社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－MPS－47 (16)
2003／12／12

適応度差分により分類された個体の分布に基づくGAの遺伝子座依存関係モデルの構築

辻 美和子† 棟朝 雅晴†† 赤間 清††

遺伝的アルゴリズム (GA) においては，互いに依存関係があり同一のビルディングブロックを構成する変数をあらかじめ同定することで，効率的な探索が可能となる．変数間の依存関係を調査する手法としては，値の摂動による適応度の差分を利用する手法や，有望な部分個体群に存在するストリングの持つ値の分布を調査する手法が提案されている．本論文では，これらの両者の特長をあわせもつ手法として，値の摂動による適応度の差分から分類された部分個体群の分布を調査する手法を提案する．提案手法は，より少ない計算量で，正確な依存関係を同定することが可能である．

Modeling dependency from distributions of strings classified according to fitness change

Miwako Tsuji†, Masaharu Munetomo†† and Kiyoshi Akama††

Genetic Algorithms(GAs) can perform effectively by identifying a set of loci tightly linked and strongly interdepended to form a same building block. Various methods are already proposed to detect such dependencies. Some of them investigate the fitness changes from the perturbation of gene value and some others estimate the distribution of strings in promising sub population. In this paper, we propose a new method combining both of them, which detects dependencies through the estimation of the distribution of strings classified according to fitness change. The proposed method can detect dependencies accurately by a little more than $O(l)$ fitness evaluations.

1. はじめに

GA の探索性能を解析したビルディングブロック仮説によれば，GA は問題を部分問題に分割し，各部分問題における適応度の高い部分解 (ビルディングブロック) を組み合わせることで全体の最適解を得るアルゴリズムである．しかし，一点交叉あるいは多点交叉を用いた単純 GA の場合，この処理は変数のストリング上での位置に大きく依存する．同一の部分問題に属する変数がストリング上で離れて配置されているとき，交叉はこれらの変数の集合が同一の子個体に受け継がれることを保証せず，部分解を適切に組み合わせることができない．同一の部分解を構成するような遺伝子の集合をリンケージ集合とよぶ．リンケージ集合を構成する遺伝子をあらかじめ調べることで，より効率的な探索が可能になると考えられる．この手法はリンケージ同定と呼ばれる．

リンケージ同定の手法としては以下のようなものが存在する．

(1) 遺伝子の値の摂動による適応度の変化量を利用する手法

(2) 有望な部分個体群の分布を評価し，変数間の依存関係を含めた確率モデルを構成する手法

(1) の手法である GEMGA [3] は，各ストリングの各遺伝子の値を摂動させて適応度の変化を調べることで，局所最適に属する可能性のある遺伝子を検出し，間接的にリンケージ同定を行うアルゴリズムである．また，LINC [5] は，2 つの遺伝子の値をまとめて摂動させたときの適応度の変化量が，遺伝子をそれぞれ個別に摂動させたときの適応度の変化量の和に等しいときこれらの遺伝子は個別に最適化し，後から組み合わせることが可能であるとしている．そうでないとき，すなわち適応度の変化量が非線形性を示すとき，これらの遺伝子の組はリンケージ集合として交叉の際に同時に交換されるべきであるとされる．さらに Heckendorn ら [2] により，遺伝子の値の摂動を利用する手法は Walsh 解析を利用して一般化された．

これらの手法は，適応度の差分を利用していることから，各部分解の適応度全体への寄与の大きさの違いの影響を受けにくいという利点がある．また，探索の前処理としてリンケージ同定を行うために，その後の探索が効率よく実行できる．しかし，正確なリンケージ情報を得るためには，$O(l^2)$ の適応度評価コストが必要となる．

(2) の手法は分布評価アルゴリズム (EDA) と呼ばれ，優れた個体群における個体の分布を評価し，確率モデル化する．EDA は，PBIL [7] のようにそれぞれの遺伝子の値の分布を独立に評価する単純なモデルからはじまり，現在では BOA [6] のような多変数間の依存関係をベイシアンネットワークによりモデル化し複雑な問題構造を表現できるアルゴリズムが提案されている．EDA は，リンケージ同定 (確率モデル構築) に適応度の計算のオーバヘットを含まない．しかし，適切なモデルを得るためには数世代の評価をくり返す必要があり，この間に一部の部分問題において有望な部分解が失われ，適切なモデルが構築できない可能性がある．

† 北海道大学工学研究科システム情報工学専攻
Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University.
†† 北海道大学情報基盤センター 大規模計算システム研究部門
Division of Large-Scale Computational Systems, Information Initiative Center, Hokkaido University.

(1) 個体群を初期化する
(2) 各遺伝子座 i に対して
  (a) 各ストリングで i の値を反転させ，適応度の差分の絶対値を記録する
  (b) 遺伝子 i の値を 1 に統一する
  (c) 適応度の差分の絶対値に基づき，ストリングを分類する (**図 2**)
  (d) 分類した各部分個体を調査し，リンケージ集合を構成する
  (e) 得られたリンケージ集合から最適なものを選択する
  (f) 遺伝子 i の値を初期状態にもどす

図 1 アルゴリズム

本論文では，部分解の寄与度の大きさの影響を受けず，かつリンケージ同定のための適応度評価コストの小さい手法として，両者の手法を組み合わせて利用する手法を提案する．提案手法では，リンケージ同定は EDA と同様に個体群の分布をよりよく表現するモデルを選択することで行われる．ただし，モデル構築のための部分個体群は，適応度の値そのものではなく，遺伝子の値の摂動による変化量を基準として選択される．また，モデルは BOA のような次世代の遺伝子の値の確率分布ではなく，LINC のような関連する遺伝子の集合で表現される．次世代の個体はこの集合を用いた交叉により生成される．

2. 遺伝子摂動と確率モデルを利用したリンケージ同定

各遺伝子座 (変数) は識別番号 $i = 1, 2, \cdots, l$ で表現される．1 次元に符号化されたストリングの場合，識別番号 i は遺伝子座のストリング上での物理的な位置 (右から i 番目など) に対応する．しかし，提案手法は遺伝子のストリング上での物理的な位置に依存しないため，ストリングは必ずしも 1 次元に符号化される必要はなく，各遺伝子座に一意に識別番号が割り当てられていればよい．

2.1 アルゴリズム

提案手法のアルゴリズムを図 1 に示す．はじめに，個体群中の各ストリングで遺伝子座 i の値を摂動させ，適応度の差分の絶対値

$$|df_i(s)| = |f(s) - f(s_i)| \tag{1}$$

を記録する．ここで s_i は i 番目の遺伝子座の値を摂動 ($1 \to 0$ もしくは $0 \to 1$) させたストリングである．$|df_i(s)|$ に基き，各個体 s を部分個体群に分類する．同一，もしくは近い $|df_i(s)|$ を持つ個体 s は同一の部分個体群に分類される．分類方法については，2.2 で述べる．また，後のリンケージ同定の作業を容易にするために，遺伝子 i の値を 1 に統一する．リンケージ同定の基準となるのは適応度の差分の絶対値であり，遺伝子 i の値そのものはアルゴリズムにとって本質的でないためである．

各部分個体群に対して，リンケージ集合 I を構成した場合の分布のエントロピー $E(M_I)$ を計算する．ただし I は必ず遺伝子 i を含む．$E(M_I)$ は ECGA [1] でも用いられた個体群の表現に際する複雑さを定義する尺度であり，次式で計算される．

$$E(M_I) = -\sum_{x=1}^{\#\_schema} p \log_2 p \tag{2}$$

$$p = \mathrm{counts}_x / n \tag{3}$$

n は部分個体群サイズ，$\#\_schema$ はリンケージ集合 I で定義される可能なスキーマ数，counts_x はスキーマ x の部分個体群中での頻度をあらわす．

I は i のみを要素として持つ集合として初期化される．$\forall j \notin I$ に関して $I = I + \{j\}$ としたときの式 (2) を計算し，最も小さい $E(M_I)$ を与える遺伝子座 j が集合 I に追加される．集合の要素数が k(k はユーザ定義の依存関係の最大次数) になるまで，この操作は繰り返される．

各部分個体群に対して上記のリンケージ集合 I を構成し，もっとも小さな $E(M_I)$ を与えた部分個体群から得られたリンケージ集合を i に関するリンケージ集合とする．

2.2 適応度の差分による個体群分割

EDA において個体群分布から変数間の依存関係を同定するためには，個体群分布が問題構造に関する情報を持つ必要がある．完全にランダムに初期化された初期個体群は，問題に関するいかなる情報も持たない．多くの EDA では初期個体群から適応度の高い部分個体群を選択し，部分個体群内部のストリングの偏りから問題構造に関する情報を得ている．しかし，適応度の高さはストリング全体に対する評価であり，部分解に対する評価ではないため，問題構造を正確に与えるような部分個体群を得るためには，数世代の操作を繰り返さなければならないことが多い．

以下で適応度の差分により分割された部分個体群が個体分布のエントロピーによるリンケージ同定を可能にする個体分布の低い不確定度を与える理由について述べる．

問題は部分問題に (準) 分割可能であるとする．よって，適応度がいくつかの部分問題の適応度の線形和に近い形で表現可能である．言い換えれば，全体の適応度を f，各部分問題 I の適応度を $f_I(s_I)$ とすれば，

$$f(s) = \sum_{\forall I} f_I(s) \tag{4}$$

なる構造を持つ．

対象とする遺伝子座 i の含まれるリンケージ集合を $\hat{I}$ とする．式 (4) は

$$f(s) = f_{\hat{I}}(s) + \sum_{\forall I \neq \hat{I}} f_I(s) \tag{5}$$

と書ける．ゆえに式 (1) は

$$|df_i(s)| = |(f_{\hat{I}}(s) + \sum_{\forall I \neq \hat{I}} f_I(s)) - (f_{\hat{I}}(s_i) + \sum_{\forall I \neq \hat{I}} f_I(s_i))|$$

である．ここで $f_I(s)$ は $i \notin I$ のとき遺伝子 i の値とは無

部分解	
111***	f_1
110***	f_2
101***	f_3
100***	f_4
011***	f_5
010***	f_6
001***	f_7
000***	f_8

表 1 部分解と適応度

$\|df_3(s)\|$	s	部分個体群
$\|f_1-f_2\|$	111***	{111***}
	110***	
$\|f_3-f_4\|$	101***	{101***}
	100***	
$\|f_5-f_6\|$	011***	{011***}
	010***	
$\|f_7-f_8\|$	001***	{001***}
	000***	

表 2 適応度の変化量

(1) 1 個体を 1 クラスとして初期化
(2) すべての個体の組 (s_p, s_q) に対して $||df_i(s_p)| - |df_i(s_p)||$ を調査し，値が最も小さい組を 1 つのクラスにまとめる
(3) 新たなクラスの $|df_i(s)|$ をそのクラスに属するすべての個体の $|df_i(s_p)|$ の平均に更新する
(4) (2),(3) を終了条件が満足されるまでくり返す.

図 2 分類アルゴリズム

関係に決定されるため

$$\sum_{\forall I \neq \hat{I}} f_I(s) = \sum_{\forall I \neq \hat{I}} f_I(s_i)$$

である．よって，上式は

$$|df_i(s)| = |f_{\hat{j}}(s) - f_{\hat{j}}(s_i)| \tag{6}$$

となる．

部分問題の次数の上界が k であるとき，この部分問題は 2^k 個の解を持ち，最大で 2^k 個の部分適応度を持つ．リンケージ集合 $I-\{i\}$ に含まれる $k-1$ 個の遺伝子の値の可能な組み合わせ (スキーマ) は，2^{k-1} 通りである．$|df_i(s)|$ は，最大で 2^{k-1} 通りの値を持ち，個体群は $|df_i(s)|$ の大きさによって 2^{k-1} 個の部分個体群に分類される．同一もしくは近い値の $|df_i(s)|$ を持つ部分個体群を $\mathrm{SP}_{|df_i(s)|}$ と書く．$|df_i(s)|$ は式 (6) より $\hat{I}$ に含まれる遺伝子座の値のみに依存して決定されるため，$\mathrm{SP}_{|df_i(s)|}$ に含まれるストリング上の遺伝子でのうち $\hat{I}$ に含まれない遺伝子は，$|df_i(s)|$ の大きさに影響を与えない．よって，$\mathrm{SP}_{|df_i(s)|}$ においてこれら遺伝子の値は偏りなく分布する．一方，各 $\mathrm{SP}_{|df_i(s)|}$ において，部分問題を構成する遺伝子座は，ある部分適応度の差分 $|df_i(s)| = |f_{\hat{j}}(s) - f_{\hat{j}}(s_i)|$ を与えるような特定の値の組み合わせのみを取る．ゆえに，$\hat{I}$ に含まれる遺伝子座は，適応度の差分から分類された部分個体群において低い不確定性を示し，$\hat{I}$ に含まれない遺伝子座は高い不確定性を示す．このような低い不確定性を与える遺伝子座の集合を検出することでリンケージ同定が可能となる．

例として，遺伝子座 $i=1,2,3$ から構成される部分問題 **fff*****の適応度がそれぞれ表 1 のような値をとる場合を考える．遺伝子座 $i=3$ の摂動による適応度の変化量の絶対値 $|df_3(s)|$ は表 2 で与えられる．$|df_3(s)|$ の値により個体群は部分個体群 $\mathrm{SP}_{|df_3(s)|}$ に分類される．ただし，**図 1** の (2b) より，同一の $|df_3(s)|$ で 3 番目の遺伝子の値はすべて 1 に統一される．各部分個体群 $\mathrm{SP}_{|df_3(s)|}$ には，表 2 の s' で表現される個体が分類される．$\{i=1,2,3\}$ は各部分個体群内部で収束し，それ以外の*部分はランダムな値であると考えられる．よって，これらの部分個体群を調査することで $i=3$ に関するリンケージ集合 $\{1,2,3\}$ が得られる．

つぎに，i に関連する $k-1$ ビットの異なる組み合わせが i の摂動によりすべて同一の $|df_i(s)|$ を与えるような場合を考える．$|df_i(s)|$ により分類が不可能な関数の場合，提案手法でリンケージを同定することは不可能である．このような問題は $|df_i(s)|$ が様々に異なる問題と比較して，平坦な適応度地形を持ち，極端な騙し性や干し草のなかの針のような困難さを持たない．

さらに，i に関連する $k-1$ ビットの異なる組み合わせの幾つかが i の摂動により同一の $|df_i(s)|$ を与えるような問題を考える．このような問題に対しても，多くの場合で提案手法はリンケージを同定することが可能である．表 2 の例で，$|f_1-f_2|$ 以外のすべてで，$|df_i(s)|$ がほぼ等しい値であったと仮定する．このとき，$|f_1-f_2|$ を与える部分個体群と $|f_3-f_4| = |f_5-f_6| = |f_7-f_8|$ を与える部分個体群が得られる．明らかに，後者の部分個体群によるモデル化は大きな $E(M_I)$ を与え，前者の部分個体群によるモデル化では小さな $E(M_I)$ を与える．$E(M_I)$ の値を基準として，得られたモデルを選択することで，適切なリンケージ集合を得ることができる．さらに，部分個体群が 2 つのスキーマを含む場合でも，これらのスキーマの遺伝子座の集合は小さな $E(M_I)$ を与える．

2.3 部分個体群への分類

個体群を部分個体群に分類する手法としては重心法を用いたが，別の方法を用いることも可能である．4 の終了条件は，最小の $||df_i(s_p)| - |df_i(s_p)||$ があらかじめ定められた閾値 θ よりも大きくなったか，もしくは，クラス数があらかじめ定められた数よりも少なくなったときとする．関数全体の非線形性や部分関数のオーバラップが大きいとき，$|df_i(s)|$ の分散が大きくなると考えられるため，θ は比較的大きな値に設定される．逆に，関数全体の非線形性が小さいとき，θ は比較的小さい値に設定される．重心法による個体群分類アルゴリズムを図 2 に示す．

3. 実　　験

3.1 トラップ関数の線形和

テスト関数として 5 ビットのトラップ関数の和

$$f(s) = \sum_{i=1}^{m} \mathrm{trap}(s_{5\cdot i}, \cdots, s_{5\cdot i+4}) \tag{7}$$

$$\mathrm{trap}(s_{5\cdot i}, \cdots, s_{5\cdot i+4}) = \begin{cases} 5 \ (u=5) \\ 4-u(\mathrm{otherwise}) \end{cases}$$

を用いた．ただし m は部分関数の個数である．また，u は $s_{5\cdot i}, \cdots, s_{5\cdot i+4}$ に含まれる 1 の個数である．

また，同様の各部分関数の適応度全体への寄与を変化させた指数関数的に変化させた重みつきトラップ関数を用い

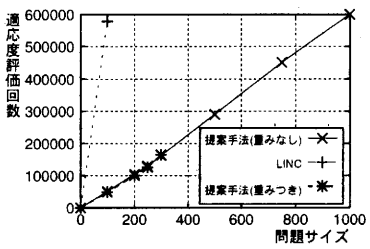

図 3 トラップ関数による実験結果：問題サイズ – 適応度評価回数

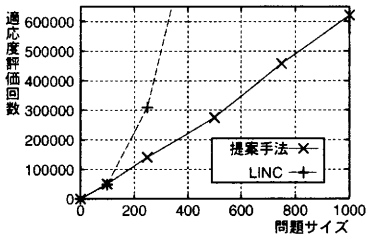

図 4 GW2 関数による実験結果：問題サイズ – 適応度評価回数

た実験を行った．重みつきトラップ関数とは，式 (7) において部分関数を $2^{i-1} \times \mathrm{trap}(s_{5\cdot i}, \cdots, s_{5\cdot i+4})$ とした関数である．

各個体長 l に対して，10 回の試行すべてで正しくリンケージ同定ができたときの適応度評価回数を記録した．個体群サイズは 10 回すべてで正しくリンケージ同定ができるまで，徐々に増やした．

この結果を図 3 に示す．比較対象として問題サイズ l に対して $O(l^2)$ のアルゴリズムである LINC が正しくリンケージ同定をするのに必要な適応度評価回数を示す．提案手法は，問題サイズに対してほぼ線形の計算量で正確なリンケージを得ることができる．

また，提案手法では関数が指数関数的重みを持つときも，通常の場合と同等の効率で解を得ることができる．しかし，BOA などの EDA では適応度の値そのものを探索の指針としているために，重みが大きな部分関数から順番にモデル化が行われ，結果として探索の効率が落ちてしまう [6]．

3.2 GW2 関数

GEMGA [3] は，遺伝子の値の摂動を用いるアルゴリズムであり，$O(l)$ のアルゴリズムとして提案された．しかし，このアルゴリズムでは，変数間の依存関係は明確に定義されず，適応度の低下量の大きさにより暗黙に定義されていた．そのため，1 つの部分問題の解に含まれる遺伝子がそれぞれ別々の適応度変化量を与えるような問題を解くことができない．この問題を解くために GEMGA は拡張され [4]LINC と同様に $O(l^2)$ のアルゴリズムとなった．問題例として，Goldberg-Wang(GW) 関数 [4] が存在する．

GW2 関数 (**表 3**) は部分解に含まれる 1 の数 u により以下のように定義される関数を鎖状につなげた関数である．上式で k は部分問題の次数を表す．また $\bar{x}$ は x の値をすべて反転させた部分ストリングとする．

$\phi(x) = 10$ if $u = 0$	$\phi_1(x, v_1, v_2)$
$= \phi_1(x, 7, 2)$ if $u = 1$	`int i = 0;`
$= \phi_1(\bar{x}, 4, 3)$ if $u = k-1$	`while(x[i] == 0) i++;`
$= 8$ if $u = k$	`if((i%2) == 1) return v1;`
$= 0$ otherwise (8)	`else return v2;`

表 3 GW2 関数

$k = 5$ で固定し，l をさまざまに変化させたときの関数による実験の結果を**図 4** に示す．この関数に対しても提案手法は $O(l)$ で解を得ることができる．部分問題内部での凹凸や非線形性の強さから，LINC では部分関数内部に線形アトラクタを含むトラップ関数と比較して，少ない適応度評価回数で解を得ることができる．しかし，問題サイズが大きいとき，提案手法と比較して非常に大きな適応度評価回数が必要である．逆に，提案手法では，部分関数内部の適応度地形の複雑さから，正確なリンケージをするためには，トラップ関数よりも大きな個体群サイズが必要となる．

4. おわりに

対象となる遺伝子の摂動に対して，同一の適応度の変化量を与える個体群の分布を調査し，リンケージを同定する手法を提案した．通常の EDA では部分関数の適応度全体への寄与が大きく異なる場合に探索性能が低下するが，提案手法では，そのような問題であっても，各部分関数の適応度寄与がほぼ等しい場合と同様に，問題のサイズに対してほぼ線形の適応度評価回数でリンケージ同定を行うことができる．

参考文献

1) Harik, G.: Linkage learning via probabilistic modeling in the ECGA, Technical Report IlliGAL No.99010, University of Illinois(1999).
2) Heckendorn, R. B. and Wright, A. H.: Efficient Linkage Discovery by Limited Probing by Heckendorn, *Proceedings of the GECCO 2003*, pp. 1003–1014 (2003).
3) Kargupta, H.: SEARCH, evolution, and the gene expression messy genetic algorithm, Report LA-UR 96-60, Los Alamos National Laboratory(1996).
4) Kargupta, H., Goldberg, D. E. and Wang, L.: Extending The Class of Order-k Delineable Problems For The Gene Experession Messy Genetic Algorithm, *Proceedings Of Genetic Programming Conference*, pp. 364–369 (1997).
5) Munetomo, M. and Goldberg, D. E.: Identifying Linkage Groups by Nonlinearity/Non-monotonicity Detection, *Proceedings of the GECCO 1999*, pp. 433–440 (1999).
6) Pelikan, M.: *Bayesian Optimization Algorithm: From Single Level to Hierarchy*, Doctroral dessertation, University of Illionis(2002).
7) Salustowicz, R. P. and Schmidhuber, J.: Probabilistic Incremental Program Evolution: Stochastic Search Through Program Space, *Machine Learning: ECML-97*, Vol. 1224, pp. 213–220 (1997).