Analysis of neutrality in GP Boolean Fitness Landscapes

Abstract

Yuri Pirola
matr. n. 060962

Supervisors

Dott. Leonardo Vanneschi

Prof. Giancarlo Mauri

Academic year 2004/2005

Corso di Laurea Magistrale in Informatica
Dipartimento di Informatica, Sistemistica e Comunicazione - DISCo

Facoltà di Scienze Matematiche, Fisiche e Naturali
Università degli Studi Milano Bicocca

Goals and motivations

Some real-world problems have successfully been solved by Genetic Programming (GP) during last years, but some difficulties persist. First of all, a method to predict problem difficulty for GP (i.e. the ability of GP to find a "good" solution) is not yet known. However, a common supposition in this research field claims a relation between *problem difficulty* and *neutrality* of the associated fitness landscape. Fitness landscapes are useful models to develop an insight about the working of GP. Neutrality is the phenomenon of having *different* solutions at the *same quality level*. The thesis presents an investigation of neutrality in a specific class of GP fitness landscapes. Comparing fitness landscapes induced by different problems, we try to find some differences that can justify the different difficulties for GP to solve the problem.
Our study differs from previous works because (1) it considers standard tree-based GP and (2) we study neutrality of the fitness landscape without modifying it.

The main original contributions are: (a) the theoretical study of some relevant characteristics of some fitness landscapes, (b) the definition of some new measures that help characterizing landscapes neutrality and (c) the experimental analysis of some landscapes.
First of all some "small" landscapes have been considered, then a sampling analysis of "larger" landscapes has been performed. Well-known sampling methodologies are not suitable to generate samples of the fitness landscapes studied here because they disregard some of their interesting properties. Therefore, a new methodology was developed and used.

The experimental results support the suitability of the sampling method presented here for the first time for the specific class of landscapes considered. Moreover, the different characteristics of the landscapes seem to justify the different difficulties for GP to solve the associated problems.

Genetic Programming and Fitness Landscapes

Genetic Programming (GP) is a machine learning technique, introduced by Koza in 1992 [1], to automatically generate computer programs that perform a specific task. GP is inspired by the Darwinian theory of evolution and it works by selecting and varying a *population* of programs. The selection process is driven by a quality criterion expressed by means of a fitness function. The variation phase is performed by a set of genetic operators.

Different representations for the solutions have been proposed in GP lit-

erature. The most commonly used is the one based on tree structures, as originally defined by Koza. Other interesting variants include representations based on linear or graph structures.

A fitness landscape can be defined as a triple $\mathcal{L} = (\mathcal{S}, \mathcal{V}, \mathsf{f})$ where $\mathcal{S}$ is the set of all possible solutions, $\mathcal{V} : \mathcal{S} \rightarrow 2^{\mathcal{S}}$ is a neighbourhood function specifying, for each $\mathsf{s} \in \mathcal{S}$, the set of its neighbours $\mathcal{V}(\mathsf{s})$, and $\mathsf{f} : \mathcal{S} \rightarrow \mathbb{R}$ is the fitness function. Generally, the neighbourhood relationship is defined in terms of the variation operators used, so $\mathcal{V}$ can be defined as $\mathcal{V}(\mathsf{s}) = \{\mathsf{s}' \in \mathcal{S} | \mathsf{s}'$ can be obtained from s by a single variation$\}$.

In some cases, even though the size of the search space $\mathcal{S}$ is huge, f can only assume a limited set of values. Thus, a large number of solutions have the same fitness. In this case, we say that the landscape has a high degree of neutrality [2]. Given a solution s, a particular subset of $\mathcal{V}(\mathsf{s})$ can be defined: the one composed by neighbour solutions that are also neutral. Formally, the *neutral neighbourhood* of s is the set $\mathcal{N}(\mathsf{s}) = \{\mathsf{s}' \in \mathcal{V}(\mathsf{s}) | \mathsf{f}(\mathsf{s}') = \mathsf{f}(\mathsf{s})\}$.

Given these definitions, it is possible to imagine a fitness landscape as being composed by a set of (possibly large) *neutrality plateaus*. More formally, a *neutral network* [3] can be defined as a connected component of the graph $(\mathcal{S}, \mathsf{E}_{\mathcal{N}})$ where $\mathsf{E}_{\mathcal{N}} = \{(\mathsf{s}_1, \mathsf{s}_2) \in \mathcal{S}^2 | \mathsf{s}_2 \in \mathcal{N}(\mathsf{s}_1)\}$.

The Even Parity Problem

The goal of the even-k parity problem [1] is to find a boolean function of k variables that returns True if an even number of inputs is True and False otherwise. Fitness is computed as the number of errors over the 2^{k} fitness cases represented by all the possible combinations of the k input values.

This problem is well-known in literature because the desired boolean function is usually difficult to obtain by most of ML techniques.

The set of all possible solutions is composed by all the well–formed trees that can be built using a function set $\mathcal{F}$ and a terminal symbols set $\mathcal{T}$ and having a depth smaller or equal than a given limit. The set $\mathcal{T}$ is composed by k boolean variables (where k is the order of the problem). Two different function sets are studied in our work: $\{\text{XOR}\,;\ \text{NOT}\}$ and $\{\text{NAND}\}$. These function sets induce two fitness landscapes (that we indicate with $\mathcal{L}^{\{\text{XOR}\,;\,\text{NOT}\}}_{(\mathsf{k},\mathsf{h})}$ and $\mathcal{L}^{\{\text{NAND}\}}_{(\mathsf{k},\mathsf{h})}$, where k is the problem order and h is the prefixed tree-depth limit) with different difficulties for GP: the landscape induced by $\{\text{XOR}\,;\ \text{NOT}\}$ is easy to search, while the one induced by $\{\text{NAND}\}$ is generally hard.

To define a neighbourhood structure, have defined a simplified version of the structural mutation operators first introduced in [4] that we call *strict structural mutations* to distinguish from the former ones. This set of mu-

tations is easy enough to study and provides enough exploration power to GP.

The {XOR ; NOT} landscape presents some important properties that are derived and proved here for the first time. First of all, supposing that all fitness values have been normalized into the range $[0, 1]$, if an expression does not contain at least an occurrence of each variable, then its fitness value is exactly equal to 0.5. For this reason, the wide majority of individuals in the even parity landscapes have fitness 0.5. Secondly, an expression in the $\mathcal{L}^{\{\text{XOR}\,;\,\text{NOT}\}}$ landscape can only have a fitness value equal to 0, 0.5 or 1.

The choice of the strict structural mutation operators permits to define some other properties of the $\mathcal{L}^{\{\text{XOR}\,;\,\text{NOT}\}}$ landscape: (a) there is *only* one neutral network at fitness 0.5 (we call it the *central network*), (b) all the other networks are composed by *one* single individual (we call them the *peripheral networks*) and (c) all the peripheral networks are connected with the central one by one mutation. The proof of these properties is omitted for lack of space but it has been extensively discussed in the thesis.

The importance of this theoretical characterization of $\mathcal{L}^{\{\text{XOR}\,;\,\text{NOT}\}}$ is mainly related to the analysis of the sampling effects, as explained below.

Sampling Methodology

Sampling fitness landscapes is often a necessary task because the number of solutions in the search space is often huge.During the thesis, a difference equation that counts the number of solutions in boolean search spaces has been derived. It shows that the size of the search space grows very quickly (actually more quickly than a^x) as the maximum tree-depth parameter increases. Moreover, also the terminals and functions used have a great influence on the size of the search space.

Generating suitable samples to perform analyses of boolean (and, in particular, even parity) fitness landscapes is a difficult task. These landscapes are characterised by the presence of a large majority of individuals with fitness equal to 0.5. Well-known sampling techniques do not offer a useful "view" of the fitness landscapes because they do not consider some important parts of the landscape (e.g. individuals at fitness different from 0.5) or they do not "respect" the real structure of the originating landscapes (e.g. they generate many isolated individuals). Thus, we propose a new sampling process that combines a well-known technique with other steps designed to reproduce the "local" structure of the landscape. The aim of this process is the generation of samples containing trees of many (possibly all the) different fitness values and forming connected neutral networks, if possible. This pro-

cess is composed by three steps: we have called them *modified Metropolis*, *vertical expansion* and *horizontal expansion*. Modified Metropolis generates a sample C of individuals. The vertical expansion tries to enrich C by adding to it some *non-neutral* neighbours of its individuals. Finally, the horizontal expansion tries to enrich C by enlarging each "small" incomplete neutral network[1].

Main results and future work

During this work, we have performed two different kind of analysis: the former has investigated fitness landscapes of small size, in order to be able to exhaustively generate all their individuals, the latter has studied samples (obtained by our sampling methodology) of larger landscapes. Exhaustive analysis has considered $\mathcal{L}_{(2,3)}^{\{\text{XOR}\,;\,\text{NOT}\}}$ and $\mathcal{L}_{(2,3)}^{\{\text{NAND}\}}$ landscapes, i.e. the two landscapes associated to the even-2 parity problem and with a maximum tree-depth equal to 3. A first sampling analysis has again involved even-2 parity landscapes with "taller" trees (maximum depth equal to 7). By comparing these results with the previous ones and with the theoretical characterization of $\mathcal{L}^{\{\text{XOR}\,;\,\text{NOT}\}}$, we were able to empirically evaluate the sampling bias. As a results, our sampling methodology seems to behave well on these landscapes. The next step has been the study of samples of the landscapes associated with the even-4 parity problem. Also the maximum trees-depth was increased to 8, in order to include at least one optimal solution.

All the analyses have considered the same set of *network measures*, i.e. measures calculated on a neutral network. This set contains some "traditional" measures (such as network size, network fitness, etc.) and some new characteristics related to neutrality that were specifically defined. In particular, we have defined: (1) the average neutrality ratio of a neutral network, which quantifies the amount of possible neutral mutations of its individuals; (2) the average Δ-fitness of a neutral network, which quantifies the average fitness gain achieved by mutating its individuals; (3) the non-improvable solutions ratio, which quantifies the amount of solutions that cannot generate better offspring in a neutral network; (4) the "non-worsenable" solutions ratio, which quantifies the amount of solutions that cannot generate worse offspring in a neutral network.

Studying measure (1), has been observed that networks with bad fitness values seem to be "more neutral" than networks with good fitness values if {NAND} is used as the set of operators, while this is not the case if {XOR ; NOT}

[1] A neutral network is called incomplete when the sample does not contain all its individuals.

is used. Studying measures (2), (3) and (4), has been observed that it is unlikely to improve fitness mutating individuals of neutral networks with good fitness values if {NAND} is used, which is not the case if {XOR ; NOT} is used. These results may help explain why the even parity problem is easy for GP if {XOR ; NOT} is used and hard if {NAND} is used. These results hold both for the "small" fitness landscapes that have been studied exhaustively and for the "large" fitness landscapes that have been sampled using the new methodology here introduced. This fact may suggest the suitability of the sampling methodology for the boolean parity problems.

Since the presented techniques are general and can be used for any GP program space, future work includes extending this kind of study to other problems and possibly defining new measures of problem hardness based on neutrality.

References

[1] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[2] C. M. Reidys and P. F. Stadler. Neutrality in fitness landscapes. *Applied Mathematics and Computation*, 117(2–3):321–350, 2001.

[3] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. In *Proc. R. Soc. London B.*, volume 255, pages 279–284, 1994.

[4] L. Vanneschi, M. Tomassini, P. Collard, and M. Clergue. Fitness distance correlation in structural mutation genetic programming. In Ryan Conor *et al.*, editor, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 455–464, Essex, 14-16 Apr. 2003. Springer-Verlag.