The effects of training set size for performance of support vector machines and decision trees

Taskin Kavzoglu[1] *and Ismail Colkesen*[2]

Gebze Institute of Technology, Department of Geodetic and Photogrammetric Engineering,
Cayirova Campus, 41400, Gebze-KOCAELI, TURKEY.
[1]kavzoglu@gyte.edu.tr, [2] icolkesen@gyte.edu.tr

Abstract

Thematic maps representing the characteristics of the Earth's surface have been widely used as a primary input in many land related studies. Classification of remotely sensed images is an effective way to produce these maps. Selecting proper number of samples and classification method are essential issues to produce accurate thematic maps. In the literature, many classification algorithms have been developed and their performances have been analyzed for different data sets. In this study, support vector machines (SVMs) and decision trees (DTs), relatively new and widely used methods, were applied to produce land use/land cover thematic map of the study area, which covers the center of Trabzon province of Turkey. Training data sets at various sizes were used to investigate the effect of the training set size on the classification accuracy. Variations in the classification performances were analyzed using overall classification accuracy and Kappa coefficient derived from the error matrix. Furthermore, McNemar's and z tests were employed to determine the statistical significance of differences in classifier performances depending on the training sample size. Results showed that classification performances of SVMs and DTs improved till a certain level

Keywords: Support vector machines, decision trees, accuracy comparison, McNemar's test.

1. Introduction

Remote sensing technologies provide valuable information for various land related studies. One of the effective and commonly used methods for extracting such information from the remotely sensing imagery is classification process. The primary output of the classification process is a thematic map representing different features of the Earth's surface. These maps are used as a base map in many global and regional studies (e.g. environmental modeling and land-use planning). Therefore, having an accurate thematic map has crucial importance in order to achieve the desired goals of the studies. Classification of remotely sensed imagery consists of complex and multi-stage steps including determining an appropriate classification method, training sample selection and accuracy assessment. Several factors (e.g. classification strategy, parameter selection and data characteristics) have significant impacts on the classification results, which should be considered carefully by the analyst in the each step (Kavzoglu, 2009). A suitable classification procedure and a sufficient number of training samples are essential prerequisites for accurate classification result (Lu and Weng, 2007).

There are two fundamental procedures that can be used for classification of an image, which are supervised and unsupervised classification techniques. Supervised classification, which is the process of using training data for assigning class labels to unknown pixels, has been widely used in remote sensing arena. In the literature, it is generally underlined that there is a strong relationship between classification accuracy and training data sets used in the learning stage of supervised classification method (Zhuang *et al.*, 1994; Foody, 1999; Pal and Foody, 2010). Foody and Mathur (2006) indicated that the accuracy of a supervised image classification is a function of the training data used.

In this study, it was focused on the size of the training set, and its impacts on classification performances of two well-known and relatively new supervised classification methods, namely, support vector machines (SVMs) and decision trees (DTs). Classification results were compared using overall accuracies and Kappa coefficient values. In addition of these comparisons, McNemar's and *z* tests were used to determine whether there was a statistically significant difference between classification results.

2. Study area and Data Sources

The study area chosen for this research covers approximately 2.5 km^2 area located in the south part of Trabzon province, Turkey. A pan-sharpened Quickbird satellite image of 1735-pixel by 1442-pixel covering the study area was used to determine land cover and land use types. The image, which was acquired in May 2005, was registered to the UTM (zone 37, ED50 datum) projection system by applying a first-order polynomial transformation. Several maps and aerial photographs were used to create ground reference maps. Additionally, field surveys were applied using a handheld GPS to collect ground reference information. After the detailed analysis of ground reference data, it was decided that mainly seven land use and land cover types covers the study area, which are water, bare soil, urban, road, pasture, forest and shadow.

3. Methodology

In the literature, numerous classification algorithms have been introduced and their classification performances were analysed with different satellite images (Jain *et al.*, 2000; Lu and Weng, 2007; Jia *et al.*, 2011). In this study, two outstanding methods, namely support vector machines and decision trees, were applied for the classification of high resolution Quickbird imagery to produce the land use and land cover thematic map of the chosen study area.

Support vector machines (SVMs) are one of the supervised learning algorithms based on statistical learning theory (Vapnik, 1995). The main goal of the SVMs for any classification problems is to construct a hyperplane separating two classes optimally. The method employs optimization algorithms to locate the optimal boundaries between classes (Huang *et al*, 2002). SVMs algorithms have been successfully used in may remote sensing studies (Kavzoglu and Colkesen, 2011a; Mountrakis *et al.*, 2011). Kernel functions are used for mapping nonlinearly separable data (e.g. remotely sensed images) into a higher dimensional space to define the optimal hyperplane. Kernel functions commonly used in SVMs can be generally aggregated into four groups; namely, linear, polynomial, radial basis function and sigmoid kernels. However, radial basis functions have been widely preferred for the classification of the classification of remotely sensed images. It was also used in this study to classification of high resolution Quickbird image.

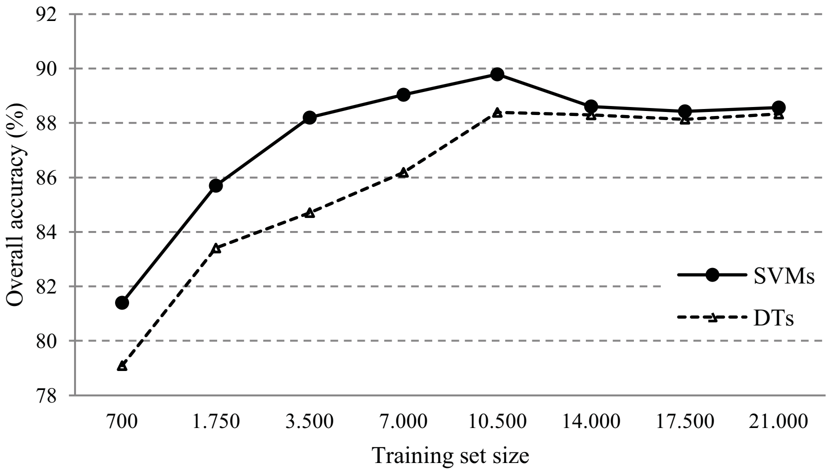
Decision trees (DTs) are defined as a classification technique that recursively partitions a data set into smaller subdivisions on the basis of a set of tests defined at each node in the tree (Friedl and Brodley, 1997). DTs based on the "divide and conquer" strategy are supervised learning algorithms that have been recently used in remote sensing applications (e.g. Friedl *et al.*, 2010, Boulila, 2011). DTs use a multi-stage approach for the class label assignment. The labelling process is considered to be a chain of simple decisions based on the results of sequential tests rather than a single, complex decision (Pal and Mather, 2003). In this study, C4.5 algorithm was employed for the construction of decision trees. C4.5 builds decision trees from a set of training data using the concept of information entropy.

Variations in the classification performances of the each method were analyzed using overall classification accuracy and Kappa coefficient derived from the error matrix. In addition to these accuracy measures, significance of the differences in classification performances was analyzed through McNemar's tests based on normal and chi-squared distributions. McNemar's test is a non-parametric test based on confusion matrices at 2 by 2 in dimension. For related samples, McNemar's test could be useful to estimate the statistical significance of the difference between two proportions (Foody, 2004; Kavzoglu and Colkesen, 2011b). Since Kappa coefficient does not satisfy the assumption of independence, Kappa based z-test is not suitable for cases in which related samples are utilized. If the statistic values are greater than critical table value (i.e. critical value of chi-squared test is 3.84 and critical value of the normal test statistic is 1.96 at 95% confidence interval), the difference in performances in terms of classification accuracy is said to be statistically significant.

4. Results and Discussions

In this study, performance of support vector machines and decision trees was investigated with the various sizes of training data on a land use and land cover identification problem. In order to achieve desired objectives of this study, different size of training data sets were created by random pixel selection strategy using an in-house program. As a result, eight training data sets of various sizes (i.e. 700, 1.750, 3.500, 7.000, 10.500, 14.000, 17.500 and 21.000) were used to investigate the effect of the training set size on the classification accuracy. In addition, a test data set including 14.000 pixels was produced to evaluate the performances of classification methods for any given training data set. At this point, it should be noted that the same training and test data sets were employed for all classification experiments.

SVMs with radial basis function kernel were used for the classification of the satellite image. For the selection of optimum parameters of SVMs (i.e. regularization parameter and kernel width) cross validation approaches were performed. It was observed that the regularization parameter (*C*) takes values between 500 and 1.500 for the variation in training data. Moreover, the optimal values of the kernel width were determined for different sample sizes ranging from 0.1 to 1. The changes in overall accuracies and Kappa coefficients corresponding to increasing training set size were given in Figure 1. It can be seen from the figure that classification performances of SVMs improved till a certain level considering training set size. Highest overall accuracy of 89.79% was achieved with the training data set containing totally 10,500 pixels (1,500 pixels per class). It should be noted that after this critical point, the classification accuracy showed downward trend, that is, it was negatively affected with the increasing number of training pixels. Moreover, it can be said that in the case of limited number of training pixels, SVMs produced higher classification accuracies than DTs.



SVMs	81.40 (0.78)	85.70 (0.88)	88.20 (0.86)	89.04 (0.87)	89.79 (0.87)	88.61 (0.87)	88.43 (0.86)	88.57 (0.86)
DTs	79.09 (0.75)	83.41 (0.81)	84.71 (0.82)	86.19 (0.84)	88.39 (0.86)	88.30 (0.86)	88.14 (0.86)	88.33 (0.86)

Figure 1: Classification accuracies for SVMs and DTs methods related to number of training samples. Note that values within parentheses indicate Kappa values.

With the use of eight training data sets, DTs models were constructed and their classification performances on the test data set were shown in Figure 1. Similar to SVMs performance, the highest overall accuracy of 88.39% was calculated with the use of 10,500 training pixels. After that point, DTs models produced lower overall accuracies same as SVMs classification. It should be noted that performances of both methods for more than 10,500 training samples were similar to each other. At this point, McNemar's test was employed to determine the statistical significance of differences in classifier performances (Table 1). When these results were analyzed, it was found that differences in classification accuracies of SVMs were statistically significant till 10,500 samples. After this critical number differences in classification accuracies were statistically insignificant. On the other hand, test results also showed that the classification performances of DTs were unstable with increasing training set size till the critical pixel size of 10,500.

McNemar's tests (i.e. normal test statistic (z) and chi-squared distribution (χ^2)) were also employed to determine differences between SVMs and DTs classification results to test whether classification performances statistically significant or insignificant with respect to increasing training data size (Table 1). The test statistics greater than the critical values are shown as "Yes" in Table 1. According to z-statistic, accuracy differences of SVMs classifications were significant until the critical training set size of 7,000. Considering chi-square test results, 10,500 training samples were determines as critical in SVMs classification. The difference in the performances by the DTs classification was statistically significant up to a critical level of 10,500 samples, in parallel to z-statistic results. On the other hand, two training set sizes (i.e. 3,500 and 10,500 pixels) were determined as critical for the DTs classification in terms of chi-square test statistic. When classification performances of SVMs and DTs were analyzed, it was found that after the critical training size of 10,500, both classifiers showed similar classification performances and differences in classification performances were statistically insignificant.

Table 1: Statistical test results for the classification methods. Note that all tests were two-tailed, critical values for z-statistic and McNemar's tests for 95% confidence interval are 1.96 and 3.84, respectively.

	Result-1	Result-2	$z = \frac{n_{ij} - n_{ji}}{\sqrt{n_{ij} + n_{ji}}}$	$\chi^2 = \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$	Significant ?	
					z	χ^2
SVM1 vs. SVM2	81.40	85.70	6.8992	11.8335	Yes	Yes
SVM2 vs. SVM3	85.70	88.20	4.4053	8.5088	Yes	Yes
SVM3 vs. SVM4	88.20	89.04	1.5726	4.5385	No	Yes
SVM4 vs. SVM5	89.04	89.79	0.4851	1.6164	No	No
SVM5 vs. SVM6	89.79	88.61	0.3207	1.3093	No	No
SVM6 vs. SVM7	88.61	88.43	0.3450	1.1272	No	No
SVM7 vs. SVM8	88.43	88.57	0.2652	1.0660	No	No
DT1 vs. DT2	79.09	83.41	6.5861	8.8926	Yes	Yes
DT2 vs. DT3	83.41	84.71	2.1072	2.6753	Yes	No
DT3 vs. DT4	84.71	86.19	2.4698	3.2785	Yes	No
DT4 vs. DT5	86.19	88.39	3.9101	5.2514	Yes	Yes
DT5 vs. DT6	88.39	88.30	0.1581	0.2236	No	No
DT6 vs. DT7	88.30	88.14	0.2885	0.4102	No	No
DT7 vs. DT8	88.14	88.33	0.3411	0.4808	No	No
SVM1 vs. DT1	81.40	79.09	3.4464	4.8624	Yes	Yes
SVM2 vs. DT2	85.70	83.41	3.7603	6.1813	Yes	Yes
SVM3 vs. DT3	88.20	84.71	6.0389	8.3300	Yes	Yes
SVM4 vs. DT4	89.04	86.19	5.1394	7.0888	Yes	Yes
SVM5 vs. DT5	89.79	88.39	0.7447	1.1318	No	No
SVM6 vs. DT6	88.61	88.30	0.5823	0.8952	No	No
SVM7 vs. DT7	88.43	88.14	0.5259	0.7527	No	No
SVM8 vs. DT8	88.57	88.33	0.4499	0.6704	No	No

5. Conclusion

Supervised classification algorithms require adequate number of training samples to determine model parameters. Therefore, training sample size has crucial role in classification accuracy estimated for resulted maps in the case of supervised classification strategy. In this study, the effect of training set size was investigated for the performance of two popular classification methods, namely support vector machines (SVMs) and decision trees (DTs). Results verified the robustness of the SVMs classifier in the case of limited training samples. In other words, it was observed that SVMs produced more accurate results for limited number of samples, compared to DTs. Moreover, SVMs and DTs classifiers showed similar performance for more than 10,500 training pixels that can be regarded as optimal for the data set considered here. In addition, statistical tests results supported the finding that classification performances of SVMs and DTs improved till a certain level and the difference in classification accuracy was statistically significant when training samples are limited.

References

Boulila, W., Farah, I.R., Ettabaa, K.S., Solaiman, B., Ben Ghezala, H. (2011), "A data mining based approach to predict spatiotemporal changes in satellite images". *International Journal of Applied Earth Observation and Geoinformation*, Vol. 13(3): 386-395.

Foody, G.M. (1999), "The significance of border training patterns in classification by a feed forward neural network using back propagation learning". *International Journal of Remote Sensing*, Vol. 20(18): 3549-3562.

Foody, G.M. (2004), "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy". *Photogrammetric Engineering and Remote Sensing*, Vol. 70(5): 627-633.

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*, Vol. 103(2): 179-189.

Friedl, M.A., Brodley, C.E. (1997), "Decision tree classification of land cover from remotely sensed data." *Remote Sensing of Environment*, Vol. 61(3): 399-409.

Friedl, M.A., Sulla-Menashe, D.,Tan, B.,Schneider, A., Ramankutty, N., Sibley, A., Huang, X.M. (2010), "MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets". *Remote Sensing of Environment*, Vol. 114(1): 168-182.

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*, Vol. 23(4): 725-749.

Jain, A.K., Duin, R.P.W., Mao, J.C. (2000), "Statistical pattern recognition: A review". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22(1): 4-37.

Jia, K., Li, Q.Z., Tian, Y.C., Wu, B.F. (2011), "A Review of classification methods of remote sensing imagery". *Spectroscopy and Spectral Analysis*, Vol. 31(10): 2618-2623.

Kavzoglu, T. (2009), "Increasing the accuracy of neural network classification using refined training data". *Environmental Modelling & Software*, Vol. 24(7): 850-858.

Kavzoglu, T., Colkesen, I. (2011a), "Assessment of environmental change and land degradation using time series of remote sensing images". *Fresenius Environmental Bulletin*, 20(1A): 274-281.

Kavzoglu, T., Colkesen, I. (2011b). "Entropic distance based K-Star algorithm for remote sensing image classification". *Fresenius Environmental Bulletin,* Vol. 20(5): 1200-1207.

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

Mountrakis, G., Im, J., Ogole, C. (2011), "Support vector machines in remote sensing: A review". *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 66(3): 247-259.

Pal, M., Foody, G.M. (2010), "Feature selection for classification of hyperspectral data by SVM". *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48(5): 2297-2307.

Pal, M., Mather, P.M. (2003), "An assessment of the effectiveness of decision tree methods for land cover classification". *Remote Sensing of Environment*, Vol. 86(4): 554-565.

Zhuang, X., B.A., Engel, Lozanogarcia D.F., Fernandez R.N., Johannsen C.J. (1994), "Optimization of training data required for neuro-classification". *International Journal of Remote Sensing*, Vol. 15(16): 3271-3277.