Comparing performance of Random Forest and SVM for shallow landslides identification

Luca Eiholzer [1]

Abstract

Landslides are a common geo-hazard especially in mountain regions, they are trigged by intense precipitations and can have negative impacts on human society in terms of costs and fatalities. In this work, the predictive performance of RF classifier and Support Vector Machine for spontaneous shallow landslide detection is compared. The results show that RF is the more reliable algorithm. In addition, noticeable differences can be seen between default's version of RF and that after a parametric search.

1. Introduction

Landslides are a common geo-hazard especially in mountain regions, and they can have negative impacts on human society in terms of costs and fatalities (Dai etal.,2002). As a result, landslide identification plays an important role in landslide risk assessment, management, and mitigation. Landslides can be identified by imagery, DEMs and field observation but those methods are expensive and time consuming. Machine learning is an efficient and accurate alternative to identify landslides (Wang et al., 2021), especially in the presence of a good quality dataset and open access as is the case in this work.

For this assignment, spontaneous shallow landslides will be considered. They are triggered by rainfall, the most common and the ones that cause the most damage. Two machine learning algorithms will be used and compared for their detection. These are Random Forest (RF) and Support Vector Machines (SVM).

In 2001, Breiman developed the first robust RF, an out-of-bag ensemble method that exploits uncorrelated forest of decision trees to solve classification and regression problems. For each decision tree, a subset of the training dataset is generated by bootstrapping: random sampling with replacement. It is a very robust and precise tool, which allows uncertainty quantification and probabilistic predictions. RF is used for a wide variety of tasks, including detecting landslides (e.g., Dai et al., 2002; Riese, 2021; Kong et al., 2021).

SVM can solve classification but also regression problems. It is an algorithm that is capable of dividing data points mapped in a high-dimensional feature. As RF, SVM can solve nonlinear problems, as most environmental phenomena are. These include landslides, and SVM is also used to detect this type of mass movements (Huang Zhao, 2018).

The aim of this work is to compare the predictive performance of RF classifier and Support Vector Machine for shallow landslide detection.

2. Dataset

The dataset was provided and pre-processed by Riese (2021), it contains 5188 observations, half of which denote the presence shallow landslides and the remaining half the absences. The presence/absence of landslides is the dependent variable, and Table 1 shows the associated predictor variables used for this assignment.

Feature name	Unit
DEM	mamsl
slope	°
planCurv	1/m
profCurv	1/m
distRoad	m
landCover	7 categories
TWI	/
geology	9 categories

Table 1. predictor variables, in the form found in the table, cf. code.

DEM correspond to the elevation, *slope* to the slope, *planCurv* to the rate of change of aspect along contour, *profCurv* to the rate of change of slope down a line, *distRoad* to the

[1]MSc student, Institute of Earth Surface Dynamics, University of Lausanne, Canton of Vaud, Switzerland. Correspondence to: Luca Eiholzer <luca.eiholzer@unil.ch>.

distance to communication routes, *landcover* to land cover information, *TWI* to the Topographic Wetness Index and *geology* to different lithologies. The choice of predictive variables was made by Riese (2021) based on data availability and literature.

The information regarding x and y-coordinates was not included in the model; it might be of interest for larger study areas (e.g., continental or world scale) but is not relevant for a study region of a few thousand km2 as is the case in this work. The latter corresponds to the entire Canton of Vaud.

3. Methods

The python language in a Google Colab environment was used to write the code for this paper. First, the dependent variable was separated from the predictor variables, also divided in numerical and categorical. The last procedure was carried out since SVM is not capable of recognizing scalars as categories, so a one-hot encoding was applied to the categorical variables ("landcover" and "geology"). Then, the dataset was randomly split into train (67%) and test (33%) subsets. Since SVM is a Euclidean distance-based approach to multi-class classification, is important to standardize all the features. The standardisation was fitted with the train subset and applied by removing the mean and scaling to unit variance.

After the pre-processing SVM and RF were implemented. Since RF performed better than SVM, a hyperparametric search was carried out for this algorithm. The best number of trees (*n_estimators*) and the best number of features to consider when looking for the best split (*max_features*) were defined using *RandomizedSearchCV* from the library *sklearn.model_selection*. According to Tonini et al. (2020), these are the two most important hyperparameters that need to be specified. An optimized version of the RF algorithm was implemented. In the case of SVM, no parametric search was performed since there is a significant difference with RF in performance with the default parameters.

As for the comparison between the three algorithms, an accuracy score, a root-mean-square error (RMSE), and a confusion matrix were computed. In addition, a receiver operating characteristic curve and a histogram of the presence of landslide probabilities will help visualize the predictive performance for shallow landslide detection.

4. Results

Table 2 shows the accuracy score and the RSME for the three algorithms implemented for shallow landslide detection.

Accuracy corresponds to the number of data correctly classified (true positive and true negative) over the total number of data instances. RMSE is a common indicator of the difference between the values predicted by a model and the true values.

	SVM	**RF**	**RF_{opt}**
Accuracy	0.811	0.836	0.842
RMSE	0.434	0.405	0.397

Table 2. Accuracy score and root-mean-square deviation

	SVM		**RF**		**RF_{opt}**	
	TP	TA	TP	TA	TP	TA
PP	657	187	683	161	688	156
PA	136	733	120	749	114	755

Table 3. Confusion matrices.TP: true presence, TA: true absence, PP: predicted presence, PA: predicted absence.

Table 3 shows the confusion matrices for the three algorithms.

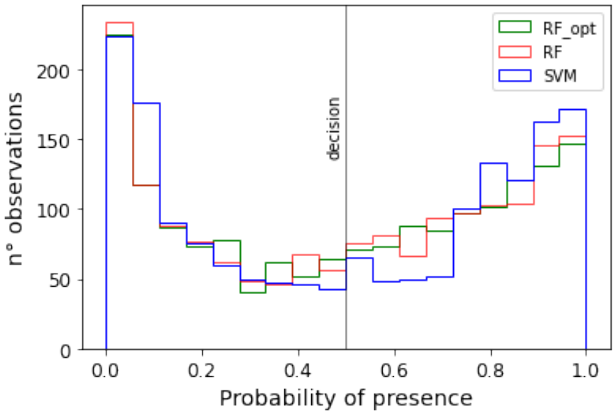

Figure 1. Presence of landslide probabilities

Figure 1 shows the presence of landslide probabilities. These probabilities are calculated based on how many times a data instance was associated with the presence and respectively absence of landslide over the total number of trees in the forest. Once finished, random forest will then make a decision based on the class that has been associated multiple times with a data instances. In this case, the threshold is set at 0.5 since the prediction is made for a binary variable. However, this threshold can be shifted if necessary.

Figure 2 shows receiver operating characteristic curves. These curves make it possible to compare the performace of binary classifiers in terms of sensitivity (true positive rate) and specificity (false positive rate).

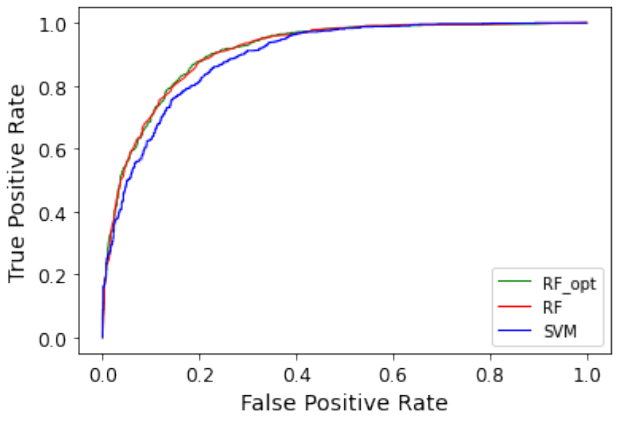

Figure 2. ROCs

5. Discussion

As shown in Table 2, the optimized version of RF is the most accurate algorithm with an accuracy score of 0.842. interestingly, the hyperparametric search allowed an improvement in this value of 0.6%. SVM has an accuracy score of 0.811 and is the least accurate algorithm in this work. As for RSME, the comparison follows the previous one: the improved version of RF has the lower RSME (0.397), followed by the default version (0.405) and finally we find SVM (0.434). Instead, looking at Table 3 it denotes that the three algorithms when making a prediction error tend to decide for a false positive rather than a false negative. The algorithm that makes the most such errors is SVM, followed by RF and its improved version. As for false negatives, the ranking among the algorithms is the same.

Regarding the presence of landslide probabilities, the three algorithms have similar behaviour, but some differences are denoted. For example, globally SVM seems to be the strictest algorithm as it has peaks at the ends. It also has the highest number of observations regarding the last column on the right (high probability of landslide presence), while the algorithm possessing the most observations in the ultimate column on the left (high probability of landslide absence) is RF. Finally, the optimized version of RF seems to be the least severe, being that there is less difference between the observations at the extremities of the graph and those located in the centre. Finally, ROCs curves show that the two RF algorithms are better classifiers than SVM in the case of this work. The optimized version of RF differs little positively from the default version. It can also be seen that the shapes of the curves resemble each other.

6. Conclusion

In this work, the predictive performance of RF classifier and Support Vector Machine for shallow landslide detection was compared. The results show that RF is the more reliable algorithm. In addition, noticeable differences can be seen between default's version of RF and that after a parametric search.

In future work, more algorithms could be compared. In addition, it would be interesting to compare the performance of the best algorithm with average of all algorithms, to see if the latter leads to better results, compensating for the defects inherent in each specific algorithm.

7. References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Dai, F. C., Lee, C. F., Ngai, Y. Y. (2002). Landslide risk assessment and management: An overview. *Engineering Geology*, *64*(1), 65-87. https://doi.org/10.1016/S0013-7952(01)00093-X

Huang, Y., Zhao, L. (2018). Review on landslide susceptibility mapping using support vector machines. *CATENA*, *165*, 520-529. https://doi.org/10.1016/j.catena.2018.03.003

Riese, J. (2021). *Landslide susceptibility mapping in the canton of Vaud using random forest, A focus on the marginal effect of different predictive variables* [MSc]. Université de Lausanne.

Kong, C., Tian, Y., Ma, X., Weng, Z., Zhang, Z., Xu, K. (2021). Landslide Susceptibility Assessment Based on Different MaChine Learning Methods in Zhaoping County of Eastern Guangxi. *Remote Sensing*, *13*(18), 3573. https://doi.org/10.3390/rs13183573

Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., Fiorucci, P. (2020). A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. *Geosciences*, *10*(3), 105. https://doi.org/10.3390/geosciences10030105

Trucchia, A., Meschi, G., Fiorucci, P., Gollini, A., Negro, D. (2022). Defining Wildfire Susceptibility Maps in Italy for Understanding Seasonal Wildfire Regimes at the National Level. *Fire*, *5*(1), 30. https://doi.org/10.3390/fire5010030

Wang, H., Zhang, L., Luo, H., He, J., Cheung, R. W. M. (2021). AI-powered landslide susceptibility assessment in Hong Kong. *Engineering Geology*, *288*, 106103. https://doi.org/10.1016/j.enggeo.2021.106103

8. Appendices

The code used for this work can be found here.

The dataset used for this work can be found here.