

## 2. СИМВОЛЬНАЯ МОДЕЛЬ И ИНТЕРПРЕТАЦИЯ ЕЕ ЭЛЕМЕНТОВ В ТЕРМИНАХ ПОПУЛЯЦИОННОЙ ГЕНЕТИКИ

### 2.1. Представление допустимых решений экстремальной задачи в виде бинарных строк

Допустимое решение  $\vec{x} \in D$  экстремальной задачи однокритериального выбора (1.3) является  $n$ -мерным вектором  $\vec{x} = (x_1, \dots, x_n)$ . В том случае, когда задача (1.3) принадлежит классу задач переборного типа, имеется конечное множество допустимых решений, в которых каждая компонента  $x_i, i = \overline{1, n}$  вектора  $\vec{x} \in D$  может быть закодирована с помощью целого неотрицательного числа:

$$\beta_i \in [0, K_i], i = \overline{1, n}, \quad (2.1)$$

где  $(K_i+1)$ - число возможных дискретных значений  $i$ -ой управляемой переменной в области поиска  $D$ . Это позволяет поставить во взаимнооднозначное соответствие каждому вектору  $\vec{x} \in D$  вектор  $\vec{\beta}$  с целочисленными компонентами:

$$(x_1, \dots, x_n) \leftrightarrow (\beta_1, \dots, \beta_n), \quad (2.2)$$

где для каждой компоненты  $\beta_i, i = \overline{1, n}$  областью возможных значений являются целые числа от 0 до  $K_i$ .

Введем алфавит  $B_2$ , содержащий только два символа 0 и 1:  $B_2 = \{0, 1\}$ . Для того, чтобы представить целочисленный вектор  $\vec{\beta} = (\beta_1, \dots, \beta_n)$  в алфавите  $B_2$  необходимо определить максимальное число двоичных символов  $\theta$ , которое достаточно для представления в двоичном коде любого значения  $\beta_i$  из области его допустимых значений  $[0, K_i]$ . Нетрудно видеть, что параметр символьной модели  $\theta$  должен удовлетворять неравенству:

$$K < 2^\theta, \quad (2.3)$$

где  $K = \max_{1 \leq i \leq n} (K_i)$ .

Запись произвольного целого неотрицательного числа  $\beta_i = (0 \leq \beta_i < 2^\theta)$  с помощью  $\theta$  двоичных символов определяется соотношением :

$$\beta_i = \sum_{l=1}^{\theta} \alpha_l 2^{\theta-l}, \quad (2.4)$$

где  $\alpha_l$  - двоичное число, равное 0 или 1;

$\theta$ -длина двоичного слова, кодирующего целое число  $\beta_i$ .

Тогда символьная запись целочисленного кода  $\beta_i$  для фиксированного значения управляемой переменной  $x_i$  в обычном двоичном коде запишется в виде следующей бинарной комбинации:

$$e_\theta(\beta_i): \quad \begin{array}{|c|c|c|c|} \hline \alpha_1 & \alpha_2 & \dots & \alpha_\theta \\ \hline \end{array} \quad (2.5)$$

←  $\theta$  →

где  $\alpha_i, i = \overline{1, \theta}$  - двоичные символы (0 или 1), полученные из соотношения (2.4).

Пример 2.1.

Пусть  $\theta=5$  и  $\beta_i=19$ . Тогда согласно соотношения (2.4) можем записать:

$$\begin{aligned} 19 &= 1 \times 2^{5-1} + 0 \times 2^{5-2} + 0 \times 2^{5-3} + 1 \times 2^{5-4} + 1 \times 2^{5-5} = \\ &= 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0. \end{aligned}$$

Следовательно, бинарная комбинация  $e_5(19)$  целого числа 19 в алфавите  $B_2$  будет иметь следующий вид:

$$e_5(19): \quad \begin{array}{c} 2^4 \quad 2^3 \quad 2^2 \quad 2^1 \quad 2^0 \\ \begin{array}{|c|c|c|c|c|} \hline 1 & 0 & 0 & 1 & 1 \\ \hline \end{array} \end{array}$$

←  $\theta$  →

Для представления допустимого решения  $\bar{x} \in D$  экстремальной задачи (1.3) в алфавите  $B_2$  объединим символьные записи  $e_\theta(\beta_i)$ , описывающие все  $n$  компонент вектора  $\bar{x}$ , в виде линейной последовательности из бинарных комбинаций (2.5):

$$E(\bar{x}) = (e_\theta(\beta_1), \dots, e_\theta(\beta_n)). \quad (2.6)$$

Записи (2.6) соответствует  $(n \times \theta)$ -битовая строка из двоичных символов (0,1):

$$E(\bar{x}); \quad \begin{array}{|c|c|c|c|c|c|c|c|c|c|} \hline e_\theta(\beta_1) & \dots & e_\theta(\beta_2) & \dots & e_\theta(\beta_n) & \dots & & & & \\ \hline \alpha_1^1 & \dots & \alpha_{\theta}^1 & \alpha_1^2 & \dots & \alpha_{\theta}^2 & \dots & \alpha_1^n & \dots & \alpha_{\theta}^n \\ \hline \end{array} \quad (2.7)$$

←  $n \times \theta$  →

Таким образом, *символьная модель экстремальной задачи переборного типа (1.3) может быть представлена в виде множества бинарных строк (2.7), которые описывают конечное множество допустимых решений  $\bar{x}$ , принадлежащих области поиска D.*

Необходимо отметить, что выбор символьной модели исходной экстремальной задачи во многом определяет эффективность и качество применяемых генетических алгоритмов. Для каждого класса задач переборного типа должна строиться своя символьная модель, отражающая специфику и особенности решаемой задачи. В качестве примера приведем символьную модель для задачи (1.12) оптимального дихотомического разбиения графа  $G(X, V, W)$ .

Представим дихотомическое разбиение  $(X_1, X_2)$  графа  $G(X, V, W)$  порядка  $n$  в виде бинарной строки  $E(X_1, X_2)$ , состоящей из  $n$  бит, расположенных в порядке возрастания их номеров. Каждому номеру бита поставим в взаимнооднозначное соответствие номер вершины графа (1-ый бит соответствует вершине  $x_1$ , 2-ой бит - вершине  $x_2$ , ...,  $n$ -ый бит - вершине  $x_n$ ). Потребуем, чтобы бинарное значение  $\alpha_1$  1-ого бита указывало, какому подмножеству вершин ( $X_1$  или  $X_2$ ) принадлежит вершина  $x_1$ :

$$\alpha_1 = \begin{cases} 1, & \text{если 1-ая вершина } x_1 \in X \text{ входит в состав} \\ & \text{подмножества вершин } X_1; \\ 0, & \text{если 1-ая вершина } x_1 \in X \text{ входит в состав} \\ & \text{подмножества вершин } X_2. \end{cases} \quad (2.8)$$

При этом каждая бинарная строка  $E(X_1, X_2)$  должна удовлетворять дополнительному требованию, связанному с сутью дихотомического разбиения: *“число битов, содержащих “1” в бинарной строке  $E(X_1, X_2)$ , должно равняться мощности подмножества вершин подграфа  $G_1(X_1, V_1, W_1)$ , равной порядку этого подграфа  $n_1$ ”*.

Так, разбиения  $(X_1, X_2)$  и  $(X_1^*, X_2^*)$ , приведенные в Таблице 1.1., имеют следующие представления в виде бинарных строк:

$E(X_1, X_2):$	1	0	0	0	0	0	1	1	0	1	1	0
$E(X_1^*, X_2^*):$	0	1	1	1	1	1	0	0	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12
	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$	$x_{11}$	$x_{12}$
	-номер бита											
												-номер
												вершины

Сравнивая построенную символьную модель экстремальной задачи (1.12) с общей символьной моделью (2.7), видим, что допустимый вектор  $\bar{X}$  включает в качестве компонент все вершины графа  $G$ , каждой из которых соответствует целое число  $\beta_i$ , принимающее только два значения 0 или 1 (т.е.  $K_i=1$  для всех  $i = \overline{1, n}$ ).

Это приводит к тому, что бинарная комбинация  $e_\theta(\beta_i)$  состоит из единственного бита, т.к. неравенство (2.3) выполняется при  $\theta=1$ . Однако, линейная последовательность (2.6) принимается в качестве бинарной строки  $E(\bar{X})$ , соответствующей допустимому разбиению  $(X_1, X_2)$ , только в том случае, если число “1” в ней равно порядку  $n_1$  графа  $G_1$ .

## 2.2. Особи и их вариабильные признаки

Наименьшей неделимой единицей биологического вида, подверженной действию факторов эволюции, является *особь*  $a_k^t$  (индекс  $k$  обозначает номер особи, а индекс  $t$  - некоторый момент времени эволюционного процесса). В качестве аналога особи  $a_k^t$  в экстремальной задаче однокритериального выбора (1.3) примем произвольное допустимое решение  $\bar{X} \in D$ , которому присвоено имя  $a_k^t$ . Действительно, вектор управляемых переменных  $(x_1, \dots, x_n)$  - это наименьшая неделимая единица, характеризующая в экстремальной задаче (1.3) внутренние параметры на каждом  $t$ -ом шаге поиска оптимального решения, которые изменяют свои значения в процессе минимизации критерия оптимальности  $Q(\bar{X})$ .

В задаче оптимального дихотомического разбиения (1.12) в качестве особи  $a_k^t$  выступает конкретное дихотомическое разбиение  $(X_1, X_2)$ , удовлетворяющее условиям (1.8) - (1.9), что позволяет интерпретировать сам процесс решения экстремальной задачи (1.12) как эволюционный процесс, связанный с перераспределением вершин  $x_i \in X$  графа  $G$  по двум подграфам  $G_1$  и  $G_2$ , соответственно, порядка  $n_1$  и  $n_2$ , с целью отыскания глобального минимума критерия оптимальности (1.11). В этом и заключается в данном случае цель эволюционного развития (эволюции) особей.

Для описания особей введем два типа *вариабельных признаков*, отражающих качественные и количественные различия между особями в степени их выраженности:

*качественные признаки* - признаки, которые позволяют однозначно разделять совокупность особей на четко различимые группы;

*количественные признаки* - признаки, проявляющие непрерывную изменчивость, в связи с чем степень их выраженности можно охарактеризовать числом.

Качественные признаки особи  $a_k^t$  определяются из символьной модели экстремальной задачи (1.3) как соответствующая точке  $\bar{X}$  с именем  $a_k^t$  бинарная строка  $E(\bar{X})$  и составляющие ее бинарные комбинации  $e_\theta(\beta_1), \dots, e_\theta(\beta_n)$ .

Приведем интерпретацию этих признаков в терминах хромосомной теории наследственности [4].

В качестве *гена* - единицы наследственного материала, ответственного за формирование альтернативных признаков особи, примем бинарную комбинацию  $e_\theta(\beta_i)$  из (2.5), которая определяет фиксированное значение целочисленного кода  $\beta_i$  управляемой переменной  $x_i$  в обычном двоичном коде. Одна особь  $a_k^t$  будет характеризоваться  $n$  генами, каждый из которых отвечает за формирование целочисленного кода соответствующей управляемой переменной. Тогда структуру бинарной строки  $E(\bar{X})$  из (2.7) можно интерпретировать *хромосомой*, содержащей  $n$  сцепленных между собой генов, которые расположены в линейной последовательности “слева - направо”. Согласно хромосомной теории наследственности передача качественных признаков  $e_\theta(\beta_i)$ ,  $i = \overline{1, n}$ , закодированных в генах, будет осуществляться через хромосомы от “родителей” к “потомкам”.

Местоположение определенного гена в хромосоме называется *локусом*, а альтернативные формы одного и того же гена, расположенные в одинаковых локусах хромосомы, называются *аллелями (аллелеформами)*:

ген 1	ген 2	...	ген n		(2.9)
$E(\bar{X}); e_\theta(\beta_1)$	$e_\theta(\beta_2)$	...	$e_\theta(\beta_n)$	,	
локус 1	локус 2		локус n		

хромосома

→

где  $e_\theta(\beta_i)$  - аллель  $i$ -го гена, находящаяся в локусе  $i$ .

Хромосому (2.9), содержащую в своих локусах конкретные значения аллелей, будем называть *генотипом (генетическим кодом)*  $E(a_k^t)$ , который содержит всю наследственную генетическую информацию об особи  $a_k^t$ , получаемую от “предков” и передаваемую затем “потомкам”. Конечное множество всех допустимых генотипов образует *генофонд*. Для дихотомического разбиения мощность генофонда равна  $C_n^{n_1}$ .

При взаимодействии особи  $a_k^t$  с внешней средой ее генотип  $E(a_k^t)$  порождает совокупность внешне наблюдаемых количественных признаков (характеристик  $\varphi_i$ ),

включающих *степень приспособленности*  $\mu(a_k^t)$  особи  $a_k^t$  к внешней среде и ее *фенотип*  $\phi(a_k^t)$ .

Приняв в качестве внешней среды критерий оптимальности  $Q(\bar{x})$ , мы можем говорить, что степенью приспособленности  $\mu(a_k^t)$  каждой особи  $a_k^t$  является численное значение функции  $Q(\bar{x})$ , вычисленное для допустимого решения  $\bar{x} \in D$  с именем  $a_k^t$ . В общем случае степень приспособленности  $\mu(a_k^t) \geq 0$  можно задать с помощью следующего выражения:

$$\mu(a_k^t) = \begin{cases} Q^2(x), \text{ если решается задача максимизации} \\ \text{функции } Q(\bar{x}); \\ 1/(Q^2(\bar{x})+1), \text{ если решается задача минимизации} \\ \text{функции } Q(\bar{x}); \end{cases} \quad (2.10)$$

Из выражения (2.10) следует, что чем больше численное значение степени приспособленности  $\mu(a_k^t)$ , тем лучше особь  $a_k^t$  приспособлена к внешней среде. Следовательно, цель эволюции особей заключается в повышении их степени приспособленности.

Фенотипом  $\phi(a_k^t)$  особи  $a_k^t$  в рамках экстремальной задачи (1.3) являются численные значения вектора управляемых переменных  $\bar{x} \in D$  и соответствующих ему характеристик  $\phi_i(\bar{x}), i = \overline{1, s}$ .

Для задачи оптимального дихотомического разбиения графа  $G$ , сформулированной как экстремальная задача (1.18), в качестве особи  $a_k^t$  выступает конкретное дихотомическое разбиение  $(X_1, X_2)$ , удовлетворяющее условиям (1.8)- (1.9). В этом случае геном является бит в бинарной строке  $E(X_1, X_2)$ , который определяет, к какой части разбиения  $X_1$  или  $X_2$  принадлежит вершина графа  $G$ , соответствующая этому биту. Линейная последовательность всех  $n$  битов составляет хромосому, в которой каждый ген определяет принадлежность вершины, соответствующей этому гену, одной из частей  $X_1$  или  $X_2$ . Введенные гены обладают свойством *диморфизма*, т.к. каждый ген может иметь только две различающиеся формы аллели: “1”, если вершина  $x_i$  принадлежит части  $X_1$  и “0”, если вершина  $x_i$  принадлежит части  $X_2$ .

Степень приспособленности  $\mu(a_k^t)$  в данном случае просто совпадает с критерием оптимальности  $F(X_1, X_2)$  - общей суммой весов ребер, входящих в подграфы  $G_1$  и  $G_2$ :  $\mu(a_k^t) = F(X_1, X_2)$ .

В состав фенотипа  $\phi(a_k^t)$  особи  $a_k^t$ , кроме разбиения  $(X_1, X_2)$ , входят следующие количественные признаки:

вес разреза  $Q(X_1, X_2)$  из (1.11); коэффициент разбиения  $K(X_1, X_2)$  из (1.13); сумма весов ребер подграфа  $G_1$   $f_1(X_1)$  из (1.16); сумма весов ребер подграфа  $G_2$   $f_2(X_2)$  из (1.17).

### 2.3. Популяции и поколения

В качестве *ареала* - области, в пределах которой только и могут встречаться особи, участвующие в эволюционном процессе, будем рассматривать область поиска  $D$ . В задаче дихотомического разбиения ареал полностью определяется структурой графа  $G(X, V, W)$ , заданной множеством вершин  $X$  и множеством ребер  $V$ , а также порядком подграфа  $G_1$  (или подграфа  $G_2$ ).

Совокупность особей  $(a_1^t, \dots, a_v^t)$ , принадлежащих ареалу, образует *популяцию*  $P^t$ . Число  $v$ , характеризующее число особей  $a_k^t$ , которые образуют популяцию, будем называть *численностью популяции*. В общем случае экстремальной задачи (1.3) популяция  $P^t = (a_1^t, \dots, a_v^t)$  соответствуют совокупности допустимых решений  $\bar{x}^k \in D$ ,  $k = \overline{1, v}$ . Для задачи оптимального разбиения графа  $G$  популяция  $P^t$  представляет собой набор из  $v$  дихотомических разбиений  $(X_1^k, X_2^k)$ ,  $k = \overline{1, v}$ , удовлетворяющих условиям (1.8) - (1.9).

Очевидно, что в популяции  $P^t$  может иметь место наличие нескольких различающихся форм того или иного варибельного признака (так называемый *полиморфизм*), что позволяет проводить разделение популяции на ряд *локальных популяций*  $P_i^t \subset P^t$ ,  $i = \overline{1, k}$ , включающих в свой состав те особи, которые имеют одинаковые или “достаточно близкие” формы тех или иных качественных или/и количественных признаков.

Так, в задаче оптимального дихотомического разбиения (1.11) для дифференциации особей  $a_k^t \in P^t$  по количественному признаку может быть выбрано, например, условие, что в локальную популяцию  $P_1^t \subset P^t$  включаются только те особи, у которых значение веса разреза  $Q(X_1, X_2)$  не превосходит некоторой заданной величины  $Q^+$ :  $Q(X_1, X_2) \leq Q^+$ . Тогда другую локальную популяцию  $P_2^t \subset P^t$  составят все те особи  $a_k^t$ ,

которые не попали в  $P_1^t$ , т.е. особи, для которых вес разреза удовлетворяет условию:  $Q(X_1, X_2) > Q^+$ .

В том случае, когда для дифференциации особей  $a_k^t \in P^t$  используется качественный признак, например, генотип  $E(X_1, X_2)$ , в качестве меры “близости” особей  $a_k^t$  и  $a_l^t$  по этому признаку можно использовать *Хэммингово расстояние*, которое определяется как число несовпадающих по своим значениям битов в  $n \times \theta$ -битовых бинарных строках  $E(X_1^k, X_2^k)$  и  $E(X_1^l, X_2^l)$ :

$$d[E(X_1^k, X_2^k), E(X_1^l, X_2^l)] = |E(X_1^k, X_2^k) \oplus E(X_1^l, X_2^l)|, \quad (2.11)$$

где  $\oplus$ - операция суммирования по mod.2. Тогда в локальную популяцию  $P_1^t \subset P^t$  будем включать только те особи, у которых Хэммингово расстояние меньше заданного неотрицательного целого числа  $\delta \geq 0$ , а в локальную популяцию  $P_2^t \subset P^t$  - те особи, для генотипов которых это условие не выполняется. При  $\delta=0$  в локальную популяцию  $P_1^t$  будут включены только те особи, генотипы которых совпадают между собой.

Будем считать, что во времени популяции  $P^t$  состоят из дискретных, неперекрывающихся между собой *поколений*, - групп особей, одинаково отдаленных в родственном отношении от общих предков, т.е. каждое последующее поколение  $P^{t+1}$  является совокупностью из  $v$  особей, которые отбираются только из особей предыдущего  $t$ -го поколения. Будем отождествлять номер поколения (верхний индекс  $t$  в обозначениях особи  $a_k^t$  и популяции  $P^t$ ) с моментом времени  $t=0, 1, \dots, T$ , где  $T$  - *жизненный цикл популяции*, определяющий период ее эволюции.

В дальнейшем *эволюцию популяции*  $P^t$  будем понимать в ограниченном смысле как чередование поколений, в процессе которого особи изменяют свои переменные признаки таким образом, чтобы каждая следующая популяция проявляла лучшую степень приспособленности к внешней среде, например, в смысле обеспечения наибольшего значения *средней степени приспособленности по популяции*  $P^t$ :

$$\mu_{cp}(t) = \frac{1}{v} \sum_{i=1}^v \mu(a_i^t). \quad (2.12)$$

Совокупность из  $v$  генотипов всех особей  $a_k^t$ , составляющих популяцию  $P^t$ , образует *хромосомный набор*, который полностью содержит в себе генетическую информацию о популяции  $P^t$  в целом. Наличие изменчивости хромосомного набора от



поколения к поколению является необходимым условием эволюции популяции  $P^t$  на генетическом уровне. Для оценки разнообразия генотипов популяции  $P^t$  введем в рассмотрение *функцию диаллейного разнообразия* по каждому биту хромосомного набора:

$$D_i = 1 - 4 \times \left[ 0.5 - \frac{v_i}{v} \right]^2, i = \overline{1, n \times \theta}, \quad (2.13)$$

где  $v_i$  - число нулей в  $i$ -ом бите хромосомного набора популяции  $P^t$ ;  $v$  - численность популяции  $P^t$ . Тогда *побитовое разнообразие* популяции  $P^t$  определим как среднее значение диаллельных разнообразий по всем  $(n \times \theta)$  битам хромосомного набора:

$$D_{B(t)} = \frac{1}{n \times \theta} \sum_{i=1}^{n \times \theta} D_i. \quad (2.14)$$

При  $D_{B(t)}=1$  имеем максимальное разнообразие генотипов в популяции  $P^t$ ; при  $D_{B(t)}=0$  все генотипы в хромосомном наборе совпадают между собой.

Обобщением побитового разнообразия на общий случай экстремальной задачи (1.3) является *генетическое разнообразие популяции  $P^t$  по всем  $n$  локусам*:

$$D_\lambda = \frac{1}{n} \sum_{i=1}^n D_\lambda(i), \quad (2.15)$$

где

$$D_\lambda(i) = 1 - \frac{v^2}{(1-v)^2} \left( \frac{1}{v} - \max_{1 \leq k \leq m_i} P(\hat{a}_0(k), i) \right)^2 \quad (2.16)$$

- *функция аллельного разнообразия в  $i$ -ом локусе*;

$P(\hat{a}_0(k), i) = \frac{v_i}{v}$  - частота аллельной формы  $e_0(k)$  в  $i$ -ом локусе;

$v_i$  - число генотипов в хромосомном наборе популяции  $P^t$ , в которых  $i$ -ый локус содержит аллельную форму  $e_0(k)$ ;

$v$  - численность популяции  $P^t$ ;

$m_i$  - число форм аллелей в  $i$ -м локусе ( $1 \leq m_i \leq v$ ).

Когда все  $v$  генотипов имеют в  $i$ -м локусе одну и ту же аллельную форму  $\hat{a}_0(k)$  ( $v_i = v$ )  $D_\lambda(i)=0$ ; если аллельные формы в  $i$ -м локусе всех генотипов хромосомного набора отличаются друг от друга ( $v_i=1$ ), то  $D_\lambda(i)=1$ .

По хромосомному набору популяции  $P^t$  можно также определить частоту генотипа  $P(E(\bar{X}))$  как долю особей, имеющих одинаковую форму генотипа в рассматриваемой популяции  $P^t$ .