ARTICLE IN PRESS

Remote Sensing of Environment xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites

Shahriar S. Heydari, Giorgos Mountrakis*

Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States

ARTICLE INFO

Keywords:
Land-cover mapping
Classifier accuracy assessment
Image classification
Image heterogeneity

ABSTRACT

A major issue in land cover mapping is classifier selection. Here we investigated classifier performance under different sample sizes, reference class distribution, and scene complexities. Twenty six 10 km × 10 km blocks with complete reference information across the continental US are used. Per-pixel classification took place using six spectral bands from Landsat imagery. The tested classifiers included Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Bootstrap-aggregation ensemble of decision trees (BagTE), artificial neural network up to 2 hidden layers, and deep neural network (DNN) up to 3 hidden layers. For the entire block, our accuracy assessment indicated that all classifiers, with the exception of NB (a Maximum Likelihood variant), performed similarly. However, when we concentrated on edge pixels (pixels at the border of adjacent land cover classes), it was clear that the SVM and KNN offer considerable accuracy advantages, especially for larger reference datasets. Because of their relatively low execution times SVM and KNN would be recommended for classifications using Landsat's spectral inputs and Anderson's 11-level classification scheme. However, both SVM and KNN demonstrated substantial accuracy degradation during the parameter grid search. For this reason, an exhaustive parameter optimization process is suggested. While the ANN and DNN neural network variants did not perform as well, their performance may have been restricted by the lack of rich contextual information in our simple six band per-pixel input space. The effect of class distribution in the training dataset was also evident on the calculated accuracy metric. Gradual accuracy degradation as edge pixel presence increased was also observed. Future work could focus on data-rich classification problems such as change detection using Landsat stacks or expand in high spectral or spatial resolution sensors.

1. Introduction

Classification of remotely sensed data is essential in generating thematic maps. Thematic maps have many applications in environmental management, agricultural planning, health studies, climate and biodiversity monitoring, and land change detection (Khatami et al. 2016). A wide range of regional and global datasets for classification are currently available, facilitating studies at unprecedented scales (Grekousis et al. 2015). The classification process, in general, is composed of different tasks, from the selection of data source and sampling design, to classification method selection and classifier performance evaluation (Lu and Weng 2007). Although all of these tasks are important and their successful implementation is dependent on each other, a major task is the selection of a suitable classification method.

One type of classification method may be more suitable for a specific target objective, problem condition, or imaging details over another method (see Table 1 in Lu and Weng 2007). The classifiers performance assessment is also highly dependent on data quality, data values distribution, and sampling design (Jin et al. 2014; Li et al. 2014a); and it can also be evaluated under various criteria like accuracy, reproducibility and/or robustness (Cihlar et al. 1998). Even for the most widely used assessment criteria for classification accuracy, there are important concerns that limit the ability to properly assess the accuracy of resulting map (see Foody 2002, for a review). This line of research has been followed by more recent papers discussing the problems arising from increasing accuracy degradation over time in temporal land cover analysis and change detection (Giles M. Foody 2010), or stressing the importance of sample size or statistical hypothesis testing when comparing different classifiers or scenarios performance (Giles M. Foody 2009).

* Corresponding author.
E-mail addresses: sshahhey@syr.edu (S.S. Heydari), gmountrakis@esf.edu (G. Mountrakis).

http://dx.doi.org/10.1016/j.rse.2017.09.035
Received 20 February 2017; Received in revised form 18 September 2017; Accepted 28 September 2017

Please cite this article as: Heydari, S.S., Remote Sensing of Environment (2017), http://dx.doi.org/10.1016/j.rse.2017.09.035

Table 1
Classifiers parameters

Classifier	Parameter	Parameter values/range
Naïve Bayesian (NB)	Probability distribution type	Normal, Kernel
	Smoothing function	Normal, Box, Triangle, Epanechnikov
K-Nearest Neighbor (KNN)	Distance metric	Chebychev, Euclidean, Mahalanobis, Minkowski
	Distance weight	Inverse, squared inverse
	Number of neighbors	1 to 40 (step of 2)
Support Vector Machine (SVM)	Kernel function	Fixed at Gaussian
	Box constraint (C)	0.01, 0.1, 0.5, 1, 2, 5, 10, 25, 50, 100, 300
	Kernel scale (gamma)	0.1, 0.5, 1, 2, 5, 10, 25, 50
Tree ensemble (BagTE)	Ensemble method	Bagging
	Number of trees	50, 100, 200, 500
	Maximum number of tree splits	10, 25, 50, 100, 200
	Minimum tree leaf size	1, 3, 5, 10, 25
	Number of simulation iterations	10
Artificial Neural Network (1 or 2 hidden layers, followed by a softmax classifier)	Training algorithm	Resilient backpropagation (trainrp)
	# of nodes in 1st hidden layer	5 to 15 (step of 1)
	# of nodes in 2nd hidden layer	0 to 8 (step of 1)
	Number of simulation iterations	100
	Training parameters (specific to chosen training algorithm):	Changed randomly in each iteration within given range:
	- Learning rate	- 0.01–1
	- Delta0	- 0.01–0.5
	- Delta_inc	- 1–5
	- Delta_dec	- 0.1–1
Deep Neural Network, autoencoder-based (1, 2, or 3 hidden layers, followed by a softmax classifier)	Training algorithm	Standard backpropagation
	# of nodes in 1st hidden layer	5 to 30 (step of 2)
	# of nodes in 2nd hidden layer	0 to 20 (step of 2)
	# of nodes in 3rd hidden layer	0 to 10 (step of 2)
	Number of simulation iterations	100
	Training parameters (specific to chosen training algorithm):	
	- Lambda	
	- Rho	- 1E-8–1E-3
	- Beta	- 0.05–0.7
		- 1–9

Therefore, it is difficult to generate a general statement to advise on classifiers ranking. One should always declare the specific conditions that the classifier performance assessment is based on. There are good review papers that introduce the classifiers in general and discuss their application conditions, strengths and weaknesses (Lu and Weng 2007; C. Li et al., 2014; X. Li et al., 2014), but they are mostly qualitative without specific quantitative results for example, best attainable classifiers accuracy. Other papers discuss classifiers for certain problem types. For example, see (Weng 2012) for a discussion on classifiers for mapping of impervious surfaces, (Mallinis and Koutsias 2012) for a comparison of ten classifiers for burned area mapping, (J. He et al. 2015), for comparing four main classifiers in generation of arctic geological maps, or (Pelletier et al. 2016) for assessing the robustness of random forest (RF) classifier for a specific area. Still, other researchers seek to review the application of a specific classifier in more detail. For example, see (Mountrakis et al. 2011) for a review of SVM classifiers; (Pal and Mather 2003), for an assessment of decision tree methods for land use classification; or (Belgiu and Drăguţ 2016), for an overview of random forest classifier. Additional processing is another focus of research which includes making ensemble of classifiers (X. Li et al. 2014), controlling of misclassification by post-processing (Marcos Martinez and Baerenklau 2015), or using ancillary data to aid in classification by field visits (Meddens et al. 2016) or other sources and sensors (Zhu et al. 2016). Based on numerous case studies, one can perform a meta-analysis of previously researched cases and assess the comparative results of case studies at a higher level. This meta-analysis has been done for a single type of classifier such as KNN (Chirici et al. 2016), or more general including pairwise comparisons among many classifiers (Khatami et al. 2016).

While fragmented comparisons between traditional classifiers can be found in existing literature, they are limited in terms of: i) number of case studies incorporated, ii) the search space of the classifier parameters (often resorting to default values), and iii) absence of a promising new classification family based on deep neural networks (DNN). To the best of our knowledge, there are just a few studies that investigate per-pixel classifier accuracy performance over multiple case studies or over a large area. For example, (Ballantine et al. 2005) performed mapping for continental North Africa using MODIS data but comparisons were restricted to a few classifiers. In (Gong et al. 2013) a global sampling and classification was implemented using four different classifiers, but they used a fixed set of parameters for each classifier. Similarly, (Lawrence and Moran 2015) tested classification accuracy for multiple classifiers for 30 data sets but they used a fixed set of classifier parameters that did not allow classifiers to reach their best potential. (Pelletier et al. 2016) performed a grid search on classifier parameters over two large areas in France, focusing on SVM and Random Forest classifiers. Finally, W. Li et al. (2016) employed numerous popular classifiers plus the new autoencoder-based DNN implementations over one composite set sampled through the entire Africa, but they only reported a fixed parameter set (except for DNN).

Our research goals fill this gap by overcoming the three aforementioned limitations. Along these lines, we: i) compared classifiers' best achievable accuracy, ii) identified the accuracy costs associated with the reduction of the parameter grid and training dataset size and iii) investigated how landscape heterogeneity influences classifier performance. We tested six different classifiers in our research: Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Bootstrap-aggregation ensemble of decision trees (BagTE), artificial neural network (ANN) up to 2 hidden layers, and autoencoder-based deep neural network (DNN) up to 3 hidden layers. We used a dataset of 26 Landsat images for classifiers comparison, and ran each classifier with a grid of parameter settings to evaluate its performance.