Extraction of Landscape Information based on a Quality Control Approach and Its Applications to Mutation in GA

Mitsukuni Matayoshi
Okinawa International Univ.
901-2701, Okinawa, Japan
matayosi@okiu.ac.jp

Morikazu Nakamura
Univ. of the Ryukyus, Japan
morikazu@ie.u-ryukyu.ac.jp

Hayao Miyagi
Univ. of the Ryukyus
903-0213, Okinawa, Japan
miyagi@ie.u-ryukyu.ac.jp

ABSTRACT

We introduce an attraction hypothesis and repulsion hypothesis on combinations of genes and we characterize "gene/ locus pair" as a "Unique Inheritance" if the pair satisfies one of the hypotheses. We propose a method based on a statistical approach to extract a set of gene-locus pairs characterized as "Unique Inheritance", and also two new genetic operations, attraction mutation and repulsion mutation.

Categories and Subject Descriptors

G.2.1 [**Combinatorics**]: Combinatorial algorithms

General Terms

Algorithms

Keywords

Genetic Algorithms, Landscape Information, Quality Control method, fast 2-opt method, Linkage.

1. INTRODUCTION

GA is applied to various combinatorial optimization problems and obtains good results. Some researches use not only LS(Local Search) but also the landscape of search area to improve performance[2]. They statistically examined solutions of the Traveling Salesman Problem, and hit on the Big Valley Hypothesis[1]. They proved that some random TSPs have Big Valley Structure, and succeeded in getting good performance using its characteristics[1].

2. HYPOTHESIS AND INHERITANCE

We take note of Boese's "edge"[1]. We propose that the concept of an "edge" suggests the following three hypotheses for more than two genes to improve fitness.

Attraction hypothesis: Combinations of genes strongly attract each other(It could be considered a linkage of bilding block within a chromosome).

Repulsion Hypothesis: Combinations of genes strongly repulse each other.

Powerless Hypothesis: Combinations of powerless genes are compared with combination of attractive/repulsive genes.

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

We next provide new hypotheses for embracing the "combining power" of gene and genetic locus.

Unique Inheritance Hypothesis: There are combinations of gene and genetic locus which strongly influence fitness, when a gene takes "unique" position on chromosome.

Based on this hypothesis, we could treat the relation between genes, genetic locus and fitness simultaneously. We call gene and genetic locus "unique gene" and "unique genetic locus". And we call the relation of a unique gene and a unique genetic locus "unique inheritance".

2.1 Flow to unique inheritance extraction

An outline of how to extract unique inheritance is:
(1) Sort random initial solutions in order of highest fitness.
(2) Select high quality solutions.
(3) Extract unique inheritances by applying a control limit technique of quality control method.

2.1.1 Unique Genetic Locus

The next expression (1) is to consider a gene pair set. We prepare individuals sorted by fitness into a high quality solution set $S(\in natural\ number)$. C_i^r means one gene on genetic locus $r(1 \leq r \leq n)$ of the chromosome C_i in S. $n(\in natural\ number)$ is the length of chromosome $C_i(i \in S)$.

$$a_{jk} = \{(l, m) : l = C_i^j, m = C_i^k; j, k \in r, j \neq k, i \in S\} \quad (1)$$

a_{jk} gives gene pair set in which genetic locus j, k appears frequently. Next, we give $n \times n$ matrix A_{jk} consists of h_{jk}^{lm} is the number of elements(l, m) of a_{jk}. where, $j \neq k, h_{jk}^{l=m} = 0$.

$$A_{jk} = \begin{pmatrix} h_{jk}^{11} = 0 & \cdots & h_{jk}^{1n} \\ \vdots & \ddots & \vdots \\ h_{jk}^{n1} & \cdots & h_{jk}^{nn} \end{pmatrix}. \quad (2)$$

Example 1 Let us consider the high quality chromosomes $C_{1,2,3,4}(S = 4, n = 4)$ given by
$C_1 = (3, 2, 1, 4)$, $C_2 = (2, 4, 1, 3)$,
$C_3 = (2, 3, 1, 4)$, $C_4 = (4, 2, 1, 3)$,
$(e.g.\ C_1 = 3, 2, 1, 4. \rightarrow C_1^1 = 3, C_1^2 = 2, C_1^3 = 1, C_1^4 = 4)$,
Then $a_{12}, a_{13}, a_{24}, a_{34}$, and $A_{12}, A_{13}, A_{24}, A_{34}$ are
$a_{12} = \{(3, 2), (2, 4), (2, 3), (4, 2)\}$,
$a_{13} = \{(3, 1), (2, 1), (2, 1), (4, 1)\}$,
$a_{24} = \{(2, 4), (4, 3), (3, 4), (2, 3)\}$,
$a_{34} = \{(1, 4), (1, 3), (1, 4), (1, 3)\}$,

$$A_{12} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad A_{13} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$
$$A_{24} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad A_{34} = \begin{pmatrix} 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next, let's consider A_{jk}^{sum} and A_{jk}^{avg}. A_{jk}^{sum} is the total sum of the matrix A_{jk} represented in (2). A_{jk}^{avg} defines the average of A_{jk} which is obtained by dividing A_{jk}^{sum} by $n \times n$.

$$A_{jk}^{sum} = \sum_{q=1}^{n} \sum_{r=1}^{n} h_{jk}^{qr}, \quad A_{jk}^{avg} = A_{jk}^{sum}/n^2. \tag{3}$$

We select a control limit by the "$\boldsymbol{C}$ control chart method" for ease of calculations, and write it as:

$$A_{jk}^{UCL} = A_{jk}^{avg} + 3\sqrt{A_{jk}^{avg}}. \tag{4}$$

where UCL means the Upper Control Limit.

The calculation of A_{12}^{UCL} for **Example 1** is $A_{12}^{UCL} = 4/4^2 + 3 \cdot \sqrt{4/4^2} = 1.75$. In this case, three frequency "2" exceed A_{jk}^{UCL} which appear in A_{13} and A_{34}. In other words, gene pair$(2,1)$ on genetic locus $j = 1, k = 3$ which relation seems to be a unique inheritance, and gene pairs$(1,3)$ and $(1,4)$ on genetic locus $j = 3, k = 4$. On the other hand, there is no frequent gene pair in A_{12} and A_{24}.

In this paper, we focus on the relation between a gene pair and its genetic locus to extract what we call the unique genetic locus. Therefore our method requires obtaining the relation between gene and genetic locus quantitatively.

Let H consist of elements of A_{jk} exceeding the UCL.

$$H = \{H_{jk}^{lm} | h_{jk}^{lm} \geq A_{jk}^{UCL}, \forall i, j = 1, 2, \cdots, n\} \tag{5}$$

And let A_{jk}^{US} (US; Upper control limit Sum) defined as (6), be the total sum of A_{jk} elements which surpass the UCL.

$$A_{jk}^{US} = \sum\nolimits_{H_{jk}^{lm} \in H} h_{jk}^{lm} \tag{6}$$

We may consider A_{jk}^{US} to be "**connection**" as a criterion of intensity of loci j and k. For **Example 1** , $A_{12}^{US} = A_{24}^{US} = 0, A_{13}^{US} = 2, A_{34}^{US} = 4$. Thus "connection=2" appears at genetic loci "1" and "3", "connection=4" appear to genetic loci "3" and "4", but A_{12}^{US} and A_{24}^{US} do not have any it.

2.1.2 Extraction of unique genetic locus

Now, let's consider $n \times n$ matrix U composed of A_{jk}^{US}.

$$U = \begin{pmatrix} A_{11}^{US} = 0 & \cdots & A_{1n}^{US} \\ \vdots & \ddots & \vdots \\ A_{n1}^{US} & \cdots & A_{nn}^{US} \end{pmatrix} \tag{7}$$

where $A_{kl} = A_{lk};\ k, l \in n,\ k \neq l, A^{US} = 0 (k = l)$.

An element is a candidate for a unique genetic locus if its value is greater than 1. However, we know that a lot of candidates exist in many cases. Therefore we try to narrow the candidates down to the appropriate ones by use of the Control Limit again. U^{sum} and U^{avg} are defined as:

$$U^{sum} = \sum^{n} \sum^{n} A_{jk}^{US}, \quad U^{avg} = U^{sum}/n^2. \tag{8}$$

The Control Limit U^{UCL} is

$$U^{UCL} = U^{avg} + 3\sqrt{U^{avg}}. \tag{9}$$

U^{UCL} works as a threshold for selecting unique genetic loci. If V is the set of A_{jk}^{US} surpass at U^{UCL}, then

$$V = \{A_{jk}^{US} | A_{jk}^{US} \geq U^{UCL}\}. \tag{10}$$

is the set of unique genetic locus. Then for **Example 1** :

$$U = \begin{pmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 \\ 2 & 2 & 0 & 4 \\ 0 & 0 & 4 & 0 \end{pmatrix}, \quad \begin{array}{l} U^{sum} = 16, \\ U^{avg} = 16/16 = 1, \\ U^{UCL} = 1 + 3\sqrt{1} = 4, \end{array}$$

$V = \{A_{3,4}^{UCL}, A_{4,3}^{UCL}\} = \{(3,4),(4,3)\}$.

In the above case, $V = \{(3,4),(4,3)\}$ are extracted as unique genetic loci by recognizing the "connection" of genetic loci "3" and "4". At this time, note that genetic locus pair (1,3) has been screened out. If we have four chromosomes $(C_{1,\ldots,4})$, we are able to consider relations between gene "1" and genes "3,4" on the genetic loci "3" and "4". We may regard this as a stronger connection than between genes "1" and genes "2,3,4" on the genetic loci "1" and "3". Our method may be an alternative way to calculate linkage.

2.2 Target, Improved mutations and Results

The quadratic assignment problem (QAP) is employed as a target problem for evaluation of our proposed method.

Total population of GA: 50, ratio of mutations among the population: 30%, rate of mutations among the chromosome length: 30%, crossover rate: 50%, and limitation time: 180sec. LS is the fast 2-opt method. We used an escape operation from an undesirable convergence, as done in [2].

We devise two improved mutation strategies for considering genetic locus, which are Attraction mutation(**A**) and Repulsion mutation(**R**). **A** executes mutation at all loci except the "unique genetic locus" which is regards as strongly "connected(linked)" to improve fitness value. **R** executes mutation at the only "unique genetic locus" which is not regarded as strongly linked. We use simple mutation(**S**) for comparison with **A** and **R**. The symbols in Table 1 are the results of a non-parametric Wilcoxon rank-sum test. For brevity, tests marked by "*" required longer search times to find optimal solution, but "+" did not in comparison with **S**. Figure 1 is the landscapes of bur26a represented by U.

Table 1: Results of Three Mutations

Tests	S sec.	A sec.	R sec.	IA	IR
bur26a	4.06	3.90	5.08*	3	-25
bur26b	5.88	4.83	7.82**	17	-33
chr15c	3.08	3.68	1.70++	-19	44
chr18a	10.82	12.02	4.72++	-11	56
chr20a	58.44	99.96**	54.37	-71	6
kra30b	64.04	114.35**	70.47	-78	-10
lipa30b	1.81	1.98*	1.64+	-9	9
tai20b	1.68	1.37+	1.98*	18	-17

++,** :p<0.01/2. +,* :p<0.05/2.
IA=(S-A)*100/S. IR=(S-R)*100/S.

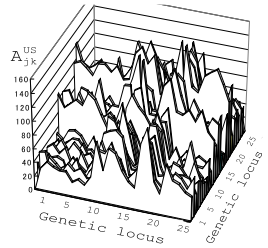

Figure 1: Landscape of bur26a (Matrix U)

3. CONCLUSIONS

The purpose of this paper is to verify four hypotheses by testing the effectiveness of two mutation schemes based upon them. Success is measured as an improvement of either approach over simple mutation. The action of attraction mutation and repulsion mutation on "unique genetic locus(loci)" are mutually different, and therefore we consider that the results do not negate our four hypotheses.

4. REFERENCES

[1] K. D. Boese and et al. New adaptive multistart techniques for combinatorial global optimizations. *Operations Research Letters*, 16(2):101–113, 1994.

[2] P. Merz and et al. Fitness landscape analysis and memetic algorithms for the quadratic assignment problem. *IEEE Trans. on EC*, 4(4):337–352, 2000.