Learning Formulas to Capture Biodiversity

Our theoretical understanding of biodiversity requires data that captures the full complexity of ecosystems. Data-sets often only tell us whether a species is present or absent from a location, with little to no information on the age structure of the populations, the spatial organization of the community, or even population sizes. As noted by Weinstein et al.: "Data acquisition currently outpaces the ability to identify individual organisms in high resolution imagery" [11]. Thus, one way to build rich data-sets with detailed information on the spatial structure of the organisms (and even their age) is to develop techniques to detect and classify living organisms from high-resolution aerial pictures.

The rise of deep learning has drastically improved our ability to detect and classify complex objects in images [3]. Arguably one of the most impressive achievement of deep learning is its ability to learn to play Atari games from the visual input alone, without prior knowledge of the rules of the games [7]. However, deep learning has a few important drawbacks, most notably: the resulting models are difficult to interpret and do not mix well with expert knowledge. An alternative approach is to use genetic algorithms [6] to learn mathematical formulas (Figure 1). This approach is called symbolic regression [10] and has already been used to learn mathematical formulas in ecology [1]. Symbolic regression was recently used to recreate the Atari paper [12], with the resulting models being clear, easy to understand mathematical formulas. Furthermore, since the resulting models are simply mathematical formulas or small programs, we can easily mix them with existing knowledge.

The project's objective is to develop and test symbolic regression in the context of identifying and classifying living organisms in images. It fits within a larger project to grow a knowledge base for theories of biodiversity [2]. Symbolic regression would be useful both to add formulas to our knowledge base but also to search for possible improvements of existing theories. The project sits at the intersection of machine learning, ecology, and even formal mathematics.

References

[1] M Alfonseca and FJ Soler Gil. Evolving a predator–prey ecosystem of mathematical expressions with grammatical evolution. *Complexity*, 20(3):66–83, 2015.

[2] P Desjardins-Proulx, T Poisot, and D Gravel. artificial intelligence for ecological and evolutionary synthesis. *Frontiers in Ecology and Evolution*, 7:402, 2019.

[3] I Goodfellow, Y Bengio, and A Courville. *Deep Learning*. MIT Press, 2016.

[4] JF Miller. An empirical study of the efficiency of learning boolean functions using a cartesian genetic programming approach. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 2*, pages 1135–1142. Morgan Kaufmann, 1999.

[5] JF Miller. *Cartesian Genetic Programming*, pages 17–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

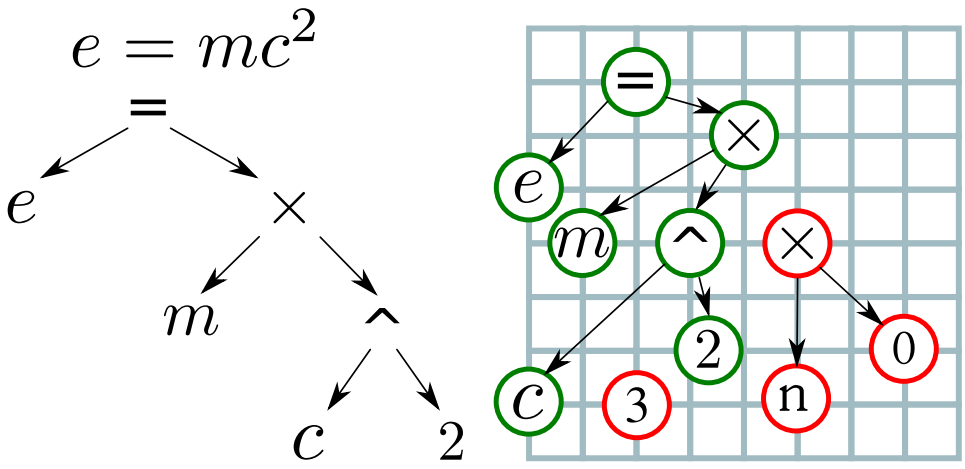

Figure 1: In genetic algorithms, a population of solutions undergoes a process of crossover (mixing solutions), mutation (random modifications to the solutions), and selection (choosing the best solution according to some objective). Genetic algorithms work better with flat fixed-size data structures for the process of mixing and mutating solutions. Unfortunately, mathematical formulas are trees (e.g. Einstein's mass-energy equivalence on the left). Several solutions have been suggested to evolve mathematical formulas, with two popular options being grammatical evolution [8, 9] and Cartesian Genetic Programming [4, 5]. On the right-side: the formula is placed on a grid for Cartesian Genetic Programming. The nodes in green are active, they represent the formula on the left. The red nodes are currently inactive, although mutations could reactivate them. Laying the formula on a grid limits the size of the formula while allowing simple techniques to be used for mixing solutions and mutating them.

[6] M Mitchell. *An introduction to genetic algorithms*. MIT Press, 1999.

[7] V Mnih, K Kavukcuoglu, D Silver, A Graves, I Antonoglou, D Wierstra, and M Riedmiller. Playing atari with deep reinforcement learning. *arXiv*, 2013.

[8] M O'Neill and C Ryan. Grammatical evolution. *IEEE Transactions on Evolutionary Computation*, 5(4):349–357, 2001.

[9] M O'Neill and C Ryan. *Grammatical Evolution*. Springer, 2003.

[10] P Orzechowski, W La Cava, and JH Moore. Where are we now? a large benchmark study of recent symbolic regression methods. 2018.

[11] BG Weinstein, S Marconi, S Bohlman, A Zare, and E White. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11:1309, 2019.

[12] DG Wilson, S Cussat-Blanc, H Luga, and JF Miller. Evolving simple programs for playing Atari games, 2018.