Scientific Process Discovery Using Scientific Approaches

Yexiang Xue, Purdue University, yexiang@purdue.edu

Automating scientific discovery, in particular, discovering scientific equations that best fit the experiment data, has been a grand goal of AI dating back to its founders (Herbert Simon et. al.) but remains a holy grail. Indeed, much effort has been made, especially in symbolic equation finding. This includes search-based methods, genetic programming, reinforcement learning, deep function approximation, and integrated systems. Most endeavors treat the task of equation finding as a standard learning task -- they directly search for the best equation in the full hypothesis space involving all the independent variables that fits a pre-collected, static dataset. This type of search can be challenging because the search space of all candidate equations is exponentially large.

Recently, a promising direction has been to accelerate AI-driven scientific discoveries using scientific approaches. This approach has been pioneered by the BEACON system [1], followed by a series of works such as Inductive Logic Programming [2], and Control Variable Genetic Programming [3]. For example, to discover the ideal gas law $pV = nRT$, the scientific approach first holds n (gas amount) and T (temperature) as constants in controlled variable experiments and finds p (pressure) is inversely proportional to V (volume). This results in a reduced-form equation $pV = constant$. This reduced-form equation then can be expanded to include n, again using controlled experiments in which only pV and n are allowed to vary. Such processes repeat until the full equation is found. The discovery of the first few reduced-form equations can be significantly cheaper than other approaches, because the searches are in reduced spaces involving a small number of independent variables. As a result, this approach has the potential to supercharge state-of-the-art approaches in modeling complex scientific phenomena with many interlocking contributing factors.

Nevertheless, in real-world applications, it is often not the number of independent variables, but the intertwined nature of many physical processes, that dominate the complexity of a physical system. For example, dendritic solidification refers to the natural phenomenon in which solid-state materials (e.g., ice) emerge from supercooled liquids (e.g., water). It involves several intertwined processes, more precisely, phase separation, temperature dependence, diffusivity along the liquid-solid interfaces, and the differences in the mobility rates among interfaces of different directions. The simultaneous existence of all these physical processes overwhelms state-of-the-art learning algorithms that aim to learn the full-fledged model in one shot.

We propose scientific process discovery using scientific approaches. The key idea is to drive the discovery using an integrated AI system composing of hypothesis generator, an experiment designer, and a model learner. Our proposed approach breaks down the learning of such complex phenomena into multiple stages. In each stage, a single physical process is learned. For example, in learning dendritic solidification, we first restrict our search within the space of simple phase separation models that do not take into account complicated factors (e.g., temperature dependence, diffusivity, etc.). Then, we start to refine this model by introducing additional physical processes at play, one at a time. We first notice the phase separation process depends on temperature, then model the diffusivity along the liquid-solid interfaces, and then the directional differences in the growth rate. In each stage, we employ three closely collaborated AI-empowered agents:

(1) **Hypothesis generator**: The hypothesis generator criticizes the current model, suggesting that certain constants are better expanded into sub-processes. It accomplishes this by fitting the current model to different portions of the experiment data and noticing the fitted values of constants vary substantially. This implies the possibility that the constant is a sub-process.

(2) Experiment designer: The experiment designer decides the best experiment data which best reveals the dependent factors of a sub-process at-study. For example, the experiment designer will actively collect experiment data relating to solidifications at interfaces of different directions when we model the directional dependence of diffusivity in the dendritic solidification process.

(3) Model learner: The model learner expands a constant in the current model into a sub-process. It employs various machine learning algorithms in the search of the best sub-process.

References

[1] Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, *5*, 31–54.

[2] Džeroski, S., & Todorovski, L. (1995). Discovering dynamics: From inductive logic programming to machine discovery. *Journal of Intelligent Information Systems*, *4*, 89–108.

[3] Jiang, N. & Xue, Y. (2023). Symbolic Regression via Control Variable Genetic Programming. *Proceedings of the 2022 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).* 4, 178 – 195.