УДК 681.3.06:007

ПРИМЕНЕНИЕ ТЕОРЕТИКО-ВЕРОЯТНОСТНОГО ОПИСАНИЯ БАЗОВЫХ ОПЕРАТОРОВ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ ДЛЯ ОЦЕНКИ ДИНАМИКИ КАЧЕСТВА ПОПУЛЯЦИИ

Петров А.В.
Кафедра информатики

Цель работы состоит в создании теоретико-вероятностного описания базовых операторов ГА для количественной оценки динамики критериев качества популяции в зависимости от вероятностных параметров алгоритма.

This paper is intended to focus on a probability-theoretical description of basic GA operators and aims at providing a stochastic parameter-dependent, quantitative estimate of criteria dynamics of the generation quality and fitness.

Введение. Практическая реализация генетических алгоритмов (ГА) и их применение для решения прикладных задач в большинстве случаев базируется на описательном материале, не содержащем математических оснований этого класса эволюционных моделей. Однако вопрос сходимости ГА представляет большой интерес не только с теоретической точки зрения. Формализация процесса эволюции может привести к вытеснению случайного поиска и замене его какими-либо детерминированными вычислительными процедурами при решении проблемы выбора оптимальных параметров ГА.

На сегодняшний день разработано несколько подходов к объяснению сходимости процесса эволюции, основанных, в том числе, на утверждениях, которые высказали Холланд (Holland, 1975) и Прайс (Price, 1970). Сформулируем и докажем первую теорему.

1. Теорема Холланда. *Теорема о схеме (Schema Theorem)* [3, 4], предложенная Дж. Холландом, рассматривается сегодня как основная теорема о ГА и является одной из самых известных попыток формального доказательства их сходимости.

Определение 1. (Схема, определенный разряд.) *Пусть пространство двоичных хромосом Ξ составлено из строк длины L. Тогда схемой $\mathfrak{H}$, по Холланду, называется строка длины L над алфавитом $\{0, 1, *\}$, где "неопределенный" символ $*$ имеет смысл "либо 0, либо 1".*

Разряды, содержащие 0 и 1, носят название определенных.

Определение 2. (Порядок, определенная длина схемы.) *Пусть, по Холланду, $\mathfrak{H}$ – схема. Тогда ее порядок $o(\mathfrak{H})$ – это количество содержащихся в $\mathfrak{H}$ определенных разрядов.*

Определенная длина схемы $\delta(\mathfrak{H})$ есть расстояние между крайними определенными разрядами $\mathfrak{H}$.

Определение 3. (Представитель схемы.) *Пусть ξ – двоичная хромосома длины L из множества Ξ, $\mathfrak{H}$ – схема той же длины. Тогда ξ принадлежит $\mathfrak{H}$ (представляет ее, или является ее представителем), если значения определенных разрядов схемы $\mathfrak{H}$ и хромосомы ξ совпадают.*

Замечание 1. *Количество представителей схемы* $\mathfrak{H}$ *в множестве* Ξ *составляет* $M_\Xi(\mathfrak{H}) = \left|\{\xi \mid \xi \in \Xi, \xi \in \mathfrak{H}\}\right| = 2^{L-o(\mathfrak{H})}$.

Определение 4. (Качество схемы.) *Качество, или приспособленность, схемы* $\mathfrak{H}$ *в популяции* $\mathfrak{G}_t$ *определяется средней приспособленностью ее представителей в этой популяции:* $\overline{F}(\mathfrak{H}) = \sum_{\mathfrak{g}_{tj} \in \mathfrak{H}} F(\mathfrak{g}_{tj}) / M_t(\mathfrak{H})$, *где* $\mathfrak{g}_{tj}$ *– j-я хромосома популяции* $\mathfrak{G}_t$, $M_t(\mathfrak{H})$ *– количество представителей* $\mathfrak{H}$ *в* $\mathfrak{G}_t$.

Теорема Холланда устанавливает соотношение между количеством представителей схемы в популяциях $\mathfrak{G}_t$ и $\mathfrak{G}_{t+1}, t \geq 1$ и формулируется следующим образом [3]:

Теорема 1. (О схеме.) *В генетическом алгоритме, использующем механизм пропорционального отбора, а также одноточечное скрещивание, происходящее с вероятностью* P_c, *при заданной кратности хромосом в популяции* $\mathfrak{G}_t$ *и среднем значении функции качества хромосом* $\overline{F}$, *для каждой схемы* $\mathfrak{H}$ *имеет место следующее неравенство:*

$$M_{t+1}(\mathfrak{H}) \geq M_t(\mathfrak{H}) \cdot \frac{\overline{F}(\mathfrak{H})}{\overline{F}} \left(1 - P_c \cdot \frac{\delta(\mathfrak{H})}{L-1}\right). \qquad (1)$$

Доказательство. При пропорциональном отборе каждая хромосома $\mathfrak{g}_{tj}$ из $\mathfrak{G}_t, 1 \leq j \leq N$, выбирается для одного акта скрещивания с вероятностью

$$\Pr(S(\mathfrak{G}_t) = j) = F(\mathfrak{g}_{tj}) \Big/ \sum_{q=1}^{N} F(\mathfrak{g}_{tq}),$$

где $S : \mathfrak{G}_t \to \{1, 2, \ldots, N\}$ – вероятностная функция отбора, N – объем популяции, что соответствует

$$NF(\mathfrak{g}_{tj}) \Big/ \sum_{q=1}^{N} F(\mathfrak{g}_{tq}) = F(\mathfrak{g}_{tj}) / \overline{F}$$

актам отбора $\mathfrak{g}_{tj}$ при формировании поколения $\mathfrak{G}_{t+1}$. Поскольку интерес представляют только $\mathfrak{g}_{tj} \in \mathfrak{H}$, то в следующее поколение будет отобрано

$$\sum_{\mathfrak{g}_{tj} \in \mathfrak{H}} \frac{F(\mathfrak{g}_{tj})}{\overline{F}} = \frac{M_t(\mathfrak{H}) \cdot \sum\limits_{\mathfrak{g}_{tj} \in \mathfrak{H}} F(\mathfrak{g}_{tj})}{M_t(\mathfrak{H}) \cdot \overline{F}} = M_t(\mathfrak{H}) \cdot \frac{\overline{F}(\mathfrak{H})}{\overline{F}}$$

представителей $\mathfrak{H}$.

При одноточечном скрещивании схема $\mathfrak{H}$ будет разрушена, если позиция скрещивания, случайно выбранная по равномерному закону распределения, окажется между определенными разрядами. При вероятности скрещивания P_c и определенной длине $\delta(\mathfrak{H})$ вероятность разрушения схемы равна

$$P_1 = P_c \cdot \frac{\delta(\mathfrak{H})}{L-1}.$$

Отсюда следует, что количество представителей $\mathfrak{H}$, сохранившихся после одноточечного скрещивания, составляет

$$M_{t+1}(\mathfrak{H}) = M_t(\mathfrak{H}) \cdot \frac{\overline{F}(\mathfrak{H})}{\overline{F}} \cdot \left(1 - P_c \cdot \frac{\delta(\mathfrak{H})}{L-1}\right).$$

Нестрогое неравенство (1) указывает на то, что представители схемы $\mathfrak{H}$ могут быть образованы при скрещивании каких-либо хромосом из $\mathfrak{G}_t$. ■

З а м е ч а н и е 2. *Введение мутации хромосом с вероятностью единичной мутации P_m приводит к разрушению схемы порядка $o(\mathfrak{H})$ с вероятностью*

$$P_2 = 1 - (1 - P_m)^{o(\mathfrak{H})} \approx P_m \cdot o(\mathfrak{H})$$

для малых P_m и равномерного закона распределения позиций мутации, так что неравенство (1) принимает вид

$$M_{t+1}(\mathfrak{H}) \geq M_t(\mathfrak{H}) \cdot \frac{\overline{F}(\mathfrak{H})}{\overline{F}} \left(1 - P_c \cdot \frac{\delta(\mathfrak{H})}{L-1} - P_m \cdot o(\mathfrak{H})\right). \qquad (2)$$

Теорема Холланда в форме (2) позволяет утверждать, что выживают схемы с низким порядком и малой определенной длиной, характеризующиеся высоким качеством. Гольдберг называет такие схемы *строительными блоками.*

Тем не менее, теорема о схеме подвергается критике, в частности, на том основании, что она обосновывает экспоненциальный рост числа строительных блоков, однако не является доказательством более быстрого, по сравнению со случайным поиском, создания новых решений с высокой приспособленностью. Более того, как отмечает Н. Радклифф, теорема Холланда справедлива даже при случайном поиске, несмотря на то, что она используется для демонстрации превосходства над ним ГА [3]. Большей доказательной силой обладает вторая из названных в п. 1 теорем, однако ее рассмотрение выходит за рамки настоящей статьи.

2. Канонический генетический алгоритм.

О п р е д е л е н и е 5. (Канонический ГА [3].) *Канонический ГА есть динамическая система вида*

$$\Pr(\xi \in \mathfrak{G}_{t+1}) = \sum_{x,\, x' \in \mathbf{X}} \Pr\left(\xi \leftarrow \varrho(x), \varrho(x')\right) \times$$
$$\times \frac{F(\varrho(x))F(\varrho(x'))}{\overline{F}^2} \cdot \Pr\left(\varrho(x) \in \mathfrak{G}_t\right) \cdot \Pr\left(\varrho(x') \in \mathfrak{G}_t\right), \qquad (3)$$

где $\Pr\left(\xi \leftarrow \varrho(x), \varrho(x')\right)$ *– вероятность того, что результатом применения операторов ГА к хромосомам $\varrho(x)$ и $\varrho(x')$, кодирующим соответственно x и x', станет хромосома ξ;* $\mathbf{X}$ *– область определения целевой функции f.*

Особенностью описания (3) является вероятностный подход к оценке степени достоверности наличия конкретной хромосомы в популяции очередного поколения. Независимо от [3] и более ранних работ, автором данной статьи получено аналогичное теоретико-вероятностное описание, позволяющее оценить вероятность наличия отдельной хромосомы в популяции следующего поколения после применения базовых операторов ГА и вероятность тождественности существующей популяции одной из возможных. С целью разумного упрощения и в точном соответствии с теоремой 1 набор базовых операторов ограничен пропорциональным отбором, одноточечным скрещиванием и единичной мутацией.

3. Создание начальной популяции. Пусть хромосома, или *генотип*, представляет собой двоичную строку длины L, и все хромосомы являются разрешенными: $N_0 = 2^L$. Пусть также ГА оперирует популяциями объемом N хромосом. Тогда существует N_0^N различных популяций, образующих универсум $\mathfrak{U} = \{u_p | u_p = \{\xi_{pj_1}, \xi_{pj_2}, \ldots, \xi_{pj_N}\}\}$, $\xi_{pj_1}, \xi_{pj_2}, \ldots, \xi_{pj_N} \in \Xi$, $1 \le j_1, j_2, \ldots, j_N \le N$. Начальный отбор равновероятен, поэтому при однократном акте начального отбора каждая хромосома $\xi \in \Xi$ может быть включена в комплект u_p с вероятностью $\Pr_1(\xi) = 1/N_0$, а значит, первое поколение $\mathfrak{G}_1$ тождественно какому-либо $u_p \in \mathfrak{U}$ с вероятностью

$$\Pr(\mathfrak{G}_1 = u_p) = \prod_{j=1}^{N} \Pr_1(\xi_{pj}) = (1/N_0)^N = 1/N_0^N .$$

Таким образом, существует два многоугольника распределения вероятностей, один из которых задает вероятность включения отдельной хромосомы в очередное поколение, второй – вероятность тождественности очередного поколения конкретной популяции.

4. Пропорциональный отбор. В условиях отсутствия скрещивания популяции $\mathfrak{G}_t$, $t > 1$ формируются прямым копированием в них хромосом предыдущего поколения. Потеря хромосомы носит безвозвратный характер, так как новый генетический материал не образуется. Если хромосома ξ принадлежит комплекту u_p, то вероятность ее попадания в следующее поколение полностью определяется ее качеством:

$$\Pr^s_{t+1}(\xi) = \sum_{u_p : \xi \in u_p} \frac{F(\xi) \cdot \#(\xi, u_p)}{N} \Pr(\mathfrak{G}_t = u_p) = \frac{}{\sum_{j=1}^{N} F(\xi_{pj})}$$

$$= \sum_{u_p : \xi \in u_p} \left(F(\xi) \cdot \#(\xi, u_p) \cdot \prod_{j=1}^{N} \Pr_t(\xi_{pj}) \Big/ \sum_{q=1}^{N} F(\xi_{pq}) \right) .$$

5. Одноточечное скрещивание. Одноточечное скрещивание пары хромосом, происходящее с вероятностью P_c, существенно изменяет динамику системы, разрешая создание нового генетического материала путем перестановки фрагментов генотипов-родителей. Пусть $P^{(I)}_{t+1}(\xi)$ – вероятность переноса хромосомы $\xi \in \Xi$ в следующее поколение путем безусловного копирования, а $P^{(II)}_{t+1}(\xi)$ – вероятность образования $\xi \in \Xi$ в результате применения одноточечного скрещивания. Тогда

$$P^{(I)}_{t+1}(\xi) = (1 - P_c) \Pr_{t+1}(\xi) = \overline{P}_c \cdot \sum_{u_p : \xi \in u_p} \frac{F(\xi) \cdot \#(\xi, u_p)}{N} \Pr(\mathfrak{G}_t = u_p) = \frac{}{\sum_{j=1}^{N} F(\xi_{pj})}$$

$$= \overline{P}_c \cdot \sum_{u_p : \xi \in u_p} \left(F(\xi) \cdot \#(\xi, u_p) \cdot \prod_{j=1}^{N} \Pr_t(\xi_{pj}) \Big/ \sum_{q=1}^{N} F(\xi_{pq}) \right) .$$

Для определения $P^{(II)}_{t+1}(\xi)$ допустим, что хромосома $\xi \in \Xi$ может быть получена из любой пары ξ_{i_1}, ξ_{i_2}, такой что $i_1 \neq i_2$, в том случае, если выбранная позиция скрещивания λ является "благоприятной", то есть скрещивание в точке λ приведет к

получению ξ. В этом случае,

$$P_{t+1}^{(II)}(\xi) = \sum_{p=1}^{N_0^N}\Bigg(\sum_{i_1,i_2,\,i_1\neq i_2} \frac{F(\xi_{pi_1})}{\sum\limits_{j=1}^{N} F(\xi_{pj})} \cdot \frac{F(\xi_{pi_2})}{\sum\limits_{q=1}^{N} F(\xi_{pq}) - F(\xi_{pi_1})} \cdot \frac{n_\lambda(\xi \leftarrow \xi_{pi_1}, \xi_{pi_2})}{L-1} \cdot P_c \Bigg) \times$$

$$\times \quad \Pr(\mathfrak{G}_t = \mathfrak{u}_p) = \frac{P_c}{L-1} \sum_{p=1}^{N_0^N}\Bigg(\sum_{i_1,i_2,\,i_1\neq i_2} \frac{F(\xi_{pi_1})F(\xi_{pi_2})n_\lambda(\xi \leftarrow \xi_{pi_1}, \xi_{pi_2})}{\sum\limits_{j=1}^{N} F(\xi_{pj})\big(\sum\limits_{q=1}^{N} F(\xi_{pq}) - F(\xi_{pi_1})\big)} \Bigg) \prod_{r=1}^{N} \Pr_t(\xi_{pr}),$$

где $n_\lambda(\xi \leftarrow \xi_{pi_1}, \xi_{pi_2})$ – количество позиций скрещивания $\Lambda(\xi \leftarrow \xi_{pi_1}, \xi_{pi_2})$, "благоприятных" для получения хромосомы ξ из пары хромосом ξ_{pi_1}, ξ_{pi_2}. Формально,

$$\Lambda(\xi \leftarrow \xi_{pi_1}, \xi_{pi_2}) = \{\lambda |\, \xi_k = \xi_{pi_1 k},\, \xi_l = \xi_{pi_2 l}\},$$
$$1 \le \lambda < L;\; 1 \le k \le \lambda;\; \lambda + 1 \le l < L\,.$$

Таким образом, при одноточечном скрещивании

$$\Pr{}_{t+1}^{c}(\xi) = P_{t+1}^{(I)}(\xi) + P_{t+1}^{(II)}(\xi)\,. \tag{4}$$

6. Единичная мутация. Поскольку оператор мутации последовательно применяется к каждому генотипу в популяции, будем считать, что безусловное копирование и скрещивание генотипов ведет к образованию популяции $\mathfrak{G}_{t+1/2}$, которая, в свою очередь, преобразуется в $\mathfrak{G}_t$ в результате действия оператора мутации. Тогда

$$\Pr{}_{t+1}^{m}(\xi) = \begin{cases} \Pr{}_{t+1}^{c}(\xi), \text{ если } P_m = 0, \\ \Pr{}_{t+1}^{c}(\xi) + P_{t+1/2}^{(III)}(\xi) - P_{t+1/2}^{(IV)}(\xi), \text{ если } 0 < P_m \le 1\,. \end{cases}$$

где $\Pr{}_{t+1}^{c}(\xi)$ определяется из равенства (4), $P_{t+1/2}^{(III)}(\xi)$ и $P_{t+1/2}^{(IV)}(\xi)$ – соответственно приращение и убыль вероятности включения хромосомы $\xi \in \Xi$ в следующее поколение в результате действия оператора мутации.

Единичная мутация[1] хромосомы $\xi \in \Xi$ ведет к ее разрушению, поэтому

$$P_{t+1/2}^{(III)}(\xi) = \sum_{\mathfrak{u}_p\,:\,\xi \in \mathfrak{u}_p} (P_m(1-P_m)^{L-1} \cdot \#(\xi, \mathfrak{u}_p) \Pr(\mathfrak{G}_{t+1/2} = \mathfrak{u}_p)) =$$

$$= P_m(1-P_m)^{L-1} \cdot \sum_{\mathfrak{u}_p\,:\,\xi \in \mathfrak{u}_p} \left(\#(\xi, \mathfrak{u}_p) \prod_{j=1}^{N} \Pr_t(\xi_{pj}) \right).$$

В то же время, хромосома $\xi \in \Xi$ может быть создана вновь:

$$P_{t+1/2}^{(IV)}(\xi) = \sum_{p=1}^{N_0^N} \left(\sum_{j=1}^{N} n_\mu(\xi \leftarrow \xi_{pj}) P_m (1-P_m)^{L-1}/L \right) \Pr(\mathfrak{G}_{t+1/2} = \mathfrak{u}_p) =$$

$$= \frac{P_m(1-P_m)^{L-1}}{L} \cdot \sum_{p=1}^{N_0^N} \left(\sum_{j=1}^{N} n_\mu(\xi \leftarrow \xi_{pj}) \right) \prod_{q=1}^{N} \Pr_t(\xi_{pq})\,,$$

[1]Случай многократных мутаций рассматривается аналогично.

где $n_\mu(\xi \leftarrow \xi_{pj})$ – количество позиций единичной мутации $\mathrm{M}(\xi \leftarrow \xi_{pj})$, "благоприятных" для получения хромосомы ξ из хромосомы ξ_{pj}. Формально,

$$\mathrm{M}(\xi \leftarrow \xi_{pj}) = \{\mu |\, \xi_k = \xi_{pjk}, k \neq \mu \text{ и } \xi_k = \neg\xi_{pjk}, k = \mu\},\ 1 \leq k, \mu \leq L\,.$$

7. Вероятностные критерии качества. С учетом сказанного, вычислим значения представляющих интерес вероятностных критериев качества популяции:

1. *средней приспособленности хромосом:*

$$\overline{F}_t(\Xi) = \sum_{i=1}^{N_0} F(\xi_i)\,\Pr_t(\xi_i)\,;$$

2. *средней приспособленности популяций:*[2]

$$\overline{F}_t(\mathfrak{U}) = \sum_{p=1}^{N_0^N} \overline{F}_t(u_p)\,\Pr(\mathfrak{G}_t = u_p) = \sum_{p=1}^{N_0^N} \left(\sum_{j=1}^{N} F(\xi_{pj}) \prod_{q=1}^{N} \Pr_t(\xi_{pq}) \right) \Big/ N\,;$$

3. *средней меры давления отбора* – отношения приспособленности наилучшего генотипа популяции к средней приспособленности всех генотипов той же популяции:

$$\begin{aligned} \overline{\alpha}_t(\mathfrak{U}) &= \sum_{p=1}^{N_0^N} \frac{\max\limits_{j=1}^{N} F(\xi_{pj})}{\overline{F}_t(u_p)} \Pr(\mathfrak{G}_t = u_p) = \\ &= \sum_{p=1}^{N_0^N} \left(N \max_{j=1}^{N} F(\xi_{pj}) \prod_{q=1}^{N} \Pr_t(\xi_{pq}) \Big/ \sum_{r=1}^{N} F(\xi_{pr}) \right). \end{aligned}$$

Замечание 3. *Можно показать, что*

$$\overline{F}_t(\Xi) = \overline{F}_t(\mathfrak{U})\,. \tag{5}$$

При $N = 1$ доказательство равенства (5) не составляет труда:

$$\overline{F}_t(\mathfrak{U}) = \sum_{p=1}^{N_0^N} \overline{F}_t(u_p)\,\Pr(\mathfrak{G}_t = u_p) = \sum_{i=1}^{N_0} F(\xi_i)\,\Pr_t(\xi_i) = \overline{F}_t(\Xi)\,.$$

В случае $N > 1$ необходимо учитывать очевидное ограничение, накладываемое на значения вероятностей включения хромосом в популяцию очередного поколения: $\sum_{i=1}^{N_0} \Pr_t(\xi_i) \equiv 1$.

---

[2]Средняя приспособленность каждой популяции определяется средним арифметическим приспособленности хромосом, входящих в ее состав, то есть так же, как в теореме Холланда.

Список литературы

[1] Петров А.В. К вопросу о формализации базовых операторов генетических алгоритмов // Вестник Твер. гос. техн. ун-та. Тверь: ТГТУ, 2003. Вып. 2. С. 67–69.

[2] Петров А.В. Формальное введение в генетические алгоритмы // Сложные системы: обработка информации, моделирование и оптимизация: Сб. науч. тр. Вып. 2. Тверь: ТвГУ, 2004. С. 78–93.

[3] Altenberg, L. The Schema Theorem and Price's Theorem. Institute of Statistics and Decision Sciences, Duke University. April 1994.

[4] Whitley D. A Genetic Algorithm Tutorial. Technical Report CS–93–103, Colorado State University, March 1993.