A Genetic Model Based on Simulated Crossover of Quaternary Genes for Quadratic Fitness

Marco Carpentieri
Dipartimento di Matematica ed Informatica,
Università di Salerno, Fisciano (SA), Italy
marcar66@virgilio.it

Vito Fedullo
Dipartimento di Matematica ed Informatica,
Università di Salerno, Fisciano (SA), Italy
vfedullo@unisa.it

ABSTRACT

We present a genetic model based on simulated recombination of fixed sequences of quaternary genes (assuming four distinct forms, or alleles). States and dynamics of the infinite population genetic system, represented by the model, are derived for quadratic fitness functions. The bivariate marginal distribution genetic algorithm, devised to simulate the system, is experimentally compared with the univariate marginal distribution genetic algorithm with bit-based simulated crossover (in the case of infinite populations) for the $\mathbf{MAX-CUT}$ problem.

Categories and Subject Descriptors

I.6.5 [**Computing Methodologies**]: Simulation and Modeling, Model Development

General Terms

Algorithms

Keywords

Marginal Distribution Genetic Algorithms

1. RESULTS

We introduce a genetic model based on simulated crossover of fixed sequences of two bit genes. Such a model represents an instance of the *Random Heuristic Search* (as defined in chapter 3 of reference [4]) and can be considered as an extension of the model presented in [1]; the main characteristic of the system that we shall consider is that the recombination of the genetic material is obtained by performing a weighted average of the alleles along each fixed two-bit locus and using such statistics to produce offspring whose alleles in distinct loci are independently generated. The genetic algorithm, simulating the model we propose, belongs to the so-called class of the *Marginal Distribution Genetic Algorithms*. Note that the interest in devising marginal distribution genetic models lies in the fact that (along with restriction of the fitness to some classes of functions) they consent efficient simulation for infinite populations; moreover, in case of univariate marginal distributions, they have been used to construct approximation algorithms to solve (hard) combinatorial problems for which error bounds can be theoretically estimated.

A population P of individuals is represented by a multiset of $n \in \mathbf{N}$ $l-$length binary strings (we suppose l even) in the set $\Omega = \{0,1\}^l = \{\omega_1, \ldots, \omega_{2^l}\}$. Each population P is associated to its frequency vector $\mathbf{F} = (F_{\omega_1}, \ldots, F_{\omega_{2^l}})$ specifying the proportion of the strings in Ω contained in P. Each individual is evaluated by his fitness that is measured by means of a fitness function $f : \Omega \longrightarrow \mathbf{R}^+$. Let $A = \{00, 01, 10, 11\}$ and $B = A - \{00\}$. For $k = 1, \ldots, \frac{l}{2}$ and $a = a_1 \cdot a_2 \in A$, consider functions $\chi_k[a] : \Omega \longrightarrow \{0,1\}$ defined by

$$\chi_k[a](\omega) = \begin{cases} 1, & \text{if } a_1, a_2 \text{ are in positions } 2k-1, 2k \text{ of } \omega; \\ 0, & \text{otherwise.} \end{cases}$$

We shall use notation $E_{\mathbf{p}}[X] = \sum_{i=1}^{2^l} X(\omega_i) p_i$ to mean the expectation of function $X : \Omega \to \mathbf{R}$ considered as a random variable along with the stochastic vector $\mathbf{p} = (p_1, \ldots, p_{2^l})$. Starting from an initial population P_0, if at time t the state of the (genetic) system is the population P, represented by its frequency vector $\mathbf{F}$, then the population at time $t+1$ is obtained as follows:

1. for every $k = 1, \ldots, \frac{l}{2}$ and $a \in A$ compute $\phi_{k,\mathbf{F}}[a] = \frac{E_{\mathbf{F}}[\chi_k[a] f]}{E_{\mathbf{F}}[f]}$;
2. generate a new population P' of n $l-$length binary strings, denoted by $P' = \{\omega_{r_1}, \ldots, \omega_{r_n}\}$, with probability $\phi_{k,\mathbf{F}}[a]$ of obtaining a_1, a_2 in positions $2k-1, 2k$ independently from r_i and k for $1 \leq k \leq \frac{l}{2}$ and $1 \leq i \leq n$.

By definition of the *recombination* process described in 2., if P is a population at time t and $\mathbf{F}$ its frequency vector, the population at time $t+1$ is obtained by selecting n strings with probability distribution $\mathbf{\Phi}(\mathbf{F}) = (\mathbf{\Phi}(\mathbf{F})_{\omega_1}, ..., \mathbf{\Phi}(\mathbf{F})_{\omega_{2^l}})$, where the probability $\mathbf{\Phi}(\mathbf{F})_{\omega_j}$ of generating the string $\omega_j = \omega_{j,1} \cdots \omega_{j,l}$ is $\mathbf{\Phi}(\mathbf{F})_{\omega_j} = \prod_{k=1}^{\frac{l}{2}} \phi_{k,\mathbf{F}}[\omega_{j,2k-1} \cdot \omega_{j,2k}]$.

In the following, for $k = 1, \ldots, \frac{l}{2}$, $j = 1, 2$, $\{i_1, \ldots, i_j\} \subseteq \{1, 2\}$ $(i_1 \leq i_j)$ and $(a_1, \ldots, a_j) \in \{0,1\}^j$ we shall adopt notation

$$\phi_{k,\mathbf{F},(i_1,\ldots,i_j)}[a_1, \ldots, a_j] = \sum_{\substack{a' \in A \\ (a'_{i_1}, \ldots, a'_{i_j}) = (a_1, \ldots, a_j)}} \phi_{k,\mathbf{F}}[a'],$$

to denote the probability distribution $\phi_{k,\mathbf{F}}[a'] = \phi_{k,\mathbf{F}}[(a'_1, a'_2)]$ viewed as a joint probability over $\{0,1\}^2$ and its marginal distributions.

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

	Vertices	10	14	18	22	26	30	34
$p=\frac{1}{7}$	**1BGSC**	6, 25	12, 00	18, 10	26, 45	39, 05	50, 05	62, 23
	2BGSC	6, 30	12, 10	18, 20	26, 65	39, 40	50, 50	63, 45
	12BGSC	6, 32	12, 25	18, 40	26, 80	39, 75	50, 85	63, 80
	Edges	6, 42	13, 00	20, 80	31, 71	46, 42	62, 14	80, 14
$p=\frac{1}{4}$	**1BGSC**	8, 85	18, 90	30, 05	43, 65	58, 10	80, 04	96, 63
	2BGSC	9, 05	19, 03	30, 20	44, 05	58, 30	80, 29	97, 33
	12BGSC	9, 25	19, 45	30, 40	44, 45	58, 80	81, 25	98, 05
	Edges	11, 25	22, 75	40, 05	57, 75	81, 25	108, 75	140, 25

Table 1: Mean Size of the cuts found by the genetic algorithms

Theorem 1 Let $n \geq 8\left(\frac{M}{\epsilon E_{\mathbf{\Phi}(\mathbf{F})}[f]}\right)^2 \log\left(\frac{24\,l}{\delta}\right)$, where $\epsilon, \delta \in (0,1]$ and M is the maximum value that the fitness function can assume; if at time t the system is in the state $\mathbf{F}$, then the state $\mathbf{F}'$ at time $t+1$ is such that for all $k = 1, \ldots \frac{l}{2}$, $j = 1, 2$, $\{i_1, \ldots, i_j\} \subseteq \{1, 2\}$ $(i_1 \leq i_j)$ and $(a_1, \ldots, a_j) \in \{0,1\}^j$ it results $\left|\phi_{k,\mathbf{F}',(i_1,\ldots,i_j)}[a_1, ..., a_j] - \phi_{k,\mathbf{\Phi}(\mathbf{F})),(i_1,\ldots,i_j)}[a_1, ..., a_j]\right| < \epsilon$ with probability at least $1-\delta$.

Proof It is a result obtained by a first-order approximation and based on the Hoeffding's inequality [3]. □

We consider fitness functions $f : \Omega \longrightarrow \mathbf{R}^+$ that can be represented by quadratic functions defined on $[0,1]^l$, coincident with f on Ω, of the form $Qf(x_1, \ldots, x_l) = \sum_{\substack{i,j=1 \\ i \neq j}}^{l} w_{i,j} x_i (1 - x_j)$. Let $q = \frac{l}{2}$. In case of infinite populations, Theorem 1 states a probability convergence result according to which the stochastic genetic system becomes an iterative deterministic system whose states are $3q-$component vectors $\mathbf{\Psi} = (\psi_{1,01}, \psi_{1,10}, \psi_{1,11}, \ldots, \psi_{q,01}, \psi_{q,10}, \psi_{q,11}) \in [0,1]^{3q}$ and whose dynamics is described by the equations $\psi_{k,z}(t+1) = \frac{\psi_{k,z}(t)\xi_{k,z}(\mathbf{\Psi}(t))}{p(\mathbf{\Psi}(t))}$ for $1 \leq k \leq q$, $z \in B$, where, by independence of the alleles generated in distinct loci and the linearity properties of the mean, we get

$$p(\mathbf{\Psi}(t)) = \sum_{u=1}^{q} \sum_{\substack{i,j=1 \\ i \neq j}}^{2} w_{2(u-1)+i,2(u-1)+j} \sum_{\substack{a \in B \\ a_i=1, a_j=0}} \psi_{u,a}(t) + \sum_{\substack{u,u'=1 \\ u \neq u'}}^{q} \sum_{i,j=1}^{2} w_{2(u-1)+i,2(u'-1)+j} \sum_{\substack{a \in B \\ a_i=1}} \psi_{u,a}(t)\bar{\psi}_{u',j}(t)$$

and

$$\begin{aligned}\xi_{k,z}(\mathbf{\Psi}(t)) &= \sum_{\substack{i,j=1 \\ i \neq j \\ z_i=1, z_j=0}}^{2} w_{2(k-1)+i,2(k-1)+j} \\ &+ \sum_{\substack{u=1 \\ u \neq k}}^{q} \sum_{\substack{i,j=1 \\ i \neq j}}^{2} w_{2(u-1)+i,2(u-1)+j} \sum_{\substack{a \in B \\ a_i=1, a_j=0}} \psi_{u,a}(t) \\ &+ \sum_{\substack{u=1 \\ u \neq k}}^{q} \sum_{j=1}^{2} \sum_{\substack{i=1 \\ z_i=1}}^{2} w_{2(k-1)+i,2(u-1)+j} \bar{\psi}_{u,j}(t) \\ &+ \sum_{\substack{u=1 \\ u \neq k}}^{q} \sum_{j=1}^{2} \sum_{\substack{i=1 \\ z_i=0}}^{2} w_{2(u-1)+j,2(k-1)+i} \sum_{\substack{a \in B \\ a_j=1}} \psi_{u,a}(t) \\ &+ \sum_{\substack{u,u'=1 \\ u,u' \neq k \\ u \neq u'}}^{q} \sum_{i,j=1}^{2} w_{2(u-1)+i,2(u'-1)+j} \sum_{\substack{a \in B \\ a_i=1}} \psi_{u,a}(t)\bar{\psi}_{u',j}(t),\end{aligned}$$

with $\bar{\psi}_{u,j}(t) = \left(1 - \sum_{\substack{a \in B \\ a_j=1}} \psi_{u,a}(t)\right)$.

Following the guidelines of reference [1], the system has been simulated to solve (in the sense of an approximation algorithm) the $\mathbf{MAX-CUT}$ (*hard* optimization) problem for $p-$random undirected graphs. In Table 1 there are the results of the simulations intended to compare the performances of the genetic algorithm with simulated crossover of sequences of two-bit genes (**2BGSC**) and that with recombination of sequences of binary genes (**1BGSC**) introduced in [1]. In the table it is also reported the expected number of edges (row Edges) of the $p-$random graphs. The algorithm **2BGSC** consents to improve (in average) the performances of **1BGSC** that in [1], in case of $p-$random graphs for $p = \frac{1}{7}, \frac{1}{4}$, exhibited performances slightly worse than that of an Hopfield's network [1]. In Table 1 there are also the mean sizes of the cuts obtained by choosing the best ones, for a same $p-$random graph, found by the two algorithms (row **12BGSC**). Note that the performances were dependent on specific $p-$random graphs; this fact, supported by the remark that the concept of gene in classic genetic algorithms is more general than that provided in univariate marginal distribution ones, evidences the need of the introduction and analysis of a more general (infinite population) dynamical system marginal distribution model.

2. ACKNOWLEDGMENTS

This paper is in memory of Professor Mario Carpentieri.

3. REFERENCES

[1] A. Bertoni, P. Campadelli, M. Carpentieri, G. Grossi, *A Genetic Model and the Hopfield Networks*, *Icann96*, pp. 463 − 468, Springer-Verlag, Berlin, Germany, 1996.

[2] A. Bertoni, P. Campadelli, M. Carpentieri, G. Grossi, *Analysis of a Genetic Model*, *ICGA97: Proceedings of the Seventh International Conference on Genetic Algorithms*, pp. 121 − 126, Morgan-Kaufmann, San Mateo, California, 1997.

[3] A. Bertoni, P. Campadelli, M. Carpentieri, G. Grossi, *A Genetic Model: Analysis and Application to MAXSAT*, *Evolutionary Computation*, pp. 291 − 309, Vol. 8, n. 3, Massachusetts Institute of Technology, Mit Press, Cambridge, MA, 2000.

[4] M. D. Vose, *The Simple Genetic Algorithm: Theory and Foundations*, MIT Press, Cambridge, MA, 1999.