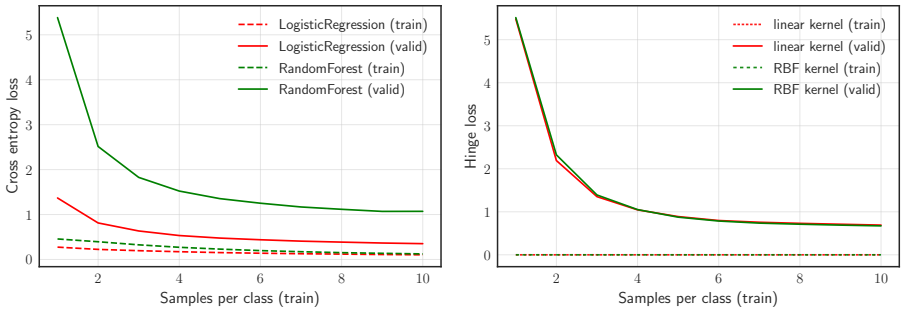
Machine learning for LC-MS medicinal plants identification

D.V. Nazarenko, P.V. Kharyuk, I.V. Oseledets, I.A. Rodin, O.A. Shpigun

April 28, 2019

Table 1: Comparative characteristics of implemented approaches on original data. Metrics were calculated for validation set.

Method	**Accuracy, %**	**Precision, %**	**Recall, %**	**F1, %**
LogisticRegression	94.77	93.71	94.56	93.45
SVM (linear)	87.88	87.05	87.75	85.84
SVM (RBF)	87.54	86.66	87.75	85.42
RandomForest	94.37	94.22	94.55	93.53



(a) (b)

Figure 2: Learning curves for tuned classification algorithms: dashed line indicates results for training set, solid line for validation, (a) in terms of cross entropy loss (logistic regression in red and random forest in green) and (b) hinge loss (SVM with linear kernel in red and RBF kernel in green)

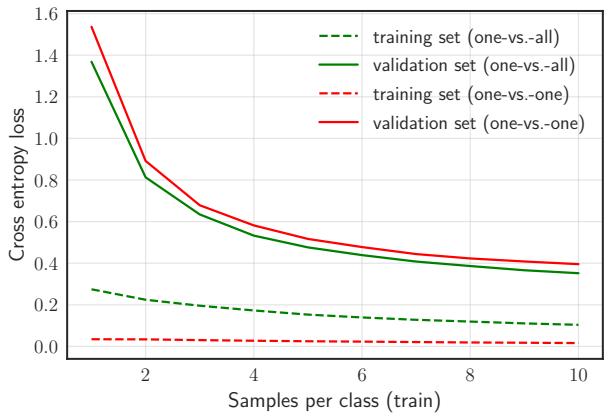


Figure 3: Comparison of "one-vs-all" and "one-vs-one" strategies for logistic regression classification.

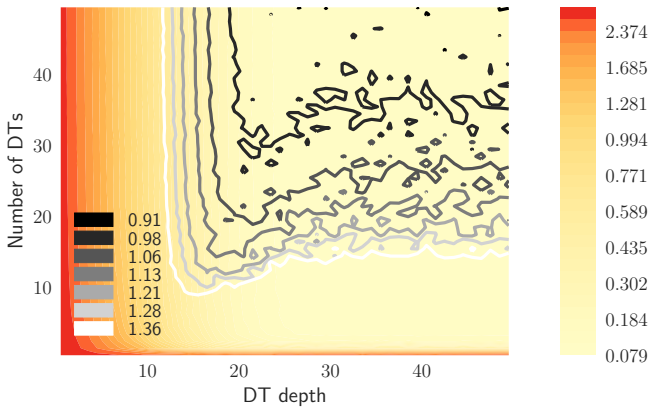


Figure 4: Logistic loss values for random forest according to number of decision trees and depth of each tree. Heat map corresponds to losses on training dataset; contour lines coincide with losses on validation data.

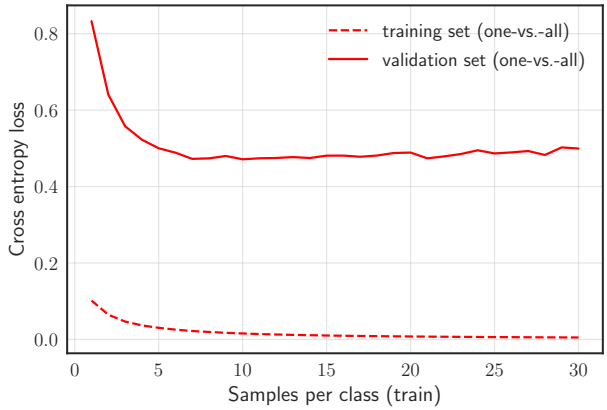


Figure 5: Learning curves for logistic regression on artificially generated data, dashed line indicates results for training set, solid line for validation set, in terms of cross entropy loss