Applied Soft Computing 67 (2018) 558–566

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Diverse partner selection with brood recombination in genetic programming

Muhammad Waqar Aslam[a,*], Zhechen Zhu[b], Asoke Kumar Nandi[b,c]

[a] Department of Computer Systems Engineering, Mirpur University of Science and Technology (MUST), Mirpur 10250, AJK, Pakistan
[b] Department of Electronic & Computer Engineering, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[c] Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35, Jyväskylä FI-40014, Finland

ARTICLE INFO

Article history:
Received 18 October 2017
Received in revised form 22 February 2018
Accepted 20 March 2018
Available online 27 March 2018

Keywords:
Genetic programming
Diversity
Partner selection
Brood recombination

ABSTRACT

The ultimate goal of learning algorithms is to find the best solution from a search space without testing each and every solution available in the search space. During the evolution process new solutions (children) are produced from existing solutions (parents), where new solutions are expected to be better than existing solutions. This paper presents a new parent selection method for the crossover operation in genetic programming. The idea is to promote crossover between two behaviourally (phenotype) diverse parents such that the probability of children being better than their parents increases. The relative phenotype strengths and weaknesses of pairs of parents are exploited to find out if their crossover is beneficial or not (diverse partner selection (DPS)). Based on the probable improvement in children compared to their parents, crossover is either allowed or disallowed. The parents qualifying for crossover through this process are expected to produce much better children and are allowed to produce more children than normal parents through brood recombination (BR). BR helps to explore the search space around diverse parents much more efficiently. Experimental results from different benchmarking problems demonstrate that the proposed method (DPS with BR) improves the performance of genetic programming significantly.

1. Introduction

Genetic programming (GP) has gained much attention in recent years because of its ability to produce human competitive solutions [1–3]. The flexibility in choosing various parameters of the algorithm and the ability to produce human interpretable solutions have made it superior to other learning algorithms and it has been applied quite frequently recently for solving real world problems. In [5] Liang, Zhang and Brownie used GP in image processing for figure ground segmentation. They demonstrated that GP based method was very successful in automatic segmentation of a variety of images. Maua and Grbac [6] used GP for software defect prediction for imbalanced datasets. Enrquez-Zrate et al. [7] used GP for predicting fuel flow and exhaust gas temperature of a gas turbine. A combination of GP and neuro-fuzzy methods was used for accurate prediction of suspended sediments in [8]. In another research GP was used for automatic scheduling of parallel unrelated machines [9]. Despite numerous advantages offered by GP there are some inherent issues in GP which limit its performance when applied to difficult tasks and there have been a variety of techniques proposed to improve its efficiency and performance. Elola et al. [10] exploited the predictive importance of intermediate solutions during a GP evolution and used it successfully for improving convergence rate. Genetic Programming in combination with fuzzy inference system was used for improving classification abilities of GP [11]. Ojha, Abraham and Snasel used diversity index measure for maintaining diversity in a multi objective GP and showed that it improved the efficiency of the proposed algorithm [13]. In [14], statistical GP using correlation based crossover and mutation was used to explore the search space more efficiently in less time. One of the main issues in GP is the premature convergence towards local optimum. It is widely accepted that the main reason for such convergence is the decrease in diversity in a fairly fit population of individuals as the population evolves. Ursem showed that a diversity guided evolutionary algorithm saved substantial amount of fitness evaluations and improved the performance remarkably [15]. Huang and Chen [16] used diversity pooling scheme for improving the convergence rate and showed that their proposed method reduced premature convergence rate. Various diversity measures were tested on standard problems for finding a relation between diversity and fitness in [17]. Important diversity measures for improving the search pro-

* Corresponding author.
E-mail addresses: waqar.cse@must.edu.pk (M.W. Aslam), zhechen.zhu@brunel.ac.uk (Z. Zhu), asoke.nandi@brunel.ac.uk (A.K. Nandi).

https://doi.org/10.1016/j.asoc.2018.03.035
1568-4946/

cess were identified in this research. A diversity rewarding fitness function was used in [18] to avoid local optimum. Similar programs were replaced with randomly generated ones in [19,20] and it was shown that it minimised the occurrence of premature convergence. Fitness and solution diversity was increased and it was found that the high fitness solution were found more quickly in [21].

One of the main factors causing loss of diversity is the selection pressure imposed on the algorithm [22]. A high selection pressure would quickly fill up the population with fitter individuals because they have better chances of survival. These fitter individuals most likely would be similar to each other either in their structure or behaviour, resulting in a loss of diversity across the population. As a consequence, the algorithm may struggle to escape from a local optimum. Diversity in any population will help to avoid such local optimum.

Maintaining diversity is particularly important if the fitness landscape have many peaks and valleys resulting in many local optimums. In such a scenario, a more diverse population will increase the likelihood of GP in finding the optimal peak or relocating individuals in a dynamic landscape. However, despite such advantages, blind promotion of diversity may result in a loss of information previously gathered by the algorithm.

In this paper pairwise diversity is used to promote crossover between two diverse individuals. The method proposed in this study is a significant extension of the methods presented in [23,24]. In [23] Day and Nandi presented the idea of binary string representing strengths and weaknesses of an individual. Using the binary strings of different individuals comparative partner selection (CPS) was used to encourage crossover between two individuals having high pairwise diversity. In one of our earlier research [24] we suggested an improvement in CPS and used shared strengths of individuals in addition to high pairwise diversity while selecting a right partner for crossover. In this study we demonstrate that just having a high pairwise diversity or sharing same strengths may not necessarily lead to the optimum solution and within pairwise diversity we have introduced the concepts of *good diversity* and *bad diversity* to find the preferred partners for crossover. The novel concept of good and bad diversity helps to follow exploration and exploitation strategy, popular in machine learning algorithms. The preferred partners found using the proposed method are allowed to produce more children through brood recombination. Brood recombination helps to explore the search space close to diverse partners more efficiently which ultimately guides the process towards the optimum solution.

The paper is organised as follows: the diversity measures and role of diversity in guiding evolution is discussed in Section 2. Evaluation of binary string is discussed in Section 3. The comparative partner selection technique and its flaws are discussed in Section 4. The proposed method is discussed in Section 5. Experiments and results are discussed in Section 6, while conclusions are drawn in Section 7.

2. Genetic programming and diversity

In this section first we highlight different measures used in literature to calculate diversity and then the promotion of diversity in finding the optimum solution is discussed.

2.1. Diversity measures

Broadly, diversity represents the level and type of variety in a population of individuals. This variety could be in the form of structure (genotype diversity) or behaviour (phenotype diversity) of individuals. The genotype diversity measures the similarity between actual structures of individuals. The most common method for measuring this diversity is the edit distance which simply counts the number of node additions and deletions required to make any two individuals identical and gives a general idea about structural resemblance of any two individuals [25].

The phenotype diversity on the other hand represents the behaviour of individuals. The most common method for evaluating this diversity is to find the distribution of fitness values in a population [26]. In this method the fitness of an individual is calculated by dividing its standard fitness by the number of individuals sharing similar fitness values. Mckay introduced the idea of evaluating individual training cases for calculating fitness [27]. This idea was later used in [23] to promote crossover between two diverse individuals.

Some other measures based on both genotype and phenotype or occasionally using the combination of the two have also been proposed [28]. Ryan [29] presented the idea of evolving two different populations together with different fitness criteria where crossover was only allowed between two different populations. An increase in diversity with reduction in code bloat was reported. Genetic lineage strategies have also been proposed [30,31] where parents for crossover are selected based on their genetic lineage to promote diversity. Teller and Veloso proposed an internal reinforcement method for GP using neural learning [32]. The neural learning was used to update specific parts of GP programs as a function programs performance. Quang Uy et al. investigated the role of semantic locality of crossover in GP [33]. They defined a semantic distance metric for defining new crossover operators to improve semantic locality. They reported substantial advantage using semantic locality. Xie and Zhang investigated the selection of optimal crossover point [34] and reported that good crossover events have a strong preference for roots and bottoms of parent program trees.

Based on the above discussion it can be concluded that while various strategies have been proposed for promoting diversity, none have been widely adopted. Each strategy has its own advantages and disadvantages, and is suited for particular applications. The genotype diversity has the advantage that it is quite objective and two solutions having exactly same structure will have exactly same behaviour. This measure, however, does not consider the behavioural differences of individuals and only implies that the actual structures of individuals are not identical. While this method is widely adopted because of its simplicity, any two individuals categorised as structurally unique by this method may behave similarly because of the presence of introns and symmetric functions. On the other hand phenotype diversity is more subjective and it is mainly calculated using fitness values. A phenotype diversity based just on fitness does not give insight into actual behaviour of the population for individual cases of a test problem. Two solutions performing well on two different cases of the same problem will be categorised as similar.

Therefore, there is a need to find a measure which is inexpensive, informative and can relate diversity to fitness improvement. In this study a phenotype based strategy for controlling diversity and guiding the search towards the optimum solution is proposed.

2.2. Promotion of diversity

Most diversity measures give an overall indication of diversity of a population but they do not give any information about the quality of individuals present in the population. It is possible to maintain a high level of diversity without getting any benefit in the quality of individuals. Moreover, the level of diversity does not indicate whether a population has suffered sub-optimal convergence or not, because diversity does not always promote optimal convergence. While in some cases it has been shown that diversity avoids early or premature convergence, it is not beneficial all the time