Primljeno / Received: 15.09.2017.
Prihvaćeno / Accepted: 13.11.2017.

UDK 528.8:83:85
Stručni rad / Professional article

EVALUATING TRAINING DATA FOR CROP TYPE CLASSIFICATION USING SUPPORT VECTOR MACHINES AND RANDOM FORESTS

PROCJENA KLASIFIKACIJE TESTNIH PODATAKA ZA POTREBE ODREĐIVANJA VRSTE KULTURE KORIŠTENJEM METODA MAŠINA VEKTORA PODRŠKE I SLUČAJNIH ŠUMA

Mustafa Ustuner, Fusun Balik Sanli

ABSTRACT

This study evaluated the effectiveness of three different training datasets for crop type classification using both support vector machines (SVMs) and random forests (RFs). In supervised classification, one of the main facing challanges is to define the training set for the full representation of land use/cover classes. The adaptation of traning data, with the implemented classifier and its characteristics (purity, size and distribution of sample pixels), are of key importance in this context. The experimental results were compared in terms of the classification accuracy with 10-fold cross validation. Results suggest that higher classification accuracies were obtained by less number of training samples. Furthermore, it is highlighted that both methods (SVMs and RFs) are proven to be the effective and powerful classifiers for crop type classification.

Keywords: training data, crop type classification, support vector machines, random forests, machine learning.

SAŽETAK

Ova studija je dala procjenu učinkovitost tri različita skupa podataka za potrebe određivanja vrste kulture pomoću support vector machines (SVM) i random forests (RF). Jedan od glavnih izazova s kojima se susrećemo kod nadgledane klasifikacije je definisanje podataka sa potpunom zastupljenošću klasa korištenja / pokrova zemljišta. Prilagođavanje podataka za implementaciju klasifikatora i njegovih karakteristika (čistoća, veličina i raspodjela pikselskih uzoraka) od ključne su važnosti u ovom kontekstu. Eksperimentalni rezultati upoređeni su s obzirom na tačnost klasifikacije s 10-strukom unakrsnom validacijom. Rezultati upućuju na to da su veće tačnosti klasifikacije dobivene sa manjim brojem testnih uzoraka po klasi. Nadalje, naglašeno je da su obje metode (SVM i RF) dokazane kao učinkoviti i moćni klasifikatori za klasifikaciju različitih vrsta kulture.

Ključne riječi: testni podaci, klasifikacija vrste kulture, support vector machines, random forests, mašinsko (strojno) učenje.

1 INTRODUCTION

Thanks to the recent development in spatial and spectral resolution of the earth observation satellite sensors, remote sensing has become more essential and powerful in several application such as environmental monitoring, precision agriculture, urban planning and many more. Image classification can still be considered as one of the most effective methods to derive information about the Earth's surface for the wide range of applications and hence has been of great interest by many researchers in remote sensing. Many advanced classifiers including deep learning algorithms, ensemble of classifiers (multiple classifiers), active learning methods and svm-based classifiers have been developed and used in remote sensing and pattern recognition. However classifying the images is still a challenging issue and complicated process since it can be influenced by many factors such as allocation of ground truth data, resolution of input imagery, feature selection and landscape heterogeneity (Foody and Mathur, 2004; Lu and Weng, 2007; Kavzoglu, 2009; Mather and Koch, 2011). The purity, size, distribution of traning data largely affect the accuracy of the supervised classification and hence the design and selection of the traning samples should be specific for the classifier to be used. The optimal size for the training data is not clear due to the uncertainty and complexity of the training stage (Foody and Mathur, 2006). Few number of training samples sometimes may not be enough to adopt the algorithm, while large number of traning samples can cause the overfitting problem (Kavzoglu and Mather, 2003; Kavzoglu 2009). The size of training data (number of sample pixels) needed for supervised classification can vary up to the characteristics of the algorithm and is of crucial importance in classification stage (Foody, Mathur, Sanchez-Hernandez and Boyd, 2006). Statistical classifiers (e.g. maximum likelihood and minimum distance classifiers), which are also called parametric classifiers, use the statistical parameters obtained from the training samples, whereas non-parametric classifiers (e.g. support vector machines and random forests classifiers) directly use the training data and make no prior assumption for the distribution of sample data. The sensitivity of classifiers for the same training set in terms of learning stage of algorithm can differ from each other. Here the significant points are the type and characteristics of training data employed for the classification (Foody and Mathur, 2006; Mather and Koch, 2011). Many studies have evaluated the impact of training set size on classification accuracy (Huang, Davis and Townshend, 2002; Pal and Mather, 2003; Foody and Mathur, 2004; Kavzoglu 2009). Huang et al. (2002) investigated the impact of training data size with four different classification methods including SVMs for land cover classification. Pal and Mather (2003) evaluated the effects of training data set size to test the effectiveness of decision tree methods for land cover classification. Foody and Mathur (2004) assessed the intelligent training sample collection for SVMs classification of agricultural crops. Kavzoglu (2009) examined the effect of a training set on artificial neural network classification. All these aforementioned studies confirm that classification accuracy is largely influenced by training data.

In this paper, two popular machine learning algorithms, support vector machines and random forests, have been preferred with the aim of evaluating the three different training dataset in terms of classification accuracy for crop type classification. The reasons for preferring these algorithms are as following: (i) their superior performance and popularity in remote sensing image classification and (ii) their non-parametric structure in terms of using training data. These

techniques make no prior assumption about the distribution of training data, unlike parametric methods, and may lead the high classification accuracy with limited number of training samples. The main purpose of this paper is to evaluate the training datasets for crop type classification with SVMs and RFs.

The rest of this paper is structured as follows. In section 2, the brief details of the classification algorithms are provided. The study area, dataset and methodology are introduced in Section 3. Section 4 presents classification accuracies and compares the results in terms of training set size. At last, Section 5 summarize results and draw the conclusion with some important remarks.

2 CLASSIFICATION ALGORITHMS

Support vector machines, a kernel-based classifier, and random forests, a tree-based ensemble learning algorithm, have gained great popularity in pattern recognition and remote sensing in last decade due to their good generalization performance on high dimensional data (Papa, Amorim, Falcão and Tavares, 2016). In this chapter, only a brief summary about the classification algorithms is provided and the reader who needs further details should refer to Gislason, Benediktsson and Sveinsson (2006) for RFs and Melgani and Bruzzone (2004) for SVMs.

Random forests which are ensemble of classification trees (multiple classifier) where each tree contributes one single vote for the appointment of most popular class. Random forests create multiple trees and the output classification is generated by combining the individual results using the majority voting strategy. Because of the low computation costs and high classification accuracy, RFs are one of the most popular classification and regression methods in pattern recognition (Waske, B., Benediktsson, J. and Sveinsson, 2012; Ghamisi, J. Plaza, Chen, Li and A. J. Plaza, 2017). Due to the ability of handling high dimensional data with small number of training samples, SVMs, one of the powerful machine learning algorithms, have been extensively used in remote sensing for classification and regression problems. SVMs were originally developed for linear classification by defining the optimal hyperplane with maximum margin width. For the non-linear cases, which is common in multi-class classification problem, SVMs use kernel methods that transform the training data into higher dimensional feature space. As a kernel type, radial basis function (RBF) kernel was employed for SVMs classification and optimum parameters (Regularization parameter and kernel width) for RBF kernel were determined by using grid search method (Melgani and Bruzzone, 2004).

3 DATASETS AND METHODS

The study area which is located in Aydin, a province in southwestern Turkey, is dominantly covered by cultivated areas (Figure 1). Major income in the region is agriculture with the great advantage of its climatic condition and fertile lands. RapidEye imagery (RapidEye Ortho product-Level 3A-) acquired on 23 August 2012 was preferred in this study since it incorporates the red-edge band sensitive to chlorophyll content. The imagery provides high resolution five spectral bands from 400 nm to 850 nm. Spatial resolution of the imagery is five meter in Ortho product but the original RapidEye data is 6.5 meter at nadir (Blackbridge,2013). Study area is covered by nine land use classes (mostly crop types) which are corn (first crop, second crop), cotton (well developed, moderate developed, weak developed), soil (wet, moist, dry) and water surface. Ground truth data collected by soil scientists at the acquisition date of imagery was used for both training and testing process. In order to evaluate the effectiveness of training datasets for crop type classification, three different traning sets were used in classification stage. First training set (Set1) has total number of 45467 sample pixels, approximately 6% of the entire study area. The second and third ones have the total number of 13718 and 2794 samples, approximately 2% and 0.4% of the entire study area, respectively (Table 1). Ground truth maps for these three different traning sets are shown in Figure 2. In all classifications, 10-fold cross validation were implemented for training and testing datasets in this study.

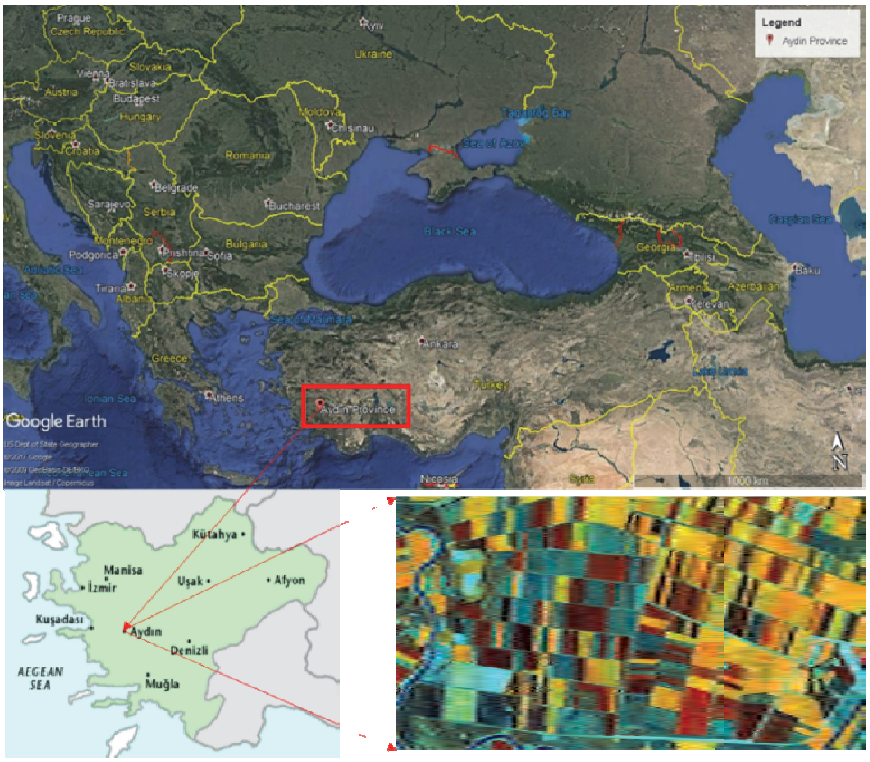

Figure 1. Study Area

Table 1.
Training set size (Number of Samples)

Class Information	Number of Samples		
Classes	Set1	Set2	Set3
First crop corn	5404	1975	236
Second crop corn	4780	1414	237
Well-developed cotton	4983	1749	302
Moderate-developed cotton	7163	1597	312
Weak-developed cotton	2540	1552	365
Wet soil	7947	1941	329
Moist soil	6086	1263	289
Dry soil	3714	1501	368
Water Body	2850	789	356
Total Number of Pixel	**45467**	**13781**	**2794**

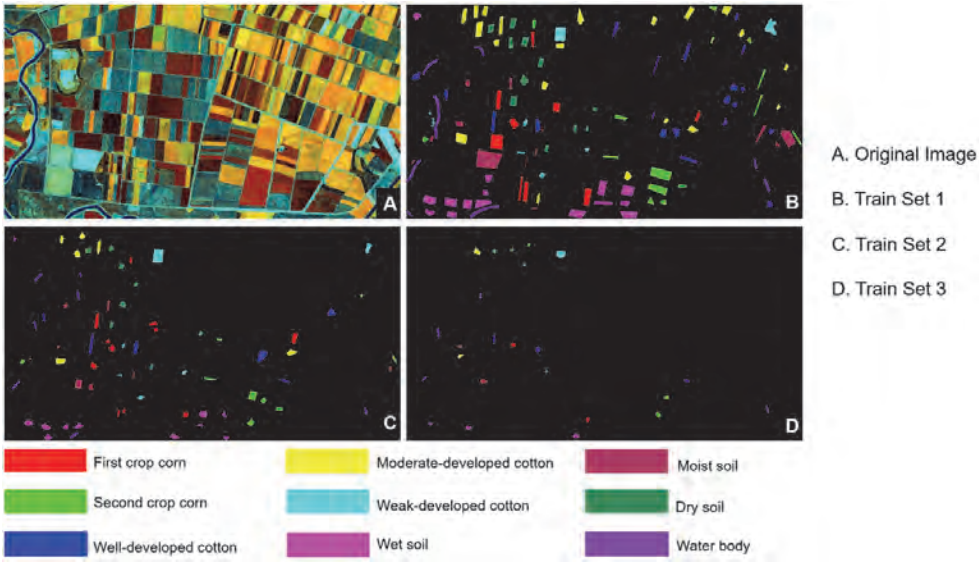

Figure 2. Original Image and Ground Truth Maps for training sets

4 RESULTS

This section presents the classification performance of the SVMs and RFs classification and demonstrates how the results are influenced based on the selection of different training sets. The experimental results were compared in terms of the classification accuracy with 10-fold cross validation. The overall accuracy and kappa coefficient were used for comparing the classification results and are provided in Table 1.

Table 1.
Classification accuracies for all training datasets

	Set1		Set2		Set3	
Type	**Overall Acc.(%)**	**Kappa**	**Overall Acc.(%)**	**Kappa**	**Overall Acc.(%)**	**Kappa**
RF	96,39	0,9588	98,43	0,9823	99,82	0,9980
SVM	95,67	0,9505	98,22	0,9799	99,93	0,9992

The results demonstrate that both methods (SVMs and RFs) are effective and powerful for crop type classification and achieve higher classification accuracies with less number of training samples in the training set. In our experience, SVMs are once again proven to achieve the higher classification accuracy with a less number of training samples, which is also proven in many studies in the literature.

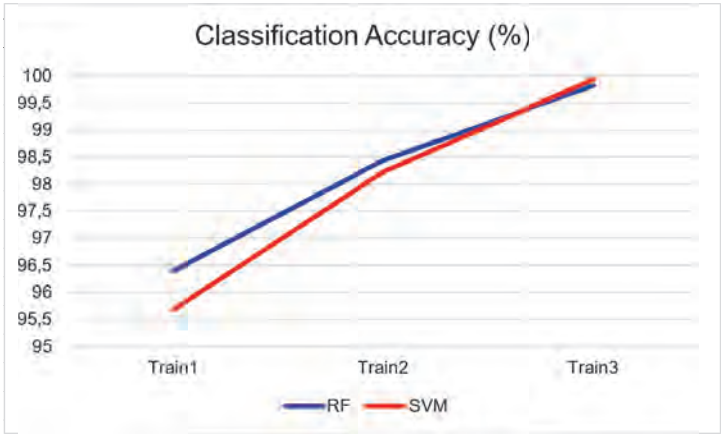

Figure 3. Changes in classification accuracies

The changes in classification accuracies (overall accuracies-%-) for both method based upon the training data can be seen in Figure 3. It can be revealed from the Figure 3 that, SVMs outperformed RFs when less number of traning samples are employed in classification. Furthermore, RFs are more successful for handling classification problem when high number of training samples are employed.

5 CONCLUSION

Training data are significant elements in supervised classification learning as they may influence the expected results by either negative or positive outcome based upon the characteristics of training data and their adaptation with the learning algorithm. Since the uncertainty of the learning stage, the optimal selection of training data is critical for supervised image classification.

In this study, effectiveness of three different training datasets for crop type classification using support vector machines (SVMs) and random forests (RFs) was comparatively evaluated and the results were presented. Experimental results suggest that both methods (SVMs and RFs) are effective and powerful for crop type classification and achieved higher classification accuracies with less number of training samples in the training set. RFs outperformed SVMs when training set 1 was employed, while SVMs obtained slightly higher classification accuracy than RF in using of training set 3. The possible reason of obtaining relatively lower classification results with training set 1 could be the overfitting problem since the SVMs only needs the support vectors to find the optimal hyperplanes separating the classes. Furthermore, the results emphasized that the training data has to be carefully in design and selection for the fully representation of the land use/cover classes as well as to meet the requirement of any particular classifier to be used.

ACKNOWLEDGMENT

The authors would like to thank Prof. Dr. Yusuf Kurucu and Assoc.Prof. Dr. M. Tolga Esetlili, Ege University, Turkey for providing free access to satellite imagery and ground truth data set.

LITERATURE AND REFERENCES

Blackbridge (2013). *Satellite Imagery Product Specifications.* [Brochure]. Retrieved from https://resa.blackbridge.com/files/2014-06/RE_Product_Specifications_ENG.pdf

Foody, G. M., Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment, 93*(1), 107-117. doi: https://doi.org/10.1016/j.rse.2004.06.017

Foody, G. M., Mathur, A. (2006). The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment, 103*(2), 179-189.
doi: http://dx.doi.org/10.1016/j.rse.2006.04.001

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment, 104*(1), 1-14. doi: https://doi.org/10.1016/j.rse.2006.03.004

Ghamisi, P., Plaza, J., Chen, Y., Li, J., Plaza, A. J. (2017). Advanced Spectral Classifiers for Hyperspectral Images: A review. *IEEE Geoscience and Remote Sensing Magazine, 5*(1), 8-32. doi: https://doi.org/10.1109/mgrs.2016.2616418

Gislason, P. O., Benediktsson, J. A., Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi: https://doi.org/10.1016/j.patrec.2005.08.011

Huang, C., Davis, L. S., Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*(4), 725-749. doi: 10.1080/01431160110040323

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software, 24*(7), 850-858.
doi: http://dx.doi.org/10.1016/j.envsoft.2008.11.012

Kavzoglu, T., Mather, P. M. (2003). The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing, 24*(23), 4907-4938. doi: 10.1080/0143116031000114851

Lu, D., Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing, 28*(5), 823-870. doi: 10.1080/01431160600746456

Mather, P. M., Koch, M. (2011). Classification. In *Computer Processing of Remotely-Sensed Images* (pp. 229-284). Chichester: John Wiley & Sons, Ltd.

Melgani, F., Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing, 42*(8), 1778-1790. doi: 10.1109/tgrs.2004.831865

Pal, M., Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment, 86*(4), 554-565.
doi: https://doi.org/10.1016/S0034-4257(03)00132-9

Papa, J. P., Amorim, W. P., Falcão, A. X., Tavares, J. M. R. S. (2016). Recent Advances On Optimum-Path Forest For Data Classification: Supervised, Semi-Supervised, And Unsupervised Learning. In C. H. Chen (ed.), *Handbook of Pattern Recognition and Computer Vision, 5th ed.* (pp. 109-123). Singapore: World Scientific.

Waske, B., Benediktsson, J., Sveinsson, J. (2012). Random Forest Classification of Remote Sensing Data. In C. H. Chen (ed.), *Signal and Image Processing for Remote Sensing, Second Edition* (pp. 365-374). Boca Raton: CRC Press.

Authors:

Mustafa Ustuner, Research Assistant
Department of Geomatic Enginering, Yildiz Technical University
34220, Esenler, Istanbul
Turkey
mustuner@yildiz.edu.tr

Assoc. Prof. Dr. Fusun Balik Sanli
Department of Geomatic Enginering, Yildiz Technical University
34220, Esenler, Istanbul
Turkey
fbalik@yildiz.edu.tr