This characteristic normally leads to a multimodal data distribution, which might represent a challenge for determining optimal classification parameters.

3.2. Experimental setup

The SVM classifier was applied to classify the Zurich image with distinct kernel functions and multi-class approaches. Reference samples were randomly extracted from the reference image illustrated in Figure 1, in order to define balanced sample datasets. We selected 13,000 samples per-class for training the models and 4,000 samples for testing their performance. We evaluated the producer's and user's accuracies (PA and UA, respectively) to assess the accuracy obtained for each class

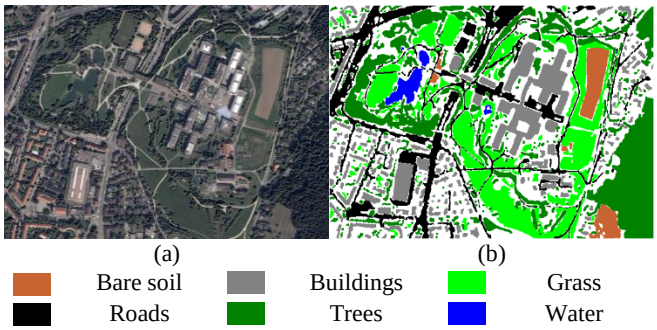

Figure 1. (a) True-colour composition of an image from the Zurich dataset, and (b) the reference image.

individually, in addition to global accuracy metrics, such as the overall accuracy (OA) and the Kappa index.

We generated multiple SVM classification models using the OAA and OAO multi-class approaches and four types of kernel functions: linear, polynomial, radial basis function (RBF), and sigmoid. The parameter set was composed of: penalty term (C), where $C \in \{0.1, 1, 10, 100, 1000\}$; degree ($d$), where $d \in \{2, 3\}$; gamma (γ), where $\gamma \in \{0.001, 0.01, 0.1, 0.25, 1, 10\}$; and, an independent term (r), where $r \in \{1, 2, 3, 4\}$. The value $\gamma = 0.25$ refers to the scale gamma, computed from the training samples by $1/(n_{feat} * \sigma^2)$, where n_{feat} is the number of features of x, and σ^2 is the variance of the data. In this study, we used the python programming language and the SVM implementation of the Scikit-learn library [8] to perform the classification tests.

4. RESULTS AND DISCUSSION

We performed a total of 790 classification tests, based on the combination of distinct parameters and the OAO and OAA strategies, as described in section 3.2. Figure 2 provides an overview of the variation of performance of the SVM classifier, in terms of OA, when combined with distinct kernel functions and multi-class strategies. This figure represents the minimum, first quartile, median, third quartile and maximum accuracy values. Outliers are not represented. The experimental results suggest that the OAO method can achieve higher overall accuracies than the OAA strategy for most kernel functions. It also shows the sensitivity of each kernel type to the selection of hyper-parameters.

Figure 2 demonstrates why the RBF kernel is frequently the choice of studies that require the classification of land cover classes, which are generally not linearly separable. It presented the highest accuracies estimates, followed by the polynomial kernel, which not only requires a larger number of parameters but also usually presents a higher computation cost. The sigmoid function presented higher sensitivity to the choice of the hyper-parameters than the remaining kernel functions. Although the polynomial kernel requires one more parameter than the sigmoid function, it presented a significantly lower variation of performance. In addition, the

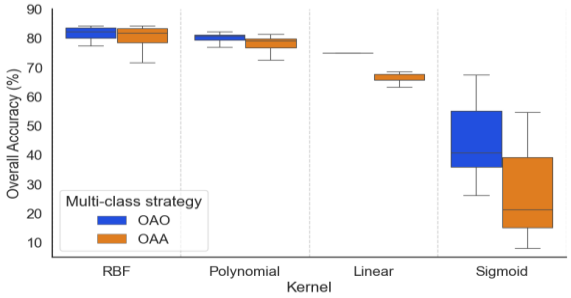

Figure 2. Variation of performance for SVM classifiers with distinct kernel functions and multiclass strategies.

OAA method appears to have a slightly higher sensitivity to parameter selection than the OAO approach.

We analysed the influence of the multi-class strategy and the kernel parameters on the classification performance of individual classes. Figure 3 provides an overview of the accuracies obtained for each class using SVM with the RBF and polynomial kernels, and the OAA and OAO strategies. Both multi-class approaches demonstrated potential to achieve similar accuracies for all the analysed classes. In this case study, we did not observe a significant influence of the multi-class strategy in the performance of the SVM classifier with the RBF and polynomial kernels. It is possible to optimize the kernel parameters in such a way that both the OAO method and the OAA approach provide similar accuracy levels for a given class of interest. On the other hand, the selection of appropriate hyper-parameters for the chosen kernel function can have a significant impact on the classification performance of some classes.

The accuracy of SVM classifiers on individual classes may vary significantly according to the initial penalty and kernel function parameters. Classification parameters used to distinguish between well separable classes might not be suitable for efficiently delineating other classes. This issue becomes more evident in classification tasks that include complex land cover classes, such as classes that are close in the available feature space or classes that present multi-modal data distribution. In this study, for instance, the class buildings includes rooftops composed of distinct materials, such as asphalt, ceramic, concrete and metal, which may increase its confusion with other classes.

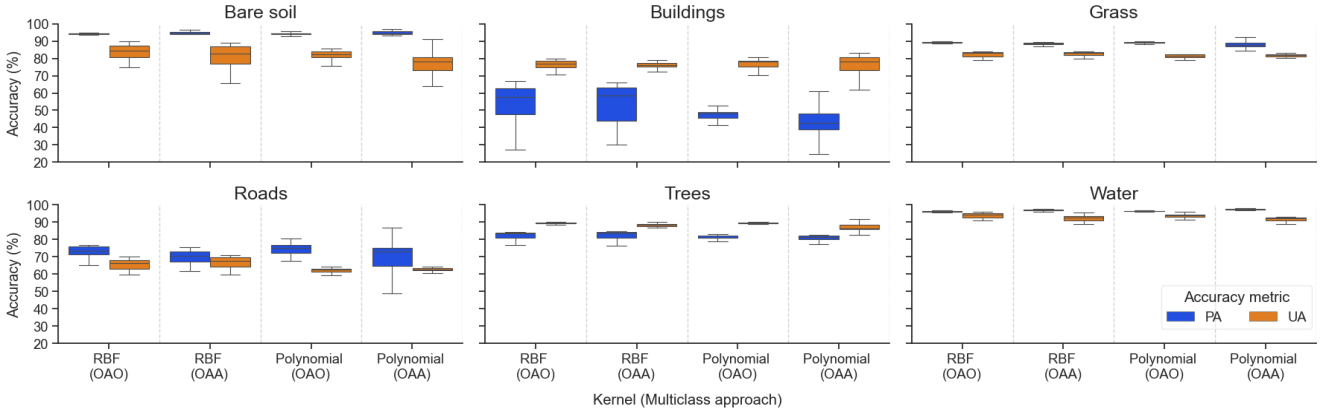

Figure 3. Classification accuracies of individual classes, provided by SVM with distinct kernels and multi-class strategies.

Classes	SVM-1 (OAO)		SVM-2 (OAA)	
	PA	UA	PA	UA
Bare soil	94.0	89.0	94.2	89.0
Buildings	66.8	78.6	65.5	78.7
Grass	89.3	83.7	89.2	83.6
Roads	76.0	70.3	75.5	70.4
Trees	84.3	89.4	84.5	88.8
Water	96.1	95.6	96.7	94.9
Kappa	0.813 ± 0.003		0.811 ± 0.003	

Table 1. Accuracies of two SVM classifiers with RBF kernel: SVM-1 (OAO, $C = 1, \gamma = 10$) and SVM-2 (OAA, $C = 10, \gamma = 10$).

The high influence of the chosen hyper-parameters on the classification performance is one of major limitations of kernel SVM, especially for classification problems that include complex class patterns. Grid search methods are normally required to improve the classification accuracy. However, they are generally time consuming and, depending on the initial parameters, they might lead to a local optimum, which can limit the effectiveness of the resulting SVM classifier [9].

Table 1 presents the classification accuracies estimated in this study for the SVM classifiers that provided the best classification performances, in terms of Kappa, using the OAO and OAA approaches. SVM-1 and SVM-2 use the RBF kernel, and the OAO and OAA multi-class strategies, respectively. Despite the use of different multi-class classification approaches, these SVM instances provided similar global accuracy estimates. Figure 4 illustrates the classification of the SVM-1 classifier. This SVM instance presented similar results to the SVM-2 classifier, but with a better balance between the PA and UA of some classes, leading to slightly higher global accuracy measures.

5. CONCLUSIONS

In this study, we investigated the influence of the choice of the kernel function and the multi-class strategy on the SVM performance for land cover classification. In this particular case study, the best classification performance, in terms of global accuracy measures, was obtained with the RBF kernel and the OAO approach. However, the SVM classifier with OAA method also achieved similar accuracy estimates. The type of multi-class classification approach did not present a significant impact on the SVM performance for the RBF and polynomial kernel functions, especially when optimized for a particular class of interest.

On the other hand, the experimental results confirmed the high sensitivity of SVMs to the choice of the kernel function and its parameters and, consequently, the importance of parameter optimization strategies. The RBF kernel presented a higher efficiency to discriminate complex land cover classes, such as building's rooftops composed of distinct materials, in comparison to the linear, polynomial, and sigmoid kernels. This work is another indicator of the efficiency of the RBF kernel for land cover classification.

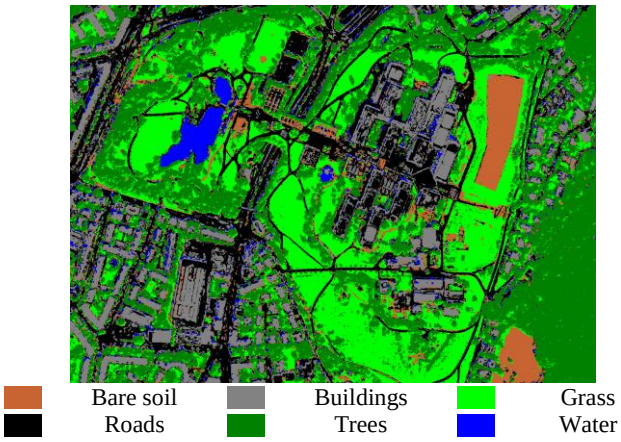

Figure 4. SVM classification result obtained with the RBF kernel ($C = 1$; $\gamma = 10$) and the OAO strategy.

6. ACKNOWLEDGEMENTS

This study was partially supported by the Coordenação de Aperfeiçoamento de Nível Superior (CAPES), Finance Code 001, and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), PQ grant #309135/2015-0.

7. REFERENCES

[1] V. Vapnik. *Estimation of dependences based on empirical data*. Springer, New York, 2006.

[2] B. Tso, and P. Mather. *Classification methods for remotely sensing data analysis.* 2nd ed. CRC Press, Boca Raton, 2009.

[3] G. Camps-Valls, and L. Bruzzone. *Kernel methods for remote sensing data analysis*. 1st ed. John Wiley & Sons, Singapore, 2009.

[4] R. G. Negri, L. V. Dutra, C. C. Freitas, and D. Lu. Exploring the capability of ALOS PALSAR L-band fully polarimetric data for land cover classification in tropical environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 9, n.12:pp. 5369-5384, 2016.

[5] G. Mountrakis, J. Im and C. Ogole. Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 66, n. 3:pp. 247–259, 2011.

[6] S. Theodoridis, and K. Koutroumbas. *Pattern recognition.* 4th ed. Academic Press, London, 2009.

[7] M. Volpi, and V. Ferrari. Semantic segmentation of urban scenes by learning local class interactions. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, *Proceedings*… Boston, 2015.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12:pp. 2825-2830, 2011.

[9] C. F. Chao, and M. H. Horng. The construction of support vector machine classifier using the firefly algorithm. *Computational Intelligence and Neuroscience*, v. 2015, n. 2:pp. 1-8, 2015.