An Improved Selection of GEP Based on CPCSC-DSC Approach

Li Wu [1, a], Yonghong Yu [2, b] and Zhou Zhou [2, c]

[1] School of Finance and Public Management, Anhui University of Finance & Economics, Bengbu 233030, China;

[2] School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu 233030, China.

[a]bbwuli@163.com, [b]ac120107@163.com, [c]1057578619@qq.com

Abstract

This paper mainly discusses the effects of duplicated chromosomes on genetic diversity, population fitness level, evolutionary efficiency, and evolutionary search retention during the selection operation of gene expression programming(GEP). It introduced the concepts of strict and non-strict implicit duplicated chromosome and proposed a novel method named creating population copy for the second choice-deleting same chromosome(CPCSC_DSC) to remove the duplicated or implicit duplicated chromosomes according to the population fitness, which can ensure population diversity, evolutionary efficiency and can avoid premature convergence. The experiment result shows that the CPCSC_DSC accelerates evolution, increases the fitness of optimal chromosome compared to the result not using CPCSC_DSC.

Keywords

Gene expression programming, Selection operation, Creating population copy for the second choice, Deleting same chromosome.

1. Introduction

After GEP[1,2,3,4] completes a certain generation of evolutionary operation, new individuals are added to the population, and the number of individuals is expanded. These individuals have different fitness, and only a part of them are added to the next generation. GEP obtains fitness of individual through the established fitness function, and takes the fitness as the criteria of selection, the higher the fitness of individuals is, the greater the probability of being choosed to add to the next generation is. Each individual in a population has a selection probability, which is generally allocated according to fitness or rank according to fitness. The proposal according to the ratio of fitness is named Monte Carlo method, and the probability of being selected depends on the proportion of individual fitness probability. The proposal according to fitness ranking sorts the target values of the population, and the fitness values are determined according to the ordinal position. The commonly used selection methods are roulette selection, random traversal sampling, elitist selection and tournament selection. The introduction of elitism strategy in roulette can guarantee the global convergence of the algorithm, but has slow evoulation speed. The tournament strategy and elite selection strategy are easy to fall into local optimum though they have fast convergence speed.

The population initialization process of the basic GEP[2,4] algorithm is completely randomized, and the selection scale of the roulette is constant, and the diversity of the GEP population is relatively simple and unevenly distributed. Reference [5] suggests that elite individuals produce strategy EPS and genetic spatial distribution of GSBS initial population strategy, by producing individuals with higher fitness and improving the genetic diversity of initial population, the evolutionary efficiency is improved, and convergence is avoided as a local optimal. Reference[6] proposed OBS algorithm which makes use of evolutionary process to eliminate close relatives and propagate distant species to enhance population diversity. Reference [7] proposed DAIP algorithm which uses the weighted diversity measure to quantify the population diversity and to maximize the population diversity. Reference [8] proposed a GDM-GEP algorithm to measure the diversity of the population by the

individual fitness variance, by designing an adaptive mutation operator, the algorithm changes the mutation rate with the population diversity, and maintains the stability of the population and the preservation of the excellent individuals.

The evolutional chromosomes same as the parent chromosome will be added to the population, and will participate in the selection operator after fitness evaluation. If the duplicated chromosome individual has a higher fitness, the probability of being selected as the parent chromosome of next generation will be increased according to the roulette selection method and the random traversal sampling method. In tournament selection and elitist strategy, as long as their fitness values reach the top of the ranking, they will be selected for the next generation.

Duplicated chromosomes as parent chromosomes are added to the next population may result in three problems, such as: (1) The diversity of excellent chromosomes in descendant populations was reduced. (2) The population fitness level and evolution efficiency of next generation before genetic operation were reduced. (3) Duplicated chromosomes assimilate other individuals malignantly, and cause evolutionary search retention. To sum up, repeated chromosomes have certain stability. Once a certain number of chromosomes are formed in the population, they are difficult to be destroyed by the genetic factor and they expand rapidly.

In order to ensure population diversity, evolutionary efficiency and avoiding premature convergence, it is necessary to eliminate duplicated chromosomes in the genetic process. This paper introduces the concepts of strict implicit duplicated chromosome and non-strict implicit duplicated chromosome, and proposes a novel approach to delete duplicated chromosomes by creating population copy for the second choice-deleting same chromosome(CPCSC-DSC). It adopts the fitness value as the judgment rule to determinate whether an individual is a duplicated chromosome or a strict implicit duplicated chromosome, and uses the CPCSC-DSC method to supplement the number of individuals to improve the GEP selection process.

2. Preliminary

Terminals and functions constitute the primitive elements of GEP. The terminals represent to the constant, input or no parameter functions in the mathematical expression, and correspond to the leaf nodes of the expression tree. The functions represent operators of espical application, components of programm and middle symbols of system, they correspond to the non leaf node of expression tree.

The expression tree(ET) is used as the phenotype of GEP individuals to represent an expression visually. In order to facilitate computer genetic manipulation, the nonlinear structure of the expression tree is translated into a linear structure which Ferreira called K-expression as chromosome of GEP, corresponding to the genotype of GEP individuals.

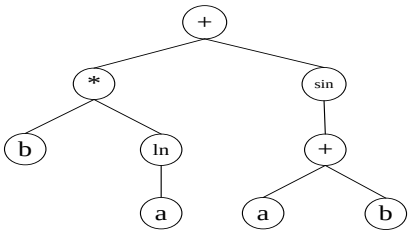
GEP genes are composed of a head and a tail[2,4]. The head contains symbols that represent both functions and terminals, whereas the tail contains only terminals. Therefore two different alphabets occur at different regions within a gene. For each problem, the length of the head h is chosen, whereas the length of the tail t is a function of h and the number of arguments of the function with the most arguments n, and is evaluated by the equation:

$$t = h^* (n\text{-}1) + 1 \tag{1}$$

Consider a gene composed of $\{+,*,ln,sin, a, b\}$. In this case $n = 2$. For instance, for $h = 7$ and $t = 8$, the length of the gene is 7+8=15. One such gene is shown below：

$$\begin{array}{c} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4 \\ +\ *\ \sin\ b\ \ln\ +\ a\ \ a\ \ b\ \ b\ \ a\ a\ b\ a\ b \end{array} \tag{2}$$

and it codes for the following ET:



The corresponding meaning of each chromosome member in the initial population created by the GEP algorithm is different. The genetic operator acts on the genetic manipulation of the population, and the higher the degree of fitness of the resulting offspring individuals, the greater the probability of being selected. GEP adopts linear equal-length coding, so long as the length of the gene is constant and the tail consists only of terminals, a legitimate offspring chromosome can be obtained. Therefore, the GEP genetic operator can be designed more simply and flexibly. The basic genetic operators of GEP mainly include selection, mutation, transposition of insertion sequence elements, root transposition, gene transposition, one-point recombination, two-point recombination, and gene recombination.

Fitness is an indicator of biological ability to measure the ability of a species to adapt to the environment, it can also apply to the GEP algorithm. The fitness function is used to calculate the individual fitness to evaluate the pros and cons. By evaluating the chromosomes, expression trees can be obtained, then the objective function values are calculated from the phenotypes, and finally, according to the conversion rules, the phenotypic (individual) fitness values can be obtained using the objective function values. Ferreira proposed two fitness evaluation models: one is that the fitness f_j of an individual chromosome j is calculated based on the absolute error, and the other is calculated based on the relative error. Let T be a dataset containing n samples, T_i is the i-th input dataset, v_i is the observed value of T_i, and $\hat{v}_i$ is the estimated value. the fitness equations are:

$$f_j = \sum_{i=1}^{n} (K - |v_i - \hat{v}_i|) \tag{3}$$

$$f_j = \sum_{i=1}^{n} (K - |\frac{v_i - \hat{v}_i}{v_i}|) \tag{4}$$

Through the constraint of the fitness function, the population retains the high-adaptation chromosomes during the evolution process. When the individual chromosomes do not have relative or absolute errors in the sample estimates, the accuracy is maximized and the fitness is maximized.

3. Creating Population Copy for the Second Choice-Deleting Same Chromosome Approach

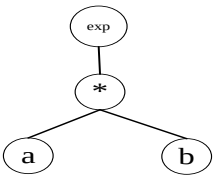
Chromosomal mutation and recombination positions are random in evolution, and the lengths involved in mutation and recombination are random, the chromosomes may not change during genetic manipulation, that is to say, duplicated chromosomes are produced, it occurs at the following cases: (1) The elements are identical after mutation. (2) The substring elements involved in the inverted string are symmetric about the middle elements. (3) Substrings have the same type and order of the elements of string being inserted during the insertion string operation, and the type and order of the elements of string being inserted are the same with the original elements of position of subsequent shift. (4) The type and order of the two string elements involved in recombination are the same. When the length of randomly generated substrings is relatively short, the probability of these situations will increase greatly.

The concept of implicit duplicated chromosomes is that the genotypes of two or more chromosomes in the population are different, but their corresponding phenotype is the same, and can eventually be converted into the same expression tree.

Definition 1. Strict Implicit Duplicated Chromosome: Two or more gene expressions have the same ORF and different noncoding regions, and can traverse to the same expression tree directly.

Equation 5 shows that two gene expressions are strict implicit duplicated chromosomes, their ORF is the same (the underlined part), and the corresponding expression tree is:

$$\begin{array}{cccccccccccc} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 \\ \underline{\exp} & * & a & b & b & b & b & b & a & b & a & \\ \underline{\exp} & * & a & b & b & b & b & b & a & b & a & \end{array} \tag{5}$$



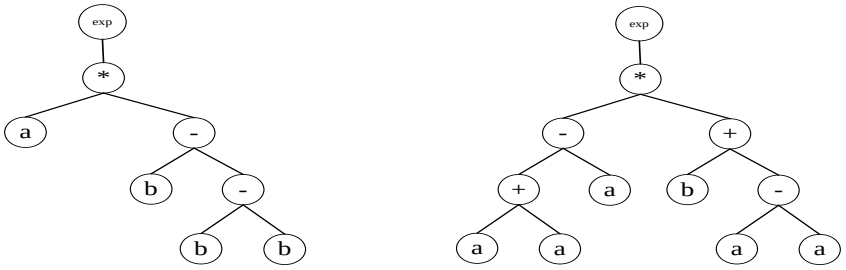
Strict implicit duplicated chromosomes entering the parent chromosome constitute descendant population may cause the following three problems：reduce genetic diversity of efficient coding segments，reduce the population fitness level and evolution efficiency of next generation before genetic operation, generate more new and strict implicit repeat chromosomes and even generate duplicated chromosomes.

The strict implicit duplicated chromosomes also have a certain stability, once there is a certain number of them in the population, they are difficult to be destroyed by the genetic factor, and they expands and even produces repeated individuals under the action of the recombination operator. In order to ensure gene fragment diversity, evolutionary efficiency and prevent the generation of duplicated chromosomes in the effective coding segment, strict implicit repeat chromosomes need to be eliminated in the genetic process.

Definition 2. Non-Strict Implicit Duplicated Chromosome: Two or more gene expressions have different ORF and noncoding regions, and the expression trees obtained by traversing directly are also different, but both the expression tree can be simplified to the same tree structure.

Equation 6 shows that two gene expressions are not strict implicit duplicate chromosomes, their ORF different (underlined parts), and their corresponding expressions trees are different as follows, but can be simplified as the tree structure shown in mentioned above.

$$\begin{array}{cccccccccccccccc} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \underline{\exp} & * & a & - & b & - & b & b & a & b & b & a & b & b & a & b & b & \\ \underline{\exp} & * & - & + & + & a & b & - & a & a & a & a & b & b & b & a & a & \end{array} \tag{6}$$



The coding regions and non-coding regions of non-strict duplicated chromosomes are different. Therefore, as a parent chromosome, it will not destroy the diversity of population gene fragments. The non-strict duplicated chromosomes have different direct phenotypes and same indirect phenotypes, and the number of existence of non-strict duplicated chromosomes is very small compared to duplicated chromosomes and strict implicit duplicated chromosomes in the population. In the process of evolution, the non-strict implicit duplicated chromosomes are easily destroyed by the genetic operator, and the probability of producing duplicated chromosomes is very low, and they often convert into total different chromosomes. Therefore, the indirect phenotype of the non-strict implicit duplicated chromosomes is unstable. To sum up, the non-strict implicit duplicated

chromosomes can be selected as the parent chromosomes of the next generation. Due to the small number, even the excellent non-strict implicit chromosomes are not selected to be the parent chromosomes, the cost is very small.

Duplicated chromosomes and strict implicit duplicated chromosomes are highly stable in the evolution process. To prevent the creation of a large number of cloned individuals from damaging the ecology of the population, it is necessary to remove the duplicated chromosomes and strict implicit duplicated chromosomes when the selection operator acts on the population. If the remove process occurs before the operation of the selection operator, the number of duplicated chromosomes and strict implicit duplicated chromosomes is reduced to 1 in the candidate population, the possibility of the genotype being selected decreases greatly. In the actual process of population genetics, individuals being generated repeatedly also reflect that the genotype has a strong survivability and convergence trend, so it is still necessary to maintain the probability of the g genotype being selected. The remove process only prevents the malignant generation of implicit duplicated individuals in next generation, but does not decrease the probability of the genotype being selected in current generation. Therefore, the remove process should occur after the selection operator operation, the objects being removed should be the parent chromosome group being selected in current generation instead of the previous generation population before the selection operator operation.

During the remove process, the basis for judging duplicated chromosomes and strict implicit duplicated chromosomes is as follows: (1) duplicated chromosomes: The gene expression is the same as the gene expression of another individual in the population. (2) strict implicit duplicated chromosomes: The ORF segment is the same as the ORF segment of another individual in the population. Since the ORF segment is a part of the gene expression, we can remove duplicated chromosome individuals according the second rule. That is, when individuals with the same ORF segments are removed, all duplicated chromosomes and strict implicit duplicated chromosomes will also be removed. To determine whether the ORF segments are the same, it is necessary to scan the gene expression to determine the length of the ORF first, then to scan the specific elements of the ORF segment from head to tail, and compare them with other existing ORF segment elements. If the elements are same, it means that the individual is a duplicated chromosome or a strict implicit duplicated chromosome.

The judgment algorithm by comparing ORFS mentioned above is less efficient. This paper adopts the fitness value as the judgment rule to determinate whether an individual is a duplicated chromosome or a strict implicit duplicated chromosome: (1) Fitness as a criterion for evaluating the quality of an individual has been calculated with the birth of a chromosome individual in genetic operator, no additional calculation is needed. (2) The fitness value is generally expressed as double data type, it has a high precision and can be used as the judgment rule to determinate whether an individual is a duplicated chromosome or a strict implicit duplicated chromosome. (3) The number of non strict implicit duplicated chromosomes in the population is extremely low. It is a small probability event that the parent chromosomes which are not selected as a descendant are deleted, and the cost is cost-effective in terms of algorithmic efficiency. The steps of algorithm of DSC by comparing fitness values are follows:

Algorithm1. Deleting same chromosomes by comparing fitness values

Input : List of chromosome populations(chromosomesList)

Output: List of chromosome populations(chromosomesList)

```
1. chromosomeFitness=1.0000;  //define and initialize the fitness of chromosome
2. chromosomesList.sort();
3. for (t=0;t<length(chromosomesList);t++)
4.    if(chromosomesList[t]==chromsomeFitness) then
5.       romove chromosomesList[t] from chromosomesList;
6.       t—;
7.    else
```

8. chromosomeFitness=chromosomesList[t];
9. end if
10. end for
11. return chromosomesList;

After deleting duplicated individuals, the individual number of parent chromosomes, which are selected as the next generation, may be reduced, and need to be supplemented to the number of individuals before deletion. We propose a novel method CPCSC to improve the GEP selection process. The improved selection process is shown in Fig. 1 :

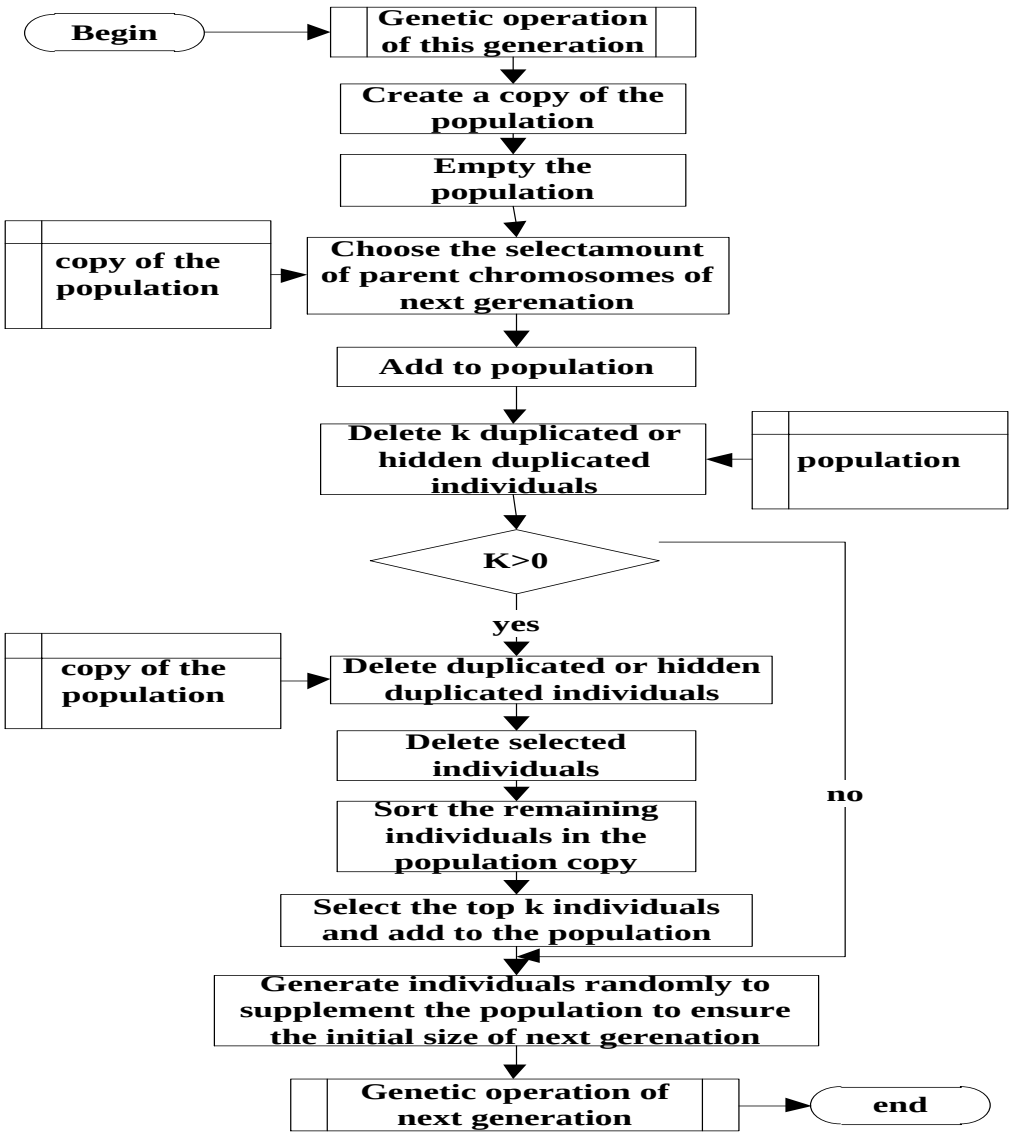


Fig.1 Selection flow of introducing CPCSC_DSC

The specific steps are: first, create the copy of population. Second, empty the population. Third, the selection operator chooses the parent chromosomes of next generation from the copy population and add them to the population, fourth, delete hidden duplicated individuals. Fifth, if the amount of population decrease after deleting, the duplicated individual and selected individuals are removed from the copy population, and the remaining dominant individuals are added to the population, and the number of the dominant individuals added to the population is the same as the number of duplicated individuals deleted. Finally, generate chromosomes randomly, and expand the parental chromosomal population to the required size.

4. Experiment and Analysis

A part of the Iris data test of UCI data set is used to prove the validity of the improved GEP algorithm compared to the standard GEP. The data contain three flower samples, setosa, versicolor and virginica. Each sample contains 4 attributes, including sepal length, sepal width, petal length and petal width. The calyx length of the setosa sample was selected as the dependent variable y, and the other attributes were used as independent variables, which were expressed as a,b and c respectively. Setosa sample data, as shown in Table 1:

Table 1. Iris-Setosa Sample Dataset.

Num	Sepal length y	Sepal width a	Petal length b	Petal width c
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.2	3.5	1.5	0.2
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1

In order to prove the effect of CPCSC_DSC algorithm on evolution, the experiment was divided into two groups, the experimental group and the control group, both groups use the same data set in Table 1, and they are executed based on reverse polish expression_sack decoding method to evaluate gene expression and based on rough multiple linear regression initialization_adaptive correction constant to initialize constant, in addition, the experimental group adopts CPCSC_DSC algorithm to delete the duplicated or hidden duplicated inidividals during generic evolution process.

Set the size of population to 60, the head length of gene expression to 15, the probability of selection to 0.3, the probability of recombination to 0.75, and the mutation probability to 0.1. Comparing the optimal fitness of the two individuals obtained by the CPCSC_DSC and no CPCSC_DSC algorithm after the 2000 generation,the individual with higher fitness wins. Fig.2 shows the results:

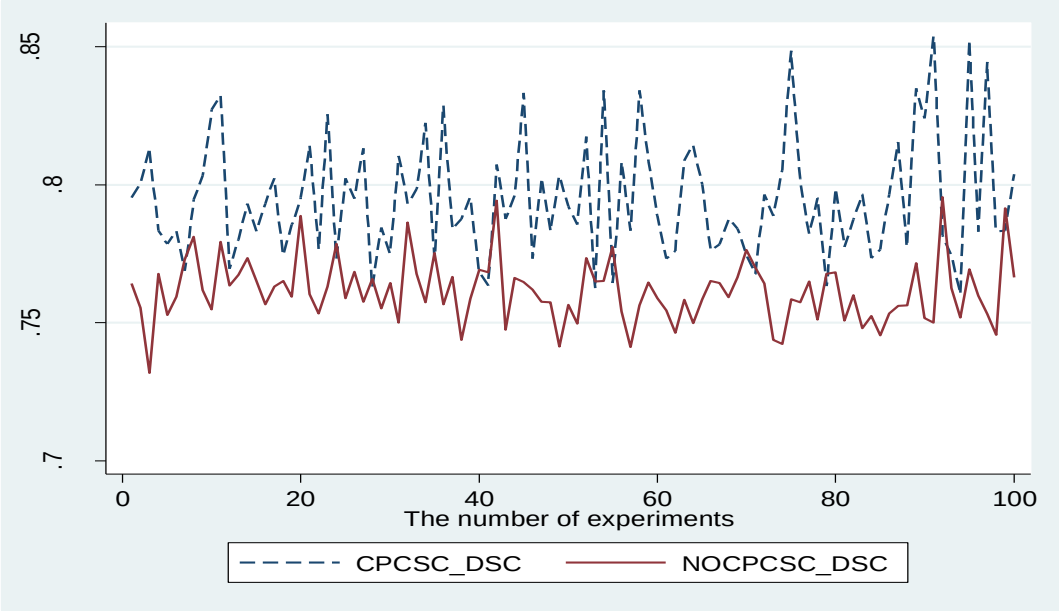


Fig.2 Fitness value of CPCSC_DSC and non CPCSC_DSC

Fig.2 shows that in the comparison of the 100 experimental results, the optimal fitness of the chromosomes obtained by the algorithm CPCSC_DSC is 88 times greater than that of the GEP without CPCSC_DSC (both group using reverse polish expression sack_ decoding method to

evaluate gene expression, and using rough multiple linear regression initialization adaptive correction constant to initialize constant) after the 2000 generation. The result shows that the CPCSC_DSC algorithm can accelerate the evolution speed. One group uses CPCSC_DSC algorithm and the other group does not use CPCSC_DSC, each group is executed 100 times respectively, and after 2000 generations of evolution, the average fitness of optimal chromosome is 0.80267, compared to 0.76258 of not using CPCSC_DSC, increased by 5.25721%.

5. Conclusion

The deletion of duplicated chromosomes in the process of selection operation of GEP plays an important role during the genetic process. An improved selection algorithm based on CPCSC_DSC dealing with the deletion of duplicated chromosomes in the process of selection is presented. The CPCSC_DSC introduced the conception of strict implicit and non-strict implicit duplicated chromosome, which can ensure the number of variable of individuals and can ensure population diversity, evolutionary efficiency and avoiding premature convergence. The experiment results show that the CPCSC_DSC accelerates evolution, increases the fitness of optimal chromosome compared to the result not using CPCSC_DSC.

References

[1] Information on https://www.gene-expression-programming.com.
[2] Ferreira Candida. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems, vol. 13(2001), p.87-129.
[3] Ferreira Candida. Genetic Representation and Genetic Neutrality in Gene Expression Programming. Advances in Complex Systems, vol. 4(2002), p.389-408.
[4] Ferreira Candida. *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence* (Springer- Verlarg, Germany 2006).
[5] Jianjun Hu: *The Researches of Key-techniques in Knowledge discovery System for TCM Pharmacology* (Ph.D., Sichuan University, China 2006).
[6] Yue Jiang, Changjie Tang and Mingxiu Zheng. Outbreeding Strategy with Dynamic Fitness in Gene Expression Programming. Journal of Sichuan University, vol. 139(2007), p.121-126. (In Chinese)
[7] Taiyong Li, Changjie Tang and Lei Duan. Adaptive Population Diversity Tuning Algorithm for Gene Expression Programming. Journal of University of Electronic Science and Technology of China, vol. 39(2010), p.279-283. (In Chinese)
[8] Bing Shan, Shihong Ni and Xiang Cha. GEP based on Population Diversity Measure by Variance of Individual's Fitness. Computer Engineering and Design, vol. 34(2013), p.3094-3098. (In Chinese)