AN APPLICATION OF MULTIPLE SPACE NEAREST NEIGHBOR CLASSIFIER IN LAND COVER CLASSIFICATION

Flávia de Toledo Martins-Bedê, Mariane Souza Reis, Eliana Pantaleão, Luciano Dutra, Sandra Sandri

Brazilian National Institute for Space Research (INPE)

{flavinha, reis, elianap, dutra}@dpi.inpe.br, sandra.sandri@inpe.br

ABSTRACT

This work presents a case study in land cover classification using ms-NN, an extension of k-NN classification algorithm. The case study focuses on an area in the Brazilian Amazon region, with data obtained from LANDSAT5 satellite Thematic Mapper (TM) sensor and Advanced Land Observing System satellite (ALOS) Phase Array L-Band Synthetic Aperture Radar (PALSAR), using Fine Beam Dual. The results obtained with ms-NN are compared with k-NN and Support Vector Machine algorithms, considering the use of a single training set, a Monte Carlo procedure for testing and an extensive number of parameterizations for the classification methods. Considering only the best results for each classifier, ms-NN obtained better results than the other methods.

Index Terms— land cover classification, multiple space nearest neighbor, classification algorithm, SAR and optical data classification.

1. INTRODUCTION

The k nearest neighbor (k-NN) classifier is one of the most popular classification and pattern recognition techniques. In this work, we propose the use of an extension to k-NN, called Multi Space Nearest Neighbors (ms-NN), in land cover classification. In this framework, the complete set of attributes is partitioned into several spaces, including a geographical one if needed [1].

This work presents a case study in land cover classification of an area in the Brazilian Amazon region, comparing the results obtained by ms-NN with k-NN and Support Vector Machine (SVM), considering the use of a single training set, a Monte Carlo procedure for testing and an exhaustive number of parameterizations for the methods.

2. MULTIPLE SPACE NEAREST NEIGHBOR

In k-NN, an unlabeled object receives the label of the majority of its k nearest neighbors. In the simplest version of k-NN, k is equal to 1 and, in this case, the algorithm is known as nearest neighbor classifier (NN) [2].

In order to classify an unlabeled sample c in a multi-dimensional domain Ω, k-NN algorithm calculates the distance in the feature space from c to all labeled samples in the training base. Then, the k smallest values are selected and c is assigned to the majority class of its k nearest neighbors. In ms-NN algorithm, there is the possibility of using multiple spaces, that can be originated from different data sources and have different ranges of values. It is also possible to use the geographic space, when the neighbor samples are either pixels or polygons. The use of the geographical space can be very useful in applications involving geo-referenced objects. It is also possible to use both the location of the object and its actual geometry as attributes for classification, thus allowing the use of topological associations.

In ms-NN, a distance function is associated to each space, as well as a neighborhood type (fixed or variable). The class of an unlabeled sample is taken from the union of the neighbors calculated from all spaces, using a predominance function, that can be the simple majority, as in traditional k-NN, or others, in particular, weighted ones. It is possible to set distinct distance functions for different spaces, such as Euclidean, Mahalanobis, Hamming and others, such as those based on fuzzy relations [3] [4].

This methodology can be seen as generalization of [5], in which it is proposed the use of a NN classifier from multiple feature subsets (MFS). MFS aims to improve classification accuracy by combining outputs from multiple NN classifiers by simple vote, with each classifier having access only to a random subset of features.

3. MATERIALS AND METHODS

The area of interest in this study, of approximately 411 km^2, is situated in the Tapajós National Forest and its surroundings, located in Belterra, in the state of Pará, in the Brazilian Amazon region (Figure 1). This area is classified in [6] as Humid Tropical Rainforest, characterized by large tress, woody lianas, epiphytes and palms. Embedded in the primary forest matrix, there exist mosaics of secondary vegetation, pasture, deforested areas and agricultural areas, large and small [7].

Funded by CNPq grants # 151201/2014-5, 150439/2014-8, 307666/2011-5.

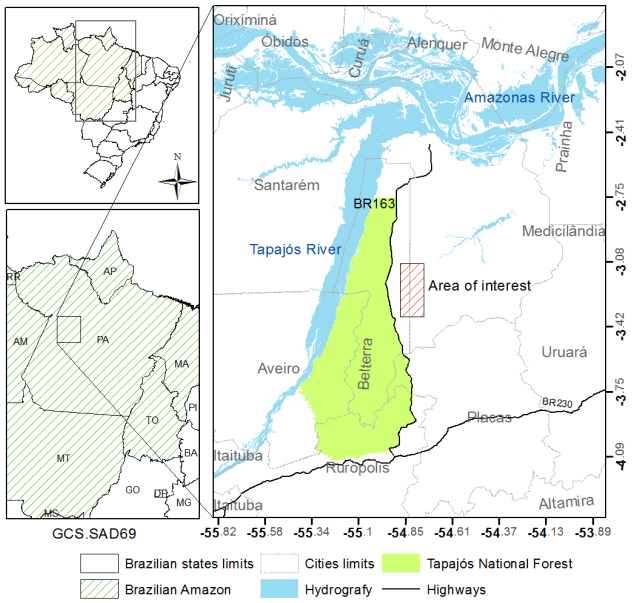

Fig. 1. Location of study area.

Two images have been used in order to compose the attribute space needed by the classifiers. The first one is a spatial subset of an image from LANDSAT5 satellite Thematic Mapper (TM) sensor, imaged in June 29th, 2010, bands 3 to 5. The second one is from the same area, acquired in June 21st, 2010, by the Advanced Land Observing System satellite (ALOS) Phase Array L-Band Synthetic Aperture Radar (PALSAR), using Fine Beam Dual (FBD). Both images were previously geometrically corrected, in order to be co-registered in the same projection and reference system (UTM WGS84, zone 21S). The ALOS-PALSAR image was used with the data in amplitude format in two polarizations (HH and HV), Estimated Number of Looks equal to 5 and pixel size 15mx15m. The soil, vegetation and shadow fractions of TM image have also been used, and were calculated using bands 1 to 5 and 7 of the original image and the method described in [8].

A segmented image was generated from LANDSAT5/TM bands 3, 4 and 5 (normalized to mean 127 and standard deviation 42) using Multiresolution Segmentation, from e-Cognition. Set parameters were scale=15, shape=0.3 and compacity=0.5. Then, mean values of the eight attribute bands were calculated for each segment. The entropy, homogeneity and dissimilarity of pixels values per segment were also calculated in the original 3, 4 and 5 bands from the TM image.

The data were classified using ms-NN, k-NN and SVM (with polynomial kernel) [2], with various parameterizations, in a region based approach. The classifiers were trained using six land cover classes (Forest, Cultivated Areas, Initial Regeneration, Advanced Regeneration, Bare Soil and Pasture), described as follows:

- Forest (FP) refers to mature (primary) forest;
- Cultivated Areas (AC) refers to agricultural areas with grown crops;
- Initial Regeneration (RI) comprises secondary vegetation in both initial and intermediate states of development;
- Advanced Regeneration (FA) refers to secondary vegetation in advanced state of development;
- Bare Soil (SE) refers to areas constituted basically by bare soil, like those prepared for plantation or recently deforested ones;
- Pasture (PA) refers to areas with typical pasture vegetation.

Considering ground knowledge, 428 polygons were handdrawn. Of these polygons, 2/3 were used for training and 1/3 for testing (Figure 2).

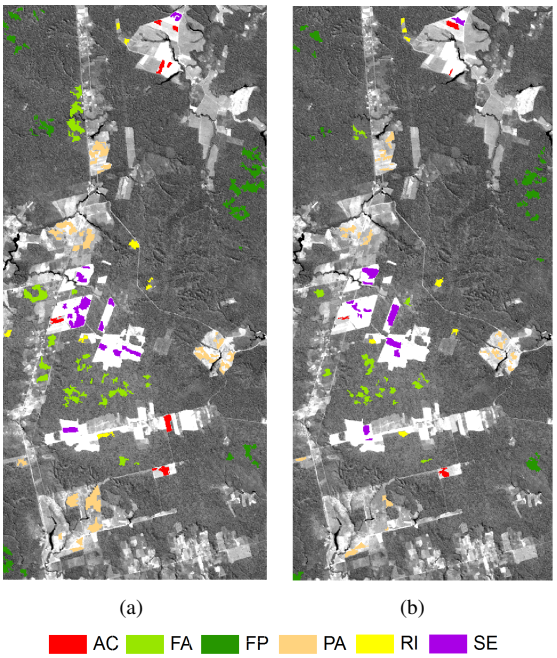

Fig. 2. Labeled samples distribution, superposed to LANDSAT5/TM image band 5, for (a) training and (b) testing.

The obtained classified images were evaluated using Overall Accuracy Index (OA), calculated by the draws of 900 pixels from the test samples, using a Monte Carlo approach, with 1000 repetitions. To choose the best classification and in order to assess if there is any statistic indication that they are similar, classification results from the same classifier were

compared using paired t-testing. The best results from each method were then compared in a paired manner using the Performance Index [1]. This index is used to account how many times one classifier is better than another, comparing their OAs, in percentage.

4. RESULTS

The best result for SVM was obtained, among 18 parameterizations, with the use of third-degree polynomial kernel and penalty equal to 10. Of the 16 k-NN classifications, the one using 1 neighbor was selected. For ms-NN classifications, tests considered 2 to 4 spaces, varying the number of neighbors in each space from 1 to 16 for 2 and 3 spaces, and from 1 to 10 for 4 spaces. In all cases, two different distance functions were used to calculate the neighbors: Euclidean (d_E) and Mahalanobis (d_M), resulting in a total of 53,200 classifications. For each number of spaces (2, 3 and 4) two classifications were selected, one using d_E and another using d_M. Table 1 shows the mean and standard deviation of OA for each selected classification.

Table 1. Mean OA of selected classifications.

Classifier	Mean	Standard Deviation
SVM	0.68	0.011
k-NN	0.88	0.008
ms-NN2\d_E	0.92	0.009
ms-NN2\d_M	0.94	0.007
ms-NN3\d_E	0.94	0.007
ms-NN3\d_M	0.95	0.007
ms-NN4\d_E	0.93	0.008
ms-NN4\d_M	0.91	0.008

Table 1 shows that the highest mean value of OA was achieved by ms-NN3\d_M. Results show that ms-NN outperforms SVM and k-NN classifiers for extensive parameterizations. When comparing only SVM and k-NN classifiers, k-NN presents the highest mean OA. However, all selected ms-NN classifications have mean OA higher than k-NN.

The Performance Index was used to assess how many times the best result (ms-NN3\d_M) was better than each of the others. Figure 3 presents the obtained values for the performance index. The index value for ms-NN3\d_M is zero because its OA value is never higher than itself. The dashed line was drawn at 50% because above, this value, the best result for ms-NN can be considered statistically better than the one being analyzed. In this case, the best classification (ms-NN3\d_M) was better than the second best (ms-NN3\d_E) in 90% of the cases.

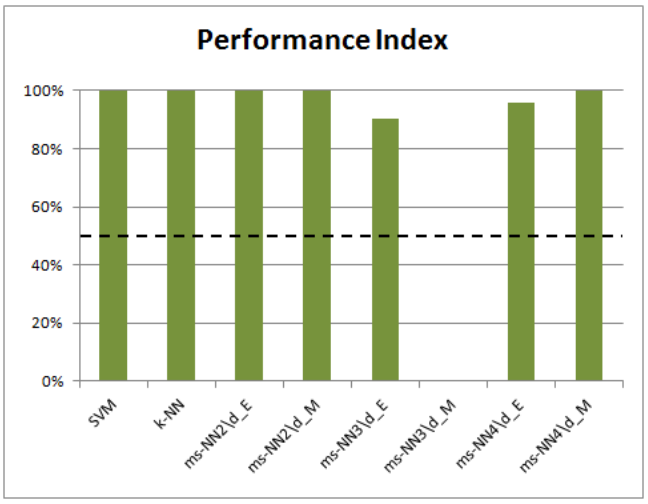

Fig. 3. Performance index: comparison of each classifier with ms-NN3\d_M.

5. CONCLUSIONS AND CONSIDERATIONS

In an exhaustive experiment, ms-NN has shown great potential in classifying land cover using real remotely sensed images. The best results of the configurations tested in ms-NN obtained higher OA than the well-known methods SVM and k-NN.

The main drawback of ms-NN is its asymptotic complexity, the same of k-NN. This work is the first step towards the use of learning algorithms, aiming at decreasing the number of parameterizations to be considered without a decrease in accuracy. The exhaustive results will be the basis of comparison for the heuristic ones. Future work also includes testing different distribution and size of test samples, as well as allowing different distance measures for each space in ms-NN parameterizations.

In a parallel study, ms-NN obtained very good results in the classification of Schistosomiasis prevalence in the state of Minas Gerais in Brazil (70% for ms-NN against 58% for SVM and 55% for k-NN). The best results were obtained with the use of the geographic spaces in addition of the attribute ones, what is something new in this approach of nearest neighbor based classifier. These results indicate a wide range of possibilities for using ms-NN in other applications besides land cover classification.

6. REFERENCES

[1] F. T. Martins-Bedê, *Extensão de classificadores k-NN para múltiplos espaços, distâncias e funções de predominância*, Ph.D. thesis, Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2014-02-26 2014.

[2] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, Academic Press, San Diego, 3 edition, 2006.

[3] S. Sandri, F.T. Martins-Bedê, and L. Dutra, "Using a fuzzy based pseudometric in classificatiom," *submitted*.

[4] S. Sandri and F.T. Martins-Bedê, "A method for deriving order compatible fuzzy relations from convex fuzzy partitions," *Fuzzy Sets and Systems*, in press, 2014.

[5] S.D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intelligent Data Analysis*, vol. 3, no. 3, pp. 181–209, 1999.

[6] IBAMA, "Floresta nacional do tapajós: Plano de manejo. volume i informações gerais," Tech. Rep. 76p, Instituto Brasileiro do Meio Ambiente e dos recursos naturais renováveis, Brasil: IBAMA, 2004.

[7] M. I. S. Escada, S. Amaral, C. D. Rennó, and T. F Pinheiro, "Levantamento do uso e cobertura da terra e da rede de infraestrutura no distrito florestal da BR-163," Tech. Rep., São José dos Campos, 2009, 2009. 52 p. (INPE-15739-RPQ/824).

[8] Y.E. Shimabukuro and A. Smith, "The least-squares mixing models to generate fraction images derived from remote sensing multispectral data," *IEEE Transactions on Geosciense and Remote Sensing*, vol. 29, no. 1, pp. 16–20, 1991.