Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 24 (2013) 121 – 125

Procedia Computer Science

17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013

Bias Reduction of Probabilistic Prototype Tree Based Estimation of Distribution Genetic Programming in Predicting Arthritis Prevalence

Kangil Kim[a,*], Hanggjun Cho[b]

[a] *Seoul National University*
Structural Complexity Laboratory
Building 302, Gwanangno 599
Seoul 151–744, Korea
[b] *Seoul National University*
Structural Complexity Laboratory
Building 302, Gwanangno 599
Seoul 151–744, Korea

Abstract

Estimation of Distribution Algorithms in Genetic Programming (EDA-GP) are algorithms applying stochastic model learning to genetic programming. In spite of various potential benefits, probabilistic prototype tree (PPT) based EDA-GPs recently appeared to have a critical problem of losing diversity easily. As an alternative learning method to reduce the effect, likelihood weighting (LW) was proposed and its results were positive to improve EDA-GP performance. In this paper, we aim to provide more generalised verification results to confirm the effects of LW. We investigate performance of PPT-based EDA-GP in a large scale problem predicting arthritis using medical data.

Selection and peer-review under responsibility of the Program Committee of IES2013

Keywords: Estimation of Distribution Algorithms, Genetic Programming, Probabilistic Prototype Tree, Arthritis, Disease Prediction, Bias

1. Introduction

In evolutionary computation, there has been a surge of research to apply stochastic models into genetic algorithms (GA), called Estimation of Distribution Algorithms (EDA)[?]. It was successful in improving performance and provided useful tools to incorporate knowledge. This approach is also applied to genetic programming (GP), called Estimation of Distribution Algorithms in Genetic Programming (EDA-GP)[?].

In EDA-GPs, however, complex data representation such as trees complicates building a stochastic model. This complexity led to developments of a variety of model representations so far. Recently, it was reported that a main representation, Probabilistic Prototype Tree (PPT)[?], may have critical problems by imposed bias[?]. According to this work, a model on PPT learns probability distribution significantly different from the model at previous generation in

* Corresponding author. Tel.: +82-2-860-6179
E-mail address: kangil.kim.01@gmail.com

Selection and peer-review under responsibility of the Program Committee of IES2013
doi:10.1016/j.procs.2013.10.034

process of using learning and sampling only. This bias is repeated at every generation and increases exponentially in terms of size of the model, which results in convergence to wrong solutions. To reduce this negative effect, modified mechanisms were proposed such as likelihood weighting. The method reduced the effect successfully and improved the performance of the EDA-GP systems.

In this paper, we aim to provide more practical evidence to support this argument, because the preliminary work is tested in well-known, but small-scale benchmark problems. For this verification, we will introduce a practical problem to predict a disease, arthritis, using medical data and then investigate the difference after reducing bias of a PPT-based EDA-GP. Arthritis is a well-known and commonly occuring disease, so predicting it is expected to have high impact for warning people and urging to manage their lifestyles and therapies. This problem is defined over partial data of National Health And Nutrition Examination Survey (NHANES) conducted by Centers of Disease Control and Prevention, USA[?].

The rest of this paper is organised as follows. Section 2 explains related background knowledge of EDA-GP, its bias problem and arthritis prediction. Section 3 describes how we define the prediction problem. Section 4 shows how to transform it to an optimisation problem to apply EDA-GP systems, and their configuration for experiments. Section 5 illustrates experiment results and analysis and we will make conclusions at Section 6.

2. Related Works

2.1. Estimation of Distribution Algorithms in Genetic Programming

Estimation of Distribution Algorithms in Genetic Programming is an algorithm of applying a stochastic model into genetic programming[?]. We may see this algorithm as an iterative learning algorithm or an evolutionary search algorithm guided by a stochastic model. The basic process of this algorithm is equal to EDA, which is repetition of generating samples from a model, selecting more fit samples, and updating the model. Detailed process is shown in algorithm 1. Replacing crossover or mutation with model learning and sampling process, this algorithm can increase

Algorithm 1 Basic Process of EDA and EDA-GP

```
Initialising Population – generating individuals from an initial probability model
while not (condition for termination) do
    Fitness Evaluation – evaluate fitness of individuals through given fitness function
    Selection – select the best individuals in terms of the fitness
    Update – modify probability or structure of the stochastic model using the best individuals
    Sampling – generate new individuals from probability stored in the model
end while
```

performance of conventional evolutionary algorithms. Moreover, its explicit representation has been expected to be benefitial to analysis of intermediate behaviour of the algorithms and guiding search incorporating prior knowledge.

Compared to EDAs adapting well-founded model representation such as Bayesian networks, EDA-GP is complication in learning a model because its data such as trees includes more information to control than linear chromosomes. Main distinctive features of the data are the number of variables and structural constraints imposed to symbols[?]. In EDA-GP, to control the aspects, a variety of model representations have been proposed such as PPT or Stochastic Grammars[?]. PPT is a major representation introduced in Probabilistic Incremental Program Evolution (PIPE) used in various EDA-GP models[????]. It may be regarded as a simple Bayesian network without any dependency between variables, but PPT has more restriction on selecting values by the constraints. An example of PPT is depicted in figure 1. PPT is basically a tree, but each node is a random variable representating a multinomial distribution over symbols. When we observe a tree, it is overlapped onto the PPT and its symbols are counted as samples for matching variables. From many observed trees, the PPT model learns new probability distribution for each variable. Generating samples is usually ancestral sampling started from the root variable of PPT. In this paper, we will mainly discuss about this PPT-based models.