Machine Learning for Species Identification: The *Hebeloma* Project from database to website

Peter Bartlett[‡], Ursula Eberhardt[§], Nicole Schütz[§], Henry J. Beker[|,¶,#]

‡ La Baraka, Gorse Hill Road, Virginia Water, United Kingdom
§ Staatliches Museum für Naturkunde Stuttgart, Rosenstein 1, D-70191, Stuttgart, Germany
| Rue Père de Deken 19, B-1040, Bruxelles, Belgium
¶ Royal Holloway University of London, London, United Kingdom
# Plantentuin Meise, Nieuwelaan 38, B-1860, Meise, Belgium

Corresponding author: Peter Bartlett (pete@pcbartlett.com)

Abstract

Attempts to use machine learning (ML) for species identification of macrofungi have usually involved the use of image recognition to deduce the species from photographs, sometimes combining this with collection metadata. Our approach is different: we use a set of quantified morphological characters (for example, the average length of the spores) and locality (GPS coordinates). Using this data alone, the machine can learn to differentiate between species.

Our case study is the genus *Hebeloma*, fungi within the order Agaricales, where species determination is renowned as a difficult problem. Whether it is as a result of recent speciation, the plasticity of the species, hybridization or stasis is a difficult question to answer. What is sure is that this has led to difficulties with species delimitation and consequently a controversial taxonomy.

The Hebeloma Project—our attempt to solve this problem by rigorously understanding the genus—has been evolving for over 20 years. We began organizing collections in a database in 2003. The database now has over 10,000 collections, from around the world, with not only metadata but also morphological descriptions and photographs, both macroscopic and microscopic, as well as molecular data including at least an internal transcribed spacer (ITS) sequence (generally, but not universally, accepted as a DNA barcode marker for fungi (Schoch et al. 2012)), and in many cases sequences of several loci. Included within this set of collections are almost all type specimens worldwide. The collections on the database have been analysed and compared. The analysis uses both the morphological and molecular data as well as information about habitat and location. In this way, almost all collections are assigned to a species. This development has been enabled and assisted by citizen scientists from around the globe, collecting and recording information about their finds as well as preserving material.

From this database, we have built a website, which updates as the database updates. The website (hebeloma.org) is currently undergoing beta testing prior to a public launch. It includes up-to-date species descriptions, which are generated by amalgamating the data from the collections of each species in the database. Additional tools allow the user to explore those species with similar habitat preferences, or those from a particular biogeographic area. The user is also able to compare a range of characters of different species via an interactive plotter.

The ML-based species identifier is featured on the website. The standardised storage of the collection data on the database forms the backbone for the identifier. A portion of the collections on the database are (almost) randomly selected as a training set for the learning phase of the algorithm. The learning is “supervised” in the sense that collections in the training set have been pre-assigned to a species by expert analysis. With the learning phase complete, the remainder of the database collections may then be used for testing. To use the species identifier on the website, a user inputs the same small number of morphological characters used to train the tool and it promptly returns the most likely species represented, ranked in order of probability.

As well as describing the neural network behind the species identifier tool, we will demonstrate it in action on the website, present the successful results it has had in testing to date and discuss its current limitations and possible generalizations.

Keywords

artificial intelligence, AI, *Agaricales*, ectomycorrhizal fungi, identification keys, neural network, type sequences

Presenting author

Peter Bartlett

Presented at

TDWG 2021

References

- Schoch C, Seifert K, Huhndorf S, Robert V, Spouge J, Levesque CA, Chen W, Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS 109 (16): 6241-6246. https://doi.org/10.1073/pnas.1117018109