Chapter 7

INVESTIGATING PROBLEM HARDNESS OF REAL LIFE APPLICATIONS

Leonardo Vanneschi[1]
[1]*Dipartimento di Informatica, Sistemistica e Comunicazione (D.I.S.Co.)*
University of Milano, Bicocca Milan, Italy

Abstract This chapter represents a first attempt to characterize the fitness landscapes of real-life Genetic Programming applications by means of a predictive algebraic difficulty indicator. The indicator used is the Negative Slope Coefficient, whose efficacy has been recently empirically demonstrated on a large set of hand-tailored theoretical test functions and well known GP benchmarks. The real-life problems studied belong to the field of Biomedical applications and consist of automatically assessing a mathematical relationship between a set of molecular descriptors from a given dataset of drugs and some important pharmacokinetic parameters. The parameters considered here are Human Oral Bioavailability, Median Oral Lethal Dose, and Plasma Protein Binding levels. The availability of good prediction tools for pharmacokinetics parameters like these is critical for optimizing the efficiency of therapies, maximizing medical success rate and minimizing toxic effects. The experimental results presented in this chapter show that the Negative Slope Coefficient seems to be a reasonable tool to characterize the difficulty of these problems, and can be used to choose the most effective Genetic Programming configuration (fitness function, representation, parameters' values) from a set of given ones.

Keywords: problem difficulty, fitness landscapes, real life applications, fitness clouds, negative slope coefficient

1. Introduction

Is Genetic Programming (GP) a suitable tool to solve my problem? How can I set parameters to make GP find better solutions? Which fitness function and representation should I choose? Although GP has been applied with success to a large number of applications of many different kinds, and besides the large amount of theory that has appeared to date – see for instance (Koza and Poli,

2003) for a short survey of GP theory and applications – still the answers to these questions are, given a particular problem, in large part unknown. What practitioners usually do when they are faced with a new complex combinatorial optimization problem is execute a set of simulations using many different GP configurations, with many different parameter settings, fitness functions and representations, and possibly other alternative Machine Learning strategies. From a comparison between the results of all these simulations, practitioners often try to empirically find the "best" algorithm and configuration for their problem. This procedure, although often successful, is very time-and computational resource-consuming. Furthermore, results of GP simulations are often difficult to interpret, given that GP, like many other optimization metaheuristics, is a stochastic (and thus non-deterministic) process. On the other hand, answering the previous questions would be much easier if an *algebraic measure* existed able, given a particular GP configuration (fitness function, representation, parameters' values), to quantify its ability to solve a given problem, *without* running GP itself.

Difficulty studies in Genetic Algorithms (GAs) have been pioneered by Goldberg and coworkers – e.g., see (Horn and Goldberg, 1995). One concept that underlies many of these studies is the notion of *fitness landscape* – e.g. see (Stadler, 2002). The fitness landscape plot can be helpful to understand the difficulty of a problem, i.e. the ability of a searcher to find good solutions for that problem – see for instance (Vanneschi, 2004; Langdon and Poli, 2002) for a deep analysis. Nevertheless, fitness landscapes are impossible to plot in practice, given the generally huge size of the space of solutions and the multi-dimentionality and complexity of the possible neighborhood structures. For this reason, in the last few years researchers have been looking for an algebraic measure able to capture some of the interesting properties of fitness landscapes. Early attemps are represented by (Weinberger, 1990; Manderick et al., 1991; Kinnear, Jr., 1994). A signicant contribution to this field has been given by Jones (Jones, 1995) with the introduction of an hardness measure for GAs called *fitness distance correlation* (*fdc*). This measure has been extended to tree-based GP and proven a suitable hardness indicator in (Vanneschi, 2004; Tomassini et al., 2005). Nevertheless, these contributions have also shown that *fdc* has some flaws, the most important one being the fact that *fdc* is not predictive, i.e. the optimal solution (or solutions) must be known beforehand, which is almost unrealistic in applied search and optimization problems. Thus, it is important to investigate other approaches based on quantities that can be measured without any explicit knowledge of the genotype of optimal solutions. Preliminary results of this enquiry can be found in (Vanneschi, 2004; Vanneschi et al., 2004; Vanneschi et al., 2006), where a new measure called *negative slope coefficient* (*nsc*) has