Selecting the most suitable classification algorithm for tiger beetle identification using morphometric data and habitat data

D. L. Abeywardhana[1], C. D. Dangalle[1]*, Anupiya Nugaliyadde[2], Y.W. Mallawarachchi[3]

[1]Department of Zoology and Environment Sciences, University of Colombo, Sri Lanka.
[2]Murdoch University, Perth, Australia.
[3]Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.

Habitat heterogeneity is a main factor in ecology which affects species diversity. Therefore, habitat details can be used as factors that influence species identification. Tiger beetles are highly habitat specific species. Different species of tiger beetles that have morphometric variation can be found restricted to different habitat types in temperate and tropical areas of the world. Therefore, habitat and morphometric data of tiger beetle species were used to develop a predictive model for the identification of tiger beetle species. Data gathered on ground-dwelling tiger beetle species collected from 45 locations by Dangalle (2002-2015), and 150 locations by Thotagamuwa (2014 -2017) were used to construct the dataset required for the study. Then data pre-processing was done to convert nominal data to numerical data, detach records with missing data and correct imbalanced data. Data resampling was also done. Further, an evaluation was performed for feature selection to determine the most important attributes for species identification predictions. Finally, a dataset containing 468 records (individuals of species) having 13 attributes of 14 species was constructed and 351 records were selected for the training set while 117 records were selected for the test set. This dataset was fed to several multiclass algorithms belonging to both single (KNN, Naïve-Bayes, SVM) and ensemble classifiers (Gradient Tree Boosting, Extra Tree Classifier and Random Forest). The performance of each algorithm was evaluated by calculating accuracy values for each model and the results revealed that ensemble classifiers yield a higher accuracy than that of single classifiers. Hence it was proven that ensemble models have a positive effect on the overall quality of predictions, in terms of accuracy, generalizability and lower misclassification costs and are more stable than single classifiers. Further, when considering the different types of ensemble classification algorithms, bagging (averaging) ensemble classification algorithm performed better than boosting methods. When considering the two bagging ensemble classification algorithms - Ensemble Extra Tree Classifier and Random Forest algorithm, both revealed almost the same overall accuracy (85%) with less than 0.12% difference. Therefore, both ensemble classification algorithms are effective for species prediction using habitat and morphometric data. However, when considering the computational time with performance, Ensemble Extra Trees Classifier can be considered as the most suitable algorithm for the scenario.

Keywords: Ensemble classifiers, single classifiers, tabular data, tiger beetles
* cddangalle@gmail.com

Acknowledgement: National Science Foundation of Sri Lanka (Grant No. RG/2017/EB/01).