New Selection Method to Improve the Population Diversity in Genetic Algorithms

K. Matsui
Dept. Electrical and Electronic Eng.,
Shizuoka Univ.
Hamamatsu, Shizuoka, 432-8561, JAPAN
E-mail: tekmats@ipc.shizuoka.ac.jp

ABSTRACT

In this paper, we present a new method of selection, in order to improve the population diversity in the genotype distribution, in Genetic Algorithms (GAs). The problem of maintaining of the population diversity is very important in designing the genetic operators, when GAs are applied to optimization problems. Therefore, we propose two types of new selection operators based on the correlations between individuals' genotypes, for improving the population diversity. The first operator is a new type of selection for reproduction, namely the correlative tournament selection. The second operator is a new type of selection for survival, namely correlative family-based selection. We applied our GA to two different problems: Royalroad problems, and Knapsack problems with non-stationary environments. We compared our method with the other representative GA model, and showed the effectiveness of the proposed GA models.

1. INTRODUCTION

Genetic Algorithms (GAs) [1] are stochastic methods of search and optimization based on the computational models of the biological evolutionary process. In designing and applying GAs to the practical tasks, the maintaining of the population diversity is an important problem. Also, generally, conventional GAs have tendencies to be losing the population diversity gradually in their searches, especially in the later phase of them. If the diversity could not be maintained sufficiently, the GAs would not find any acceptable solutions, and might trap on the premature convergence.

Therefore, we propose two types of new selection operators for the maintenance of the population diversity. These operators are based on the correlations between individuals' genotypes. The first one is an expansion of ordinary tournament selection, namely the correlative tournament selection. The second one is an modified version of ER (Elitist Recombination) model [2], namely the correlative family-based selection. The ordinary GAs involve two types of selection procedures [3]: the selection for reproduction (SFR) and the selection for survival (SFS), as shown in Fig.1. The SFR operation is the selection of parents for reproduction/crossover, whereas the SFS operation is the selection of individuals which survive in the next generation after the crossover.

We applied our GA which involves these new operators to some test beds: Royalroad [4] and non-stationary knapsack problem [7]. We compared our method with the conventional one in the following sections.

2. NEW SELECTION OPERATORS

2.1 Correlative Tournament Selection

First, we propose the correlative tournament selection (CTS), which is an expansion of the ordinary tournament selection [5]. We use the CTS as the SFR operator.

In the conventional GAs, two parents for a crossover operation are selected independently. However, for realizing a more effective crossover operation (which is the most significant feature in GAs), it is one of the promising ideas to introduce a kind of dependency, such as correlation, between two parents to be selected. In other words, by selecting a pair of parents which have higher possibilities to realize a effective crossover, the GA search will be enhanced. In this subsection, as the point of above, we propose a new selection based on the correlation between individuals, namely the CTS.

Ronald [6] proposed a method to select parents which are related to each other, namely the seduction method, as shown in Fig.2. In the seduction method, one of the two parents for crossover is selected by an ordinary way in conventional GAs. Then, the other is selected based on the first parent. However, the special case, in which the seduce function is *independent* of the first parent, is only discussed in [6], that is, this method is actually equivalent to the GA which involves two independent selection procedures.

The CTS, we propose in this paper, is similar to the seduction method. However, the selection of the second parent (i.e. the seduction) is closely depend on the first parent, and the GA will realize a new type of selection based on the correlation between individuals.

The basic idea of the CTS operator is that the offsprings born from the *similar* parents will be more similar to their parents than those from *distinct* parents.

The algorithm of the CTS is as follows:

1. Select one individual p_0, as the first parent, by the ordinary tournament selection.
2. Construct a candidate-set **P** for the second parent, by selecting N individuals p_i from the population at random.

$$\mathbf{P} = \{p_1, p_2, \dots, p_N\} \tag{1}$$

3. Evaluate correlative function $g(p_i)$ for all the individuals p_i in **P**,

0-7803-5731-0/99$10.00 ©1999 IEEE

$$g(p_i) = f(p_i) + c\ h(p_i, p_0), \qquad (2)$$

where $g(p_i)$ is the evaluation function for selection based on the correlation between candidate-individual p_i and the first parent p_0, $f(p_i)$ is the fitness of individual p_i, $h(p_i, p_0)$ is the genotype correlation between individuals p_i and p_0, and coefficient c is the parameter which determines the weight of the correlation $h(p_i, p_0)$.

4. Select one individual p_i, which has the highest value of $g(p_i)$ in **P**, as the second parent which is the mating partner of p_0.

In this paper, we use the Hamming distance as the correlation function $h(x, y)$, that is, a pair of similar individuals (x, y) have a low correlation value to each other. In our experiments, the size of **P** is set at $N = 2$.

2.2 Correlative Family-based Selection

We then propose the correlative family-based selection (CFS) as the second operator. This is an modified version of ER model [2] and a new type of selection for survival (SFS) operator, which selects individuals for the next generation after the reproduction.

In ER model, the SFS is applied on a unit, namely family, which consists of two parents and two offsprings reproduced by crossover of these parents. In the family, two individuals, which have the highest and the second-highest fitness value, survive for the next generation, whereas the other two individuals are killed. This procedure is applied for each family in the population.

In our CFS operation, we apply the following algorithm:

1. Select one individual x which has the highest fitness value in each family.
2. Calculate the Hamming distance d_i for other three individuals y_i (i = 1, 2, 3) in the family to the selected individual x,

$$d_i = H(y_i, x) \qquad (3)$$

where $H(y, x)$ is the Hamming distance of genotype between individuals y and x.

3. Select one individual y_k, which has the maximum d_k in $\{d_1, d_2, d_3\}$,

$$d_k = \max_i (d_i) \qquad (4)$$

4. Two individuals x and y_k survive for the next generation in this family, whereas other two individuals are killed.
5. Apply steps (1)-(4) for each family.

Using this CFS operation, we will be able to maintain the population diversity of genotype, as well as select individuals which have relatively higher fitness value.

3. EXPERIMENTS

3.1 Royalroad Problems

Our first experiment is the Royalroad problems (RR) [4]. The RR problems are a class of functions designed for the study of GA performance over time, especially on the building block interactions.

Some types of RR functions are proposed [4]. We use two types of RR functions: R1 and R2, as shown in Table 1. The RRs are defined as a list of schemata s_i, whose length is eight. The fitness of the genotype x is defined as follows:

$$f(x) = \sum_{i=1}^{N_s} c_i \delta_i(x) \qquad (5)$$

$$\delta_i(x) = \begin{cases} 1 & (\text{if } x \in s_i) \\ 0 & (\text{otherwise}) \end{cases} \qquad (6)$$

where coefficient c_i is a evaluation value of schema s_i. In Table 1, symbol "*" denotes a wild card (don't-care symbol). In R1, $N_s = 8$, whereas $N_s = 14$ in R2. The length of chromosome is 64. The optimum is 64 in R1, and 192 in R2, respectively (when all the bits in chromosome take a value "1").

The experimental setup is as follows: The population size is 50. Crossover rate is set at 1.0, and mutation rate is 0.02 per bit. We use the CTS operator as the SFR. The other operators consist of elitism, two-points crossover and bit-flip mutation. The parameter c in Eq.(5) takes a value from a set {1.0, 2.0, 4.0}, and we compare these three cases. We also compare our method with a conventional GA which involves a ordinary tournament selection. The difference between our GA and the conventional one is only the SFR operation: the CTS and the ordinary tournament selection.

The R1 problem: We tried 50 trials until the optimal value 64 was obtained. The results are shown in Table 2 and Fig. 3. Table 2 shows the mean/minimum/maximum generations to find the optimum in 50 trials. In this table, our methods are superior to the conventional GA by 20-30%, that is, our method could enhance the GA performance.

Fig. 3 shows the plot of the maximal fitness in this problem. Our method outperforms the conventional one over the fitness value 48 to the optimum, whereas there are less differences among four cases before reaching the fitness 48. The differences in the later phase of the searches are caused by the maintenance of the population diversity by our method.

The R2 Problem: Next, we also tried on the R2 problem in the same way. The results are shown in Table 3. Our method outperformed the conventional one, maximally by 27 %.

Computational Costs: We compared the computational costs between the proposed and the conventional method. We measured the total computational time of 20 trials of GA searches in both cases. All the trials are limited to 1000 generations. As the results, all the searching procedure were terminated within 130 sec. in the conventional method and 135 sec. in the proposed one, respectively. There is only a little difference among the both cases. All the simulations were performed on the Pentium II/233MHz. From these results, our method has another advantage of the low computational costs.

3.2 Non-stationary Knapsack Problems

The next experiment is the knapsack problems with non-stationary environments [7]. In adaptation to such environments, the maintenance of the diversity is quite essential, since the loss of the diversity reduces the ability of adaptation. So, this type of non-stationary problems are effective in testing the proposed method.

We use the following non-stationary knapsack problem: The number of items is set at 100. The weight and the value of each item are initialized in the range [1,100], and these values are changed within the range $[-\alpha,+\alpha]$ in every T evaluations. Parameter α is determined in this range at random. In our experiments, parameters T = 1000, 10000, 100000 and α = 20% are used.

The GA setups are the following: The population size is 50. Crossover rate is set at 1.0 and mutation rate is 0.02 per bit. We use the uniform crossover, bit-flip mutation and the two proposed selection operators.

We tried 20 trials with different seeds of random numbers. The results are shown in Figs. 4, 5, and 6. Fig. 4 illustrates the plots of the maximal fitness in the proposed GA and the SGA. In this figure, we found that our methods are superior to the conventional one in the non-stationary environments.

Fig.5 is the plot of the genotype variance. We used the measurement of the genotype variance v as following equation:

$$v = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{L}(\bar{x}_k - x_{ik})^2 \qquad (7)$$

where N is the population size, L is the length of chromosome, x_{ik} denotes the gene (0 or 1) at locus k in individual i, and $\bar{x}_k$ is the average value of the gene at locus k. This measurement is a variance of the distance between each individual and the average one. In Fig.5, we found that our method maintained the population diversity better than the conventional GA.

Fig. 6 shows the plot of fitness variance in this problem. Also, our proposed GA outperformed the conventional one.

From these results, our proposed method could improve the performance and the population diversity of GA. In comparison on the changing period T, these three cases have similar tendencies with each other. That is, our method has an advantage of the adaptation to various scale non-stationary-problems.

4. CONCLUSIONS

We proposed two types of selection operators, namely the correlative tournament selection and the correlative family-based selection, which are based on the correlation among individuals' genotypes, in order to maintain the population diversity.

We applied our method to the Royalroad and the non-stationary knapsack problems. Our results are superior to those of the conventional ones. Also, our method could improve the population diversity over the conventional GA.

Our method has another advantage of the computational costs. Thus, our method will be applied to other problems easily.

REFERENCES

[1] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

[2] D. Thierens, and D.E. Goldberg, "Elitist Recombination," *Proc. IEEE Conf. Evolutionary Computation*, pp. 508-512, 1994.

[3] H. Sato, et al., "A New Generation Alternation Model of Genetic Algorithms and Its Assessment," *J. of Japanese Society for Artificial Intelligence*, 12, 5, pp. 734-744, 1997 (*in Japanese*).

[4] M. Mitchell, et al., "When Will a Genetic Algorithm Outperform Hill Climbing?" Santa Fe Institute working paper 93-06-037, 1993.

[5] D.E. Goldberg and K. Deb, "A Comparative Analysis of Selection Schemes used in Genetic Algorithms," *Foundations of Genetic Algorithms*, 1991.

[6] E. Ronald, "When Selection Meets Seduction," *Proc. 6th Int'l Conf. Genetic Algorithms*, pp. 167-173, 1995.

[7] N. Mori, et al., "Adaptation to a Changing Environment by Means of the Thermodynamical Genetic Algorithm," *Proc. 4th Conf. Parallel Problem Solving from Nature*, pp. 513-522, 1996.

Table 1 Royalroad problem

i	s_i	c_i
1	11111111**	8
2	********11111111**	8
3	****************11111111**	8
4	************************11111111********************************	8
5	********************************11111111************************	8
6	**11111111****************	8
7	**11111111********	8
8	**11111111	8
9	1111111111111111**	16
10	****************1111111111111111********************************	16
11	********************************1111111111111111****************	16
12	**1111111111111111	16
13	11111111111111111111111111111111********************************	32
14	********************************11111111111111111111111111111111	32

Table 2 Simulation result in R1 problem

	mean	min.	max.
Conventional	511.2	180	1260
$c = 1.0$	372.8	160	740
$c = 2.0$	346.8	120	760
$c = 4.0$	413.2	100	900

Table 3 Simulation result in R2 problem

	mean	min.	max.
Conventional	496.4	140	1580
$c = 1.0$	410.8	60	840
$c = 2.0$	359.2	100	740
$c = 4.0$	398.4	140	800

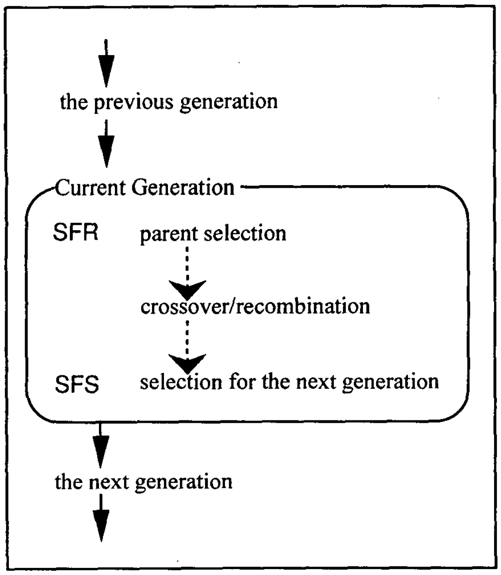

Fig.1. SFR/SFS operations

Conventional GA

```
loop {
    parent1 = select();
    parent2 = select();
    crossover and mutation;
}
```

Seduction method

```
loop {
    parent1 = select();
    parent2 = seduce(parent1);
    crossover and mutation;
}
```

Fig.2. Pseudocode of the conventional GA and the seduction method

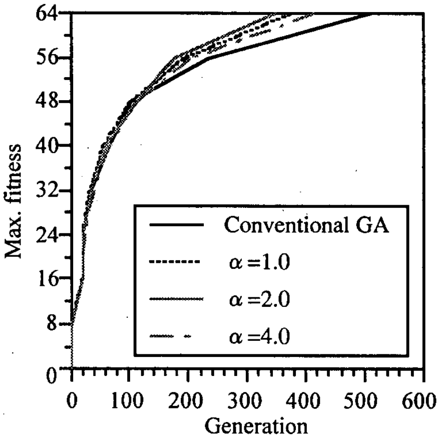

Fig.3 The plot of Maximal fitness in R1 problem

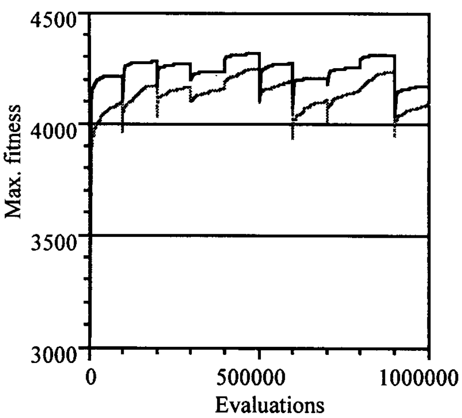

Fig. 4(c) The plot of maximal fitness at T=100000

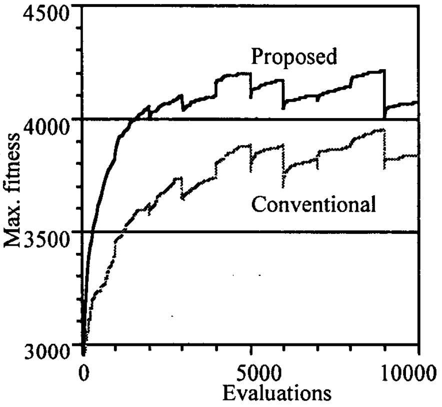

Fig. 4(a) The plot of maximal fitness at T=1000

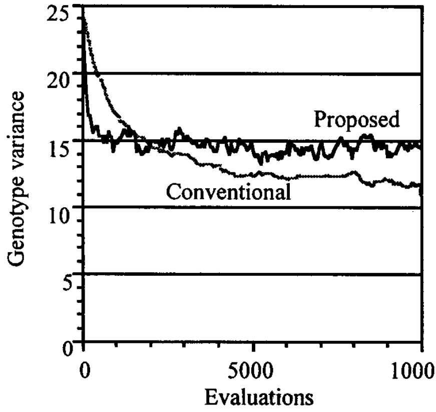

Fig.5(a) The plot of genotype variance at T=1000

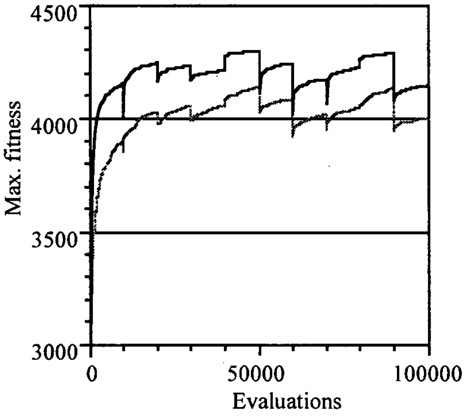

Fig. 4(b) The plot of maximal fitness at T=10000

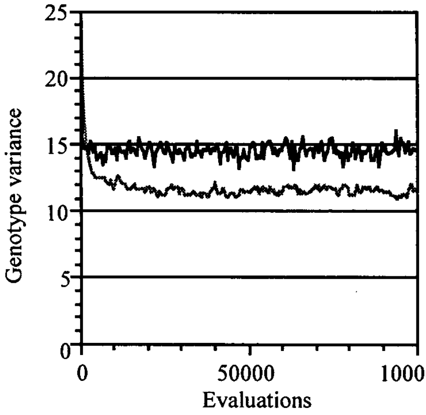

Fig.5(b) The plot of genotype variance at T=10000

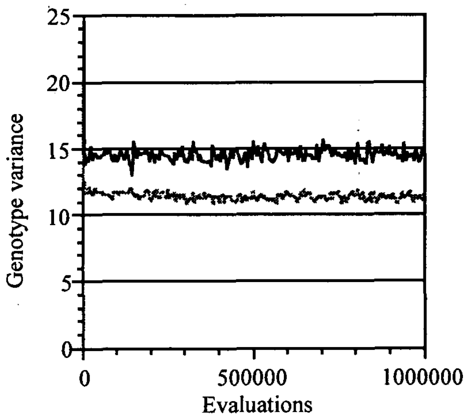

Fig.5(c) The plot of genotype variance at T=100000

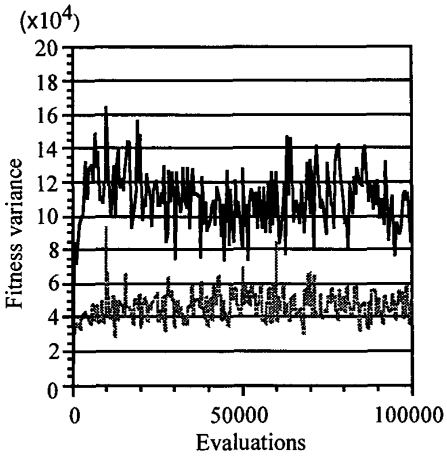

Fig.6(b) The plot of fitness variance at T=10000

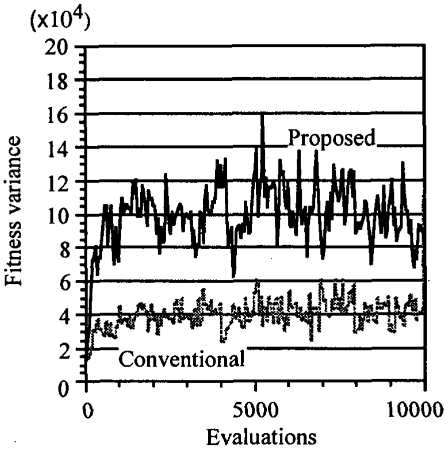

Fig.6(a) The plot of fitness variance at T=1000

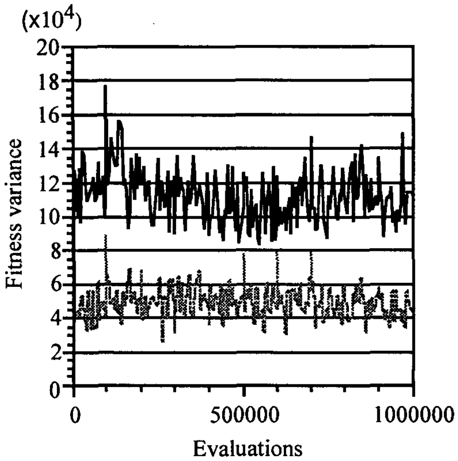

Fig.6(c) The plot of fitness variance at T=100000