Designing a Massive Dataset Framework for the Grammatical Evolution System

Miguel Nicolau
University College Dublin
Miguel.Nicolau@ucd.ie

Abstract

Data is ubiquitous. Most activities nowadays are logged in some way, such as taking a bus, connecting to wi-fi, paying dinner with a card, or even reading a book. Even mechanical devices such as light bulbs or bicycles are becoming connected, data-generating devices.

This provides fantastic opportunities to improve business processes, while posing challenges on how to extract intelligence from such amounts of data. Computer modelling is used extensively to achieve this; yet, with all the number-crunching power available, the intuition of human analysis to make sense of the underlying models is often desirable.

Genetic Programming algorithms are well suited as model-building techniques that sit between data and analysts. By generating symbolic solutions (such as mathematical formulas), they provide insight into relationships between predictors and complexity of models, and such models are easier to analyse and even modify. Variants such as Grammatical Evolution (GE) further add to this utility, by specifying the syntax of the generated solutions through a grammar.

In this study, a GE-framework is built, in an effort to apply it to huge datasets. Combining statistical techniques such as appropriate error measures and data splitting, population-based improvements such as mass parallelisation, and even specific techniques such as grammar design and repeat management, GE is applied for the first time to massive datasets, such as the Higgs dataset (eleven million samples).