The Journal of Open Source Software

sdmbench: R package for benchmarking species distribution models

Boyan Angelov[1]

1 MindMatch GmbH

DOI: 10.21105/joss.00847

Software

- Review
- Repository
- Archive

Submitted: 11 July 2018
Published: 28 September 2018

Summary

Species Distribution Modeling (SDM) is an emerging field in ecology. The effects of anthropogenic climate change, habitat destruction (deforestation, pollution) and poaching are observable in ecosystems around the world (Elith & Leathwick, 2009). SDMs have been used to address those challenges with notable success in estimating the effects of climate change on species distributions (M. P. Austin & Van Niel, 2011), natural reserve planning (Guisan et al., 2013) and predicting invasive species distributions (Descombes et al., 2016).

Steady improvements in analytical tools (new machine learning algorithms and faster processors in particular) and larger amounts of data gathered recently opened up new possibilities in the field. Although researchers benefit from these new tools, they face challenges in method selection and evaluation. In the case of machine learning algorithms, this issue is particularly relevant. How to decide which algorithm to use, knowing that model performance can vary significantly between datasets? And, how do we compare models in a consistent manner, without introducing additional bias? How can we demonstrate the improvement of a new method over the current state-of-the-art?

`sdmbench` is an R package to benchmark machine learning methods for SDM, helping researchers tackle those questions of selection and evaluation. It is inspired by similar projects in computational chemistry (Wu et al., 2017) and healthcare (Purushotham, Meng, Che, & Liu, 2017).

`sdmbench` takes a different approach to benchmarking than previously-published software packages. ENMEval (Muscarella et al., 2014) and SDMSelect (Rochette, 2017) have also addressed this challenge, but with a focus on the Maximum Entropy (MaxEnt) model and covariate selection. To summarise, the differentiating features of `sdmbench` are:

- consistent species occurrence data acquisition and preprocessing
- consistent environmental data acquisition (both current data and future projections) and preprocessing, domain-specific cross-validation
- integration of a wide variety of machine learning models and plotting utilities
- a graphical user interface (GUI) for inexperienced users

`sdmbench` obtains species occurrence data and environmental variables from GBIF (https://www.gbif.org/) and Worldclim (http://worldclim.org/), respectively. The user can also specify the type (historical data or IPCC projections) and resolution of the climate data. These popular and stable data repositories ensure high data quality. The data processing pipeline relies on external packages. The scrubr (Chamberlain, 2016) package is used to clean the occurrence data (i.e. removal of duplicates and unlikely coordinates). ENMEval provides domain-specific cross-validation to mitigate spatial autocorrelation effects that might adversely affect model accuracy. An additional processing option that can be specified is data undersampling. This feature introduces synthetic class imbalance in the data that in turn can test the effectiveness of model training on sparse datasets.

`sdmbench` v0.1.3 supports 10 popular machine learning methods in addition to MaxEnt, including neural networks (Tensorflow via Keras). Methods can be compared quantitatively by computing their Area Under the Curve (AUC, a standard procedure for machine learning classification tasks), and visually by inspecting the resulting species distribution maps. The same workflow can also be accomplished in the GUI, allowing for rapid exploration and prototyping.

`sdmbench` is available from GitHub (https://github.com/boyanangelov/sdmbench), and archived on Zenodo (http://doi.org/10.5281/zenodo.1436376).

Acknowledgements

The author would like to thank Dr. Brandon Seah and Dr. Rick Scavetta for providing feedback on the manuscript.

References

Austin, M. P., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, *38*(1), 1–8. doi:10.1111/j.1365-2699.2010.02416.x

Chamberlain, S. (2016). *Scrubr: Clean biological occurrence records*. Retrieved from https://CRAN.R-project.org/package=scrubr

Descombes, P., Petitpierre, B., Morard, E., Berthoud, M., Guisan, A., & Vittoz, P. (2016). Monitoring and distribution modelling of invasive species along riverine habitats at very high resolution. *Biological Invasions*, *18*(12), 3665–3679. doi:10.1007/s10530-016-1257-4

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. doi:10.1146/annurev.ecolsys.110308.120159

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435. doi:10.1111/ele.12189

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENM eval: An r package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models. *Methods in Ecology and Evolution*, *5*(11), 1198–1205. doi:10.1111/2041-210X.12261

Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2017). Benchmark of deep learning models on large healthcare MIMIC datasets. *CoRR*, *abs/1710.08531*. Retrieved from http://arxiv.org/abs/1710.08531

Rochette, S. (2017, September). SDMSelect: A r-package for cross-validation model selection and species distribution mapping. doi:10.5281/zenodo.894344

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., et al. (2017). MoleculeNet: A benchmark for molecular machine learning. *CoRR*, *abs/1703.00564*. doi:10.1039/C7SC02664A