Universal Journal of Computer Science and Engineering Technology
1 (2), 122-126, Nov. 2010.
© 2010 UniCSE, ISSN: 2219-2158

A New Strategy for Gene Expression Programming and Its Applications in Function Mining

Yongqiang ZHANG
The information and electricity-engineering institute,
Hebei University of Engineering, Handan, P.R.China,
yqzhang@hebeu.edu.cn

Jing XIAO
The information and electricity-engineering institute,
Hebei University of Engineering, Handan, P.R.China,
xiaojing8785@163.com.cn

Abstract: **Population diversity is one of the most important factors that influence the convergence speed and evolution efficiency of gene expression programming (GEP) algorithm. In this paper, the population diversity strategy of GEP (GEP-PDS) is presented, inheriting the advantage of superior population producing strategy and various population strategy, to increase population average fitness and decrease generations, to make the population maintain diversification throughout the evolutionary process and avoid "premature" and to ensure the convergence ability and evolution efficiency. The simulation experiments show that GEP-PDS can increase the population average fitness by 10% in function mining, and decrease the generations for convergence to the optimal solution by 30% or more compared with other improved GEP.**

Keywords: Gene Expression Programming; GEP-PDS; Function Mining; Local Optimum

I. INTRODUCTION

Ferreira developed the basic Gene Expression Programming (GEP) [1] algorithm in 2001, which has inherited the advantages of the traditional genetic algorithm (GA) and genetic programming (GP). It has been applied to many fields [2~4] for its simple coding, fast convergence speed and strong ability of solution problems. GEP creates more diverse genetic operators than GA, and in a certain extent overcomes the shortage of local optimum. But the "premature" phenomenon still exists, and the performance of the algorithm unstable in practical problems. To solve this problem, a lot of improvement strategies have been proposed. The transgenic idea of biotechnology [5] has been imported to function mining based on GEP by Tang Changjie etc., including gene injection, transgenic process and evolution intervention, to guide evolution towards the direction people expected to some extent through the integration of natural selection and artificial selection. The superior population producing strategy [3] has been presented by Hu Jianjun, to produce population with high individual fitness and genetic diversity and significantly improve the success rate and the efficiency of evolution. GEP has been combined with the clonal selection algorithm of immune system in data mining [6] by Vasileios K. Karakasis and Andreas Stafylopatis, to optimize the selection operator of GEP, so as to improve the accuracy of data prediction and evolution efficiency.

In this paper the population diversity strategy of GEP (GEP-PDS) is presented, inheriting the advantage of superior population producing strategy [9] and various population strategy [3], to make the population maintain diversification throughout the evolutionary process and avoid "premature" to ensure the convergence ability and evolution efficiency.

II. MAJOR CONCEPTS OF GEP

Unlike other genetic algorithms, GEP innovatively takes chromosome as the entity bearing genetic information, expression tree (ET) as the information expression form. It is pivotal that chromosome and ET are interconvertible so exactly that complicate formulas could be coded. Terminals of GEP provide the ending structures of chromosomes, and functions act as the intermediate structure. Ferreira applied GEP in function mining and devised two fitness computation functions [1] --- fitness based on absolute error, and on relative error. Have evaluated the evolution results of each generation fitness function, we retain individuals with high fitness and make them have a better chance of reproduction. So the cycle does not terminate until an optimal solution or certain generations appear.

III. GEP-PDS

Population diversity and selection pressure are two vital factors affecting evolution process of genetic algorithm [8]. Similarly, immature convergence phenomenon of GEP is also due to the destroyed population diversity and the lost motive power of population evolution. To ensure global convergence of the algorithm, a feasible solution is to maintain the population diversity and avoid the effective genes [9] losing.

A. *The Superior Population Producing Strategy*

To express correctly superior population producing strategy, this paper introduces some formalized descriptions as

Definition 1(GEP mode) GEP model is a 7-tuple. *GEP=<Np,Ng,h,Fs,Ts,M,F>*, where *Np* is the population size, *Ng* is the number of genomes contained in a chromosome, *h* is the head length, *Fs* is the function set, *Ts* is the terminal set, *M* is the range of selection and *F* is the linking function.

Definition 2 Suppose m sample points, M is the range of selection, the sample set SampleSet={<*s*，*z*>| *s* is the parameters set，*z* is the target values set}. If a chromosome with positive fitness meets| *vi-zi* |⩽*kM*, the chromosome is an elite individual. Where *vi* is the chromosome value set at the parameters set *si*, *zi* is the corresponding target value of *si* and *k* is a non-negative coefficients.

When *k=0*, *vi=zi* is legal and the elite individual is the finding objective function. It is equivalent to randomized method for search objective function. Set a threshold of producing times for every *k* in Elite Strategy [10]. When the random producing times reaches that threshold, if the elite individual still has not been produced, the value of *k* would increase gradually until the elite have been produced. The threshold can be set as time. If the elite has not been produced within the time, increase *k*. When *M* is set improperly, two extreme cases would happen. One is producing elite individuals difficultly, the other is too easy. In the second case the selected individual is certainly not true elite. Though the individual fitness may be high, it can not properly assess the quality status of the individual. Settings M related to reference [1].

Definition 3 Suppose GEP mode *GEP=<Np,Ng,h,Fs, Ts,M,F>*, C_j is the *j*th chromosome of population *p*, C_{ji} is the ith gene of chromosome C_j, of which $0≦j<p$, $0≤i<(h+t)$, *t* is the tail length:

(1) G_{ji} and G_{ki} are called alleles;

(2) If gene *G*∈(*Fs U Ts*),for any *j*,there is $G≠G_{ji}$,it is claimed that *G*is the lost genome on locus *i* of population *p*;

(3) If C_j = C_k, claimed C_j and C_k are repeated individuals of population p.

Having produced elite individuals, other initial population individuals are generated randomly, or through mutation of the elite individuals. In the population, keep the elite unchanged, and distribute genes uniformly in gene space (Fig.1).

```
For (test the composition of every locus){
 If (the proportion of one gene at the locus above average)
   The gene mutate to one with the lowest proportion; }
While (repeated individuals exist){
 Mutate the repeated one;
 For (test the composition of every locus) {
  If (the proportion of one gene at the locus above average)
      the gene mutate to one with the lowest proportion; } }
```

Figure 1. Distribute genes uniformly

We adopt the superior population producing strategy to optimize the initial population of GEP, to rich genetic diversity and raise individual fitness. Such population is superior.

B. *The various population strategy*

When GEP evolves to the late stage, gene convergence effect of population happens, population diversity declines, therefore results in lower efficiency. Reference [3] has proved, in the sense of probability, the evolutionary time-consuming of every generation has a positive relationship with population size. Therefore, in terms of evolutionary time, it will reduce evolution efficiency when the size is large.

Definition 4 Assume *gi=<ti,fi>* is the state of generation *gi*, of which *ti* is the time evolution to *gi*, *fi* is the maximum population fitness of *gi*. For the two evolutionary states *gi* and *gk*, suppose *i<k*. If *fi=fk*, called *gk-gi* is the stagnation generations, and *tk-ti* is the responding time. If *fi=fk* and *fi<fk+1*, said that *gk-gi* is the maximum stagnation generations, *tk-ti* is the maximum stagnation time, and the population starts to evolve again at the generation*k+1*.

Let's explain the idea of the various population strategy. In GEP, the initial population size set to *Np*, when the stagnation time reaches the maximum, if the population size has not reached the maximum population size, population size would double per evolution generation; if reached, the *Np* individuals with the worst fitness of the current population would been replaced; after evolution to the maximum stagnation generations, the population would start to evolve at the next generation and the size decreases to *Np*. Continue executing program until the optimal solution has been found or achieving the maximum generations.

C. GEP-PDS Description

Input: GEP=<*Np, Ng, h, Fs, Ts, M, F*>, fitness evaluation formula, SampleSet={<*s*，*z*>| *s* is the parameters set，*z* is the target values set }, controls parameters of GEP (maximum times of producing individuals *N*, maximum scale of population *n**Np*, maximum stagnation generations g_{top}, maximum generations G_{limit}, probability of replication, mutation and recombination etc.)

Output: optimal or approximate optimal solution

Step 1: set controls parameters of GEP;

Step 2: initialize population by superior population producing strategy;

Step 3: operate GEP(GEP mode)(Fig. 2);

Step 4: iteration end, output the optimal solution.

```
While (generations<Glimit and not evolve to an optimal solution)
      {express each chromosome of the population;
             execute program;
                        evaluate fitness;
                        execute genetic operations;
                        change population scale
                            {If (stagnation generations ==gtop)
                               {If (scale<n*Np) double scale;
                               Else replace the whole individuals}
                             If (start evolution) scale decrease to Np; }
       generations++; }
```

Figure 2. Operate GEP

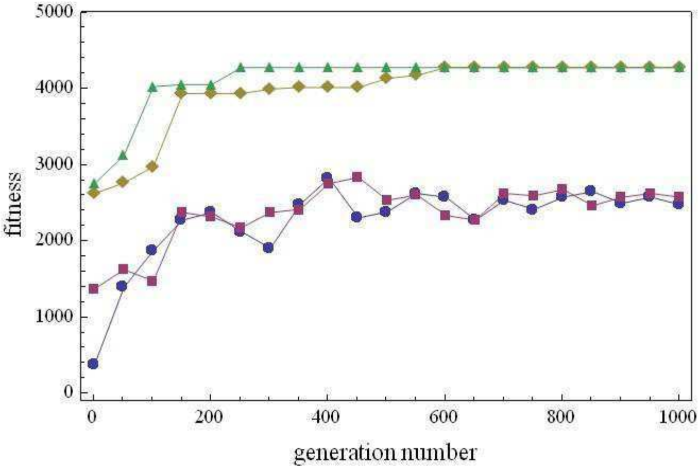
IV. EXPERIMENT AND PERFORMANCE ANALYSIS

The experiment is carried out in the VC 6.0, using C++ programming to imitate function mining process with GEP. The experimental data is imported into Mathematica 7.0 to complete simulation.

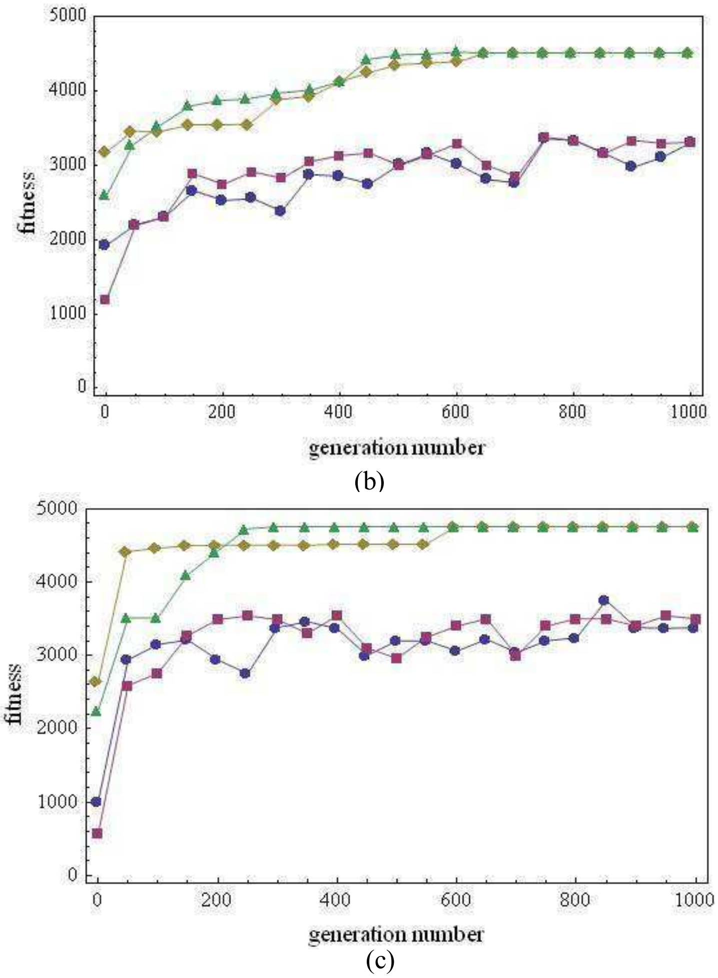
The mining processes of three commonly used standard functions are simulated in experiments. A unary quadratic function $F1 = \pi a^2$, a unary higher-order function $F2 = 5a^4 + 4a^3 + 3a^2 + 2a + 1$, and a complex trigonometric function $F3 = \frac{\sin(a)\cos(b)}{\sqrt{e^c}} + \tan(d-e)$. The functions above are from http://www.gene-expression-programming.com/GepBook/Chapter4/Section1/SS2.htm。In the experiment, the training data sets of these three functions are generated firstly. 50 independent variables of F1 and F2 are produced randomly from -50.0 to 50.0, while F3 from 0 to 1. Take them as parameter values of the training set. Target values of the set are the corresponding function values. Repeat 100 mining experiments for each data set, the average of final results are obtained as the final result. The parameters of GEP in the test are set as shown in Table 1. In the table, Q, E, S, T, C from the functions set separately means "Square root", "Exponential", "Sine", "Tangent", "Cosine".

TABLE I. PARAMETERS OF GEP IN EXPERIMENTS

	F1	*F2*	*F3*
Population Scale	40	40	40
Number of Genes	3	3	3
Function Set	+ - * /	+ - * /	+ - * / Q E S T C
Terminal Set	a	a	a b c d e
Head Length	6	6	6
maximum generations	1000	1000	1000
Linking Function	+	+	+
Selection Range	100	100	100
Mutation Rate	0.044	0.044	0.044
Recombination Rate(one-R,two-R,gene-R)	0.044	0.044	0.044
Gene Transposition Rate(IS,RIS)	0.3	0.3	0.3



(a)



(b)

(c)

Figure 3. Comparison the maximum fitness and average fitness between GEP and GEP-PDS during mining F1(a), F2(b), F3(c). ▲ stands for the maximum fitness with GEP-PDS, ◆ the maximum fitness with GEP, ■ the average fitness with GEP-PDS, ● the average fitness with GEP

As shown in Figure 3, compared with the traditional GEP, GEP-PDS produces an excellent initial population, the average fitness during evolution increased by about 10%, while generations of convergence to the optimal solution reduce about 30%. It is easy to say that the convergence to the optimal solution by GEP-PDS is significantly faster than GEP, and the evolution efficiency of GEP-PDS is higher. Although the superior population producing strategy would increase the time-consuming of initial population, the population has a high diversity, making high search efficiency, without losing its convergence rate. Simultaneously, the introduction of various population strategy at the late stage in GEP could avoid the occurrence of genetic convergence effect, injection of new genes to improve genetic diversity, thus shorten the GEP evolution stagnation time and improve efficiency.

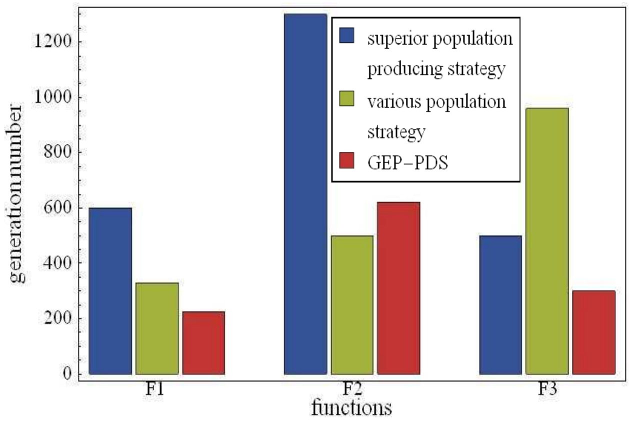


Figure 4. Comparison the average convergence generations under different strategies

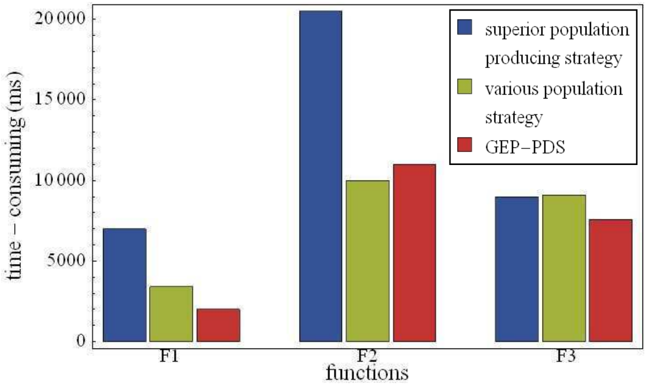


Figure 5. Comparison the average time-consuming of function mining under different strategies

Reference [7] has proved the initial population under superior population producing strategy is obviously superior to other ways. Reference [3] has stated the various population strategy precedes traditional GEP. Therefore only comparisons among GEP-PDS and superior population producing strategy and various population strategy have been done in the experiments. Figure 4 shows that GEP-PDS evolution generations is superior to the other two strategies. From figure 5，it is clear that time-consuming with GEP-PDS is the best at mining function F1 and F3.

Experiments show that, the performance of GEP-PDS precedes the traditional GEP algorithm, and superior population producing strategy and various population strategy.

V. CONCLUSIONS

Like other genetic algorithms, population diversity is one of the vital factors affecting evolution. To accelerate the efficiency and avoid local optimal, GEP-PDS has been presented in this paper to preserve high fitness and population diversity. Finally, by simulating the mining process of three standard functions, the evolution rate and convergence efficiency are compared under GEP-PDS and other strategies. The simulation experiments show that GEP-PDS can increase the population average fitness by 10%, and decrease the generations for convergence to the optimal solution by 30% or more compared with other improved GEP, so as to improve overall GEP evolutionary efficiency.

ACKNOWLEDGMENT

The authors thank the National Natural Science Foundation of Hebei Fund (F2010001040) for supporting this project.

REFERENCES

[1] Ferreira Candida. Gene expression programming: a new adaptive algorithm for solving problems [J]. Complex Systems, 13(2): 87-129(2001).

[2] Ferreira Candida. Discovery of the Boolean Functions to the Best Density-Classification Rules Using Gene Expression Programming [C]. Proceedings of the 4thEuropean Conference on Genetic Programming, Berlin: Springer-Verlag, 51-60(2002).

[3] Jianjun HU, Changjie TANG, Jing PENG, et al. VPS-GEP: skipping from local optimization fast algorithm [J]. Journal of Sichuan University (Engineering Science Edition), 39(1): 128-133(2007).

[4] Satchidananda Dehuri, Sung-Bae Cho. Multi-objective Classification Rule Mining Using Gene Expression Programming [C]. Third 2008 International Conference on Convergence and Hybrid Information Technology, ICCIT.2008.27: 754-760.

[5] Tang Changjie, Chen Yu, Zhang Huan, et al. Discover formulas based on GEP with trans-gene [J]. Journal of Computer Applications, 2007, 27(10): 2358-2360.

[6] Vasileios K. Karakasis, Andreas Stafylopatis. Efficient Evolution of Accurate Classification Rules Using a Combination of Gene Expression Programming and Clonal Selection. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 2008,12(6): 662~678.

[7] Jianjun HU, Xiaoyun WU. Superior Population Producing Strategy in Gene Expression Programming [J]. Journal of Chinese Computer Systems, 30(8): 1660-1662(2009).

[8] Whitley D. The GENITOR algorithm and selection pressure: Why rank based allocation reproduction trials is best[C]. Proc of the 3rd International Conference on Genetic Algorithm. Los Altos: Morgan Kaufmann Publishers(1989).

[9] Dong WANG, Xiangbin WU. Protect strategy for effectual gene block of genetic algorithm[J]. Application Research of Computers, 25(5)(2008).

[10] Jianjun HU, Hong PENG. Elitism-Producing Strategy in Gene Expression Programming [J]. Journal of South China University of Technology (Natural Science Edition), 37(1): 102-105(2009).

AUTHORS PROFILE

Yongqiang ZHANG (1966-), professor of Hebei University of Engineering who is studying on software reliability engineering and so on.

Jing XIAO (1987-), candidate for master degree who is studying on the GEP Algorithm and the software reliability modeling.