Abstract

This thesis deals with training dataset and validation dataset for Earth observation classification accuracy improvement. Experiments with training data and validation data for two classification algorithms (Maximum Likelihood – MLC and Support Vector Machine – SVM) are carried out from the forest-meadow landscape located in the foothill of the Giant Mountains (Podkrkonoší). The thesis is base on the assumption that 1/3 of training data and 2/3 of validation data is an ideal ratio to achieve maximal classification accuracy (Foody, 2009). Another hypothesis was that in a case of SVM classification, a lower number of training point is required to achieve the same or similar accuracy of classification, as in the case of the MLC algorithm (Foody, 2004). The main goal of the thesis was to test the influence of proportion / amount of training and validation data on the classification accuracy of Sentinel – 2A multispectral data using the MLC algorithm. The highest overal accuracy using the MLC classification algorithm was achieved for 375 training and 625 validation points. The overal accuracy for this ratio was 72,88 %. The theory of Foody (2009) that 1/3 of training data and 2/3 of validation data is an ideal ratio to achieve the highest classification accuracy, was confirmed by the overal accuracy and Kappa coefficient results for MLC. It should be noted, that resulting Producer´s and User´s Accuracies for particular classes did not reach the highest values for this ratio. While size of the training dataset is sustained, further test showed that the change in the size of the validation dataset has an effect on the stability of MLC classsification accuracy assesment result. Result of overall accuracy evaluationin the case of SVM algorithm for the ratio 375 training points and 625 validation points was 79,09 %. In the case of 50 validation points the overal accuracy had reached 76,49 %. The assumption, that SVM algorithm needs lower number of training points to achieve similar classification accuracy as MLC algorithm was confirmed for this dataset.

Keywords: training dataset, validation dataset, classification accuracy improvement, Maximum Likelihood, Support Vector Machine