УДК 32.81 ***Информатика***

Статья представлена академиком С.К. Коровиным

В.Г. Редько, Ю.Р. Цой

ОЦЕНКА ЭФФЕКТИВНОСТИ ЭВОЛЮЦИОННЫХ АЛГОРИТМОВ

В работе исследуется простая модель эволюции популяции информационных последовательностей.

Описание модели. Основные предположения модели состоят в следующем.

1. Рассматривается эволюция популяции «особей» $\{\mathbf{S}_k\}$, каждая особь $\mathbf{S}_k$ определяется последовательностью символов S_{ki}, символы принимают два значения: $S_{ki} = +1$ либо -1; $i = 1,2,\ldots,N$; $k = 1,2,\ldots,n$; N – длина последовательностей; n – численность популяции. Последовательность $\mathbf{S}_k$ можно рассматривать как модельную ДНК k-й особи.

2. На множестве последовательностей $\mathbf{S}$ определена функция приспособленности $f(\mathbf{S})$ следующим образом. Предполагается, что имеется оптимальная последовательность $\mathbf{S}_m$, а приспособленность произвольной особи $\mathbf{S}$ экспоненциально уменьшается с ростом расстояния по Хеммингу $\rho(\mathbf{S}, \mathbf{S}_m)$ между $\mathbf{S}$ и $\mathbf{S}_m$ (числа несовпадающих символов в соответствующих позициях этих последовательностей):

$$f(\mathbf{S}) = \exp[-\beta\rho(\mathbf{S}, \mathbf{S}_m)], \qquad (1)$$

где β – параметр интенсивности отбора.

3. Эволюционный процесс состоит из ряда поколений, в каждом поколении происходят а) отбор особей в следующее поколение в соответствии с их приспособленностями $f(\mathbf{S})$ и б) мутации – случайные замены символов S_{ki} .

4. Длина последовательностей N и численность популяции n в отдельном эволюционном процессе не меняются и велики: $N, n >> 1$.

Формальное описание схемы эволюции представлено в Таблице 1.

Таблица 1. Схема эволюции

Шаг 0. *Формирование начальной популяции* $\{\mathbf{S}_k\ (0)\}$. Для каждого $k = 1,...,n$ и каждого $i = 1,...,N$ выбираем случайно символ S_{ki} , полагая его равным +1 либо -1.
Шаг 1. *Отбор*
Подшаг 1.1. *Расчет приспособленностей*. Для популяции $\{\mathbf{S}_k(t)\}$ для каждого $k = 1,...,n$ вычисляем величину $f(\mathbf{S}_k)$, t – номер поколения.
Подшаг 1.2. *Формирование новой популяции* $\{\mathbf{S}_k\ (t+1)\}$. Отбираем n особей в новую популяцию $\{\mathbf{S}_k(t+1)\}$ с вероятностями, пропорциональными $f(\mathbf{S}_k)$.
Шаг 2. *Мутации*. Для каждого $k = 1,...,n$ и каждого $i = 1,...,N$ меняем знак $S_{ki}(t+1)$ на противоположный с вероятностью P; P – интенсивность мутаций.
Организация последовательности поколений. Повторяем шаги 1, 2 для $t = 1, 2,\ldots$

Подшаг 1.2 требует пояснения. Формирование новой популяции происходит следующим образом. Представим, что у нас есть рулетка. Для каждого

поколения размечаем рулетку на n секторов, долю k-го сектора (отнесенную ко всей площади круга) полагаем равной $q_k = f_k\,[\Sigma_l\, f_l\,]^{-1}$; $f_k = f(\mathbf{S}_k)$. Далее n раз крутится рулетка, каждый раз определяется номер сектора, на котором останавливается стрелка рулетки, и соответствующая этому номеру особь выбирается в популяцию следующего поколения. Таким образом, в следующее поколение будут отобраны ровно n особей. При этом для каждого вращения рулетки вероятность k-й особи попасть в следующее поколение пропорциональна ее приспособленности f_k. Некоторые особи могут быть отобраны в новое поколение несколько раз, – это означает, что в новой популяции будет несколько потомков данной особи.

Изложенная модель соответствует схеме модели квазивидов, предложенной в 1970-х годах М. Эйгеном [1,2] в качестве одной из моделей предбиологической эволюции, и схеме генетического алгоритма без скрещивания [3].

Стохастический характер эволюционного процесса. Далее предполагается, что $2^N >> n$, т.е. предполагается, что длина N последовательностей **S** (модельных ДНК) достаточно велика. При этом число последовательностей отдельных видов в популяции будет невелико, а многие виды вообще будут отсутствовать в популяции. В силу этого существенны флуктуации числа видов, и рассматриваемые эволюционные процессы имеют стохастический характер. В частности, необходимо учитывать нейтральный отбор, т.е. фиксацию особей, независящую от их приспособленностей [4,5].

Роль нейтрального отбора охарактеризуем следующей эволюционной игрой:

1. Имеется популяция черных и белых шаров, общее количество шаров в популяции равно n.

2. Эволюция состоит из последовательности поколений. Каждое поколение состоит из двух шагов. На первом шаге все шары дублируются с сохранением цвета: черный шар имеет два черных потомка, белый шар имеет два белых потомка. На втором шаге случайным образом удаляется из популяции ровно половина шаров независимо от их цвета.

Эта игра представляет собой марковский процесс, для которого показано [4], что: 1) рассматриваемый процесс сходится к одному из двух поглощающих состояний (в популяции остаются либо только белые шары, либо только черные шары); 2) при больших n характерное число поколений T_n, требуемое для сходимости к какому-либо из поглощающих состояний, равно $2n$. Для дальнейшего существенно, что характерное время нейтрального отбора T_n порядка численности популяции n.

Качественная картина эволюции. Результаты компьютерного моделирования демонстрируют, что при достаточно малой интенсивности мутаций ($1 \geq PN$, $\beta \geq PN$) эволюцию можно охарактеризовать следующим образом (пример расчета представлен на рис. 1):

- начальное распределение по ρ в популяции (при t =0) близко к нормальному распределению со средним $<\rho> = N/2$ и дисперсией $N/4$ [4,6] ($<\rho>$ - среднее по популяции расстояние по Хеммингу до оптимальной последовательности $\mathbf{S}_m$);

- процесс эволюции можно характеризовать двумя стадиями: первой – быстрой и второй – медленной; на первой стадии происходит отбор особей, расположенных на левом крыле начального распределения, и распределение сжимается; на второй стадии распределение смещается к малым значениям ρ;

- окончательное распределение формируется в окрестности оптимальной последовательности $\mathbf{S}_m$.

Аналитическая оценка. Оценим эффективность рассматриваемого алгоритма, предполагая, что численность популяции n достаточно велика:

$$T_n \geq T, \quad [1 - (1 - P)^N]^n \ll 1, \qquad (2)$$

где T_n – характерное время нейтрального отбора ($T_n \sim n$), T – характерное время сходимости всего эволюционного процесса. Первое неравенство в (2) означает, что влияние нейтрального отбора не велико. Второе неравенство соответствует пренебрежению мутационными потерями уже найденных удачных особей в популяции.

Оценим величину T, которая определяется второй (медленной) стадией. На этой стадии новые особи с меньшими значениями ρ появляются в результате мутаций и фиксируются в популяции за счет отбора. Оценим характерное время t_{-1}, за которое $<\rho>$ уменьшается на 1. Оно составляет: $t_{-1} \sim t_m + t_s$, где $t_m \sim (NP)^{-1}$ – характерное время, за которое особи популяции промутируют, $t_s \sim \beta^{-1}$ – характерное время, за которое особи, для которых $\rho = <\rho>-1$, в результате отбора вытеснят из популяции особей, для которых $\rho = <\rho>$. Полагая $T \sim t_{-1} N$, имеем:

$$T \sim (P)^{-1} + N\beta^{-1}. \qquad (3)$$

Общее число особей, участвующих в эволюции, составляет $n_{total} = nT$. Оценим величину n_{total} для заданного N, выбирая остальные параметры β, P, n так, чтобы по возможности минимизировать величину n_{total}. Интенсивность отбора считаем достаточно большой: $\beta \geq PN$, тогда можно пренебречь вторым слагаемым в (3). Полагаем $P \sim N^{-1}$, при такой интенсивности мутаций, с одной стороны, появление новых особей в популяции в результате мутаций происходит достаточно быстро и, с другой стороны, можно пренебречь мутационными потерями (выполняется второе неравенство в (2)). Тогда имеем $T \sim N$. Также полагаем, что первое из неравенств (2) выполняется на пределе: $n \sim T_n \sim T \sim N$, т.е. предполагаем минимальную допустимую численность популяции, при которой еще не существенны потери удачных особей в результате нейтрального отбора. С учетом сделанных предположений имеем:

$$T \sim N, \; n_{total} \sim N^2 . \qquad (4)$$

Результаты компьютерного моделирования. Для проверки оценок (4) проведены компьютерные расчеты для соотношений между параметрами, соответствующих условиям получения аналитических оценок: $n = N$, $P = N^{-1}$, $\beta = 1$. Расчет выполнялся следующим образом. Были получены зависимости среднего по популяции расстояния до оптимума от времени $\langle\rho\rangle(t)$ для разных значений N (рис. 2), и по этим зависимостям оценивалось характерное время сходимости эволюции T двумя способами: 1) рассчитывалось характерное время релаксации T_R в зависимостях $\langle\rho\rangle(t)$ по начальному наклону этих

кривых, 2) определялось время выхода T_S на стационарное значение $<\rho>$, которое получается при больших t (см. рис. 2). Полученные в результате зависимости T_R (N) и T_S (N) представлены на рис. 3. Кроме того, определялись значения времени T_O первого появления оптимальной последовательности $\mathbf{S}_m$ в популяции. Соответствующая зависимость $T_O(N)$ также представлена на рис. 3. Видно, что при достаточно больших N все три зависимости линейны: T_R (N) = $k_R N + T_{R0}$, T_S (N) = $k_S N + T_{S0}$, $T_O(N) = k_O N + T_{O0}$, где k_R = 0,1772, k_S = 0,3903, k_O = 0,3685, T_{R0} = 8,2709, T_{S0} = 38,7356, T_{O0} = 2,1288, что вполне согласуется с оценками (4).

Обсуждение. Сравним эволюционный метод оптимизации рассматриваемой функции приспособленности (1) с двумя простейшими методами: последовательный поиск и случайный перебор.

Последовательный поиск организуем следующим образом. Исходим из произвольной последовательности $\mathbf{S}$, символы которой равны S_i = 1 либо -1. Далее последовательно для каждого i (i = 1,2,…,N) меняем знак символа ($S_i \rightarrow - S_i$) и при увеличении приспособленности $f(\mathbf{S})$ принимаем новое значение символа, при уменьшении $f(\mathbf{S})$ – возвращаемся к старому. В результате после N испытаний найдем оптимальную последовательность $\mathbf{S}_m$. Итак, для последовательного поиска имеем: $n_{total} = N$.

При случайном переборе для нахождения оптимальной последовательности необходимо испытать порядка 2^N последовательностей: $n_{total} \sim 2^N$.

Полученные оценки приведены в таблице 2.

Таблица 2. Оценки эффективности методов поиска.

Метод поиска	n_{total}	n_{total} при $N = 1000$
Последовательный	N	1000
Эволюционный	$\sim N^2$	$\sim 10^6$
Случайный	$\sim 2^N$	$\sim 10^{300}$

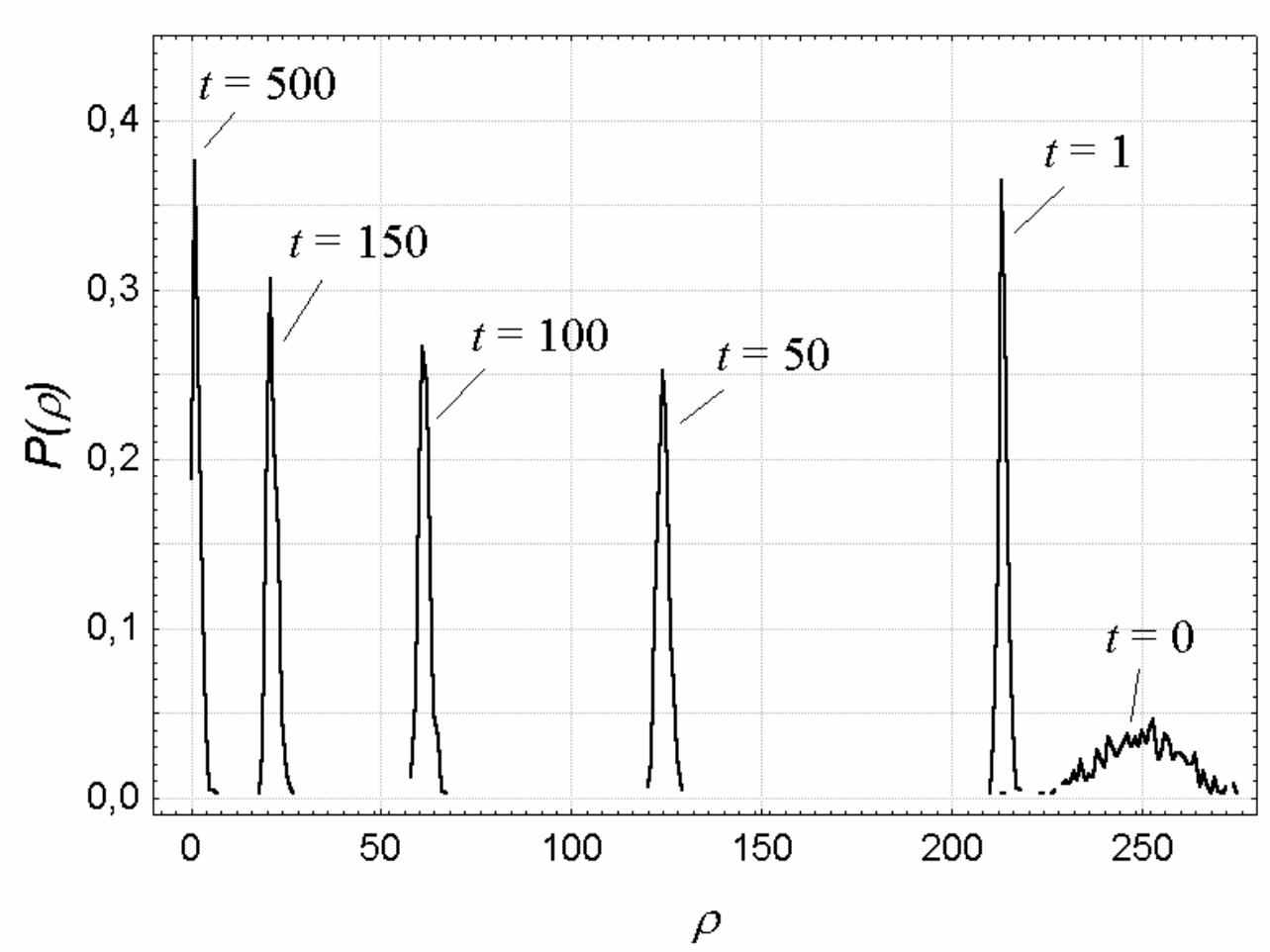
Отметим, что хотя здесь оценки получены для случая унимодальной функции приспособленности (1), аналогичные оценки могут быть сделаны и для спин-стекольной модели эволюции, когда число локальных максимумов приспособленности экспоненциально растет с размерностью задачи N [4,7].

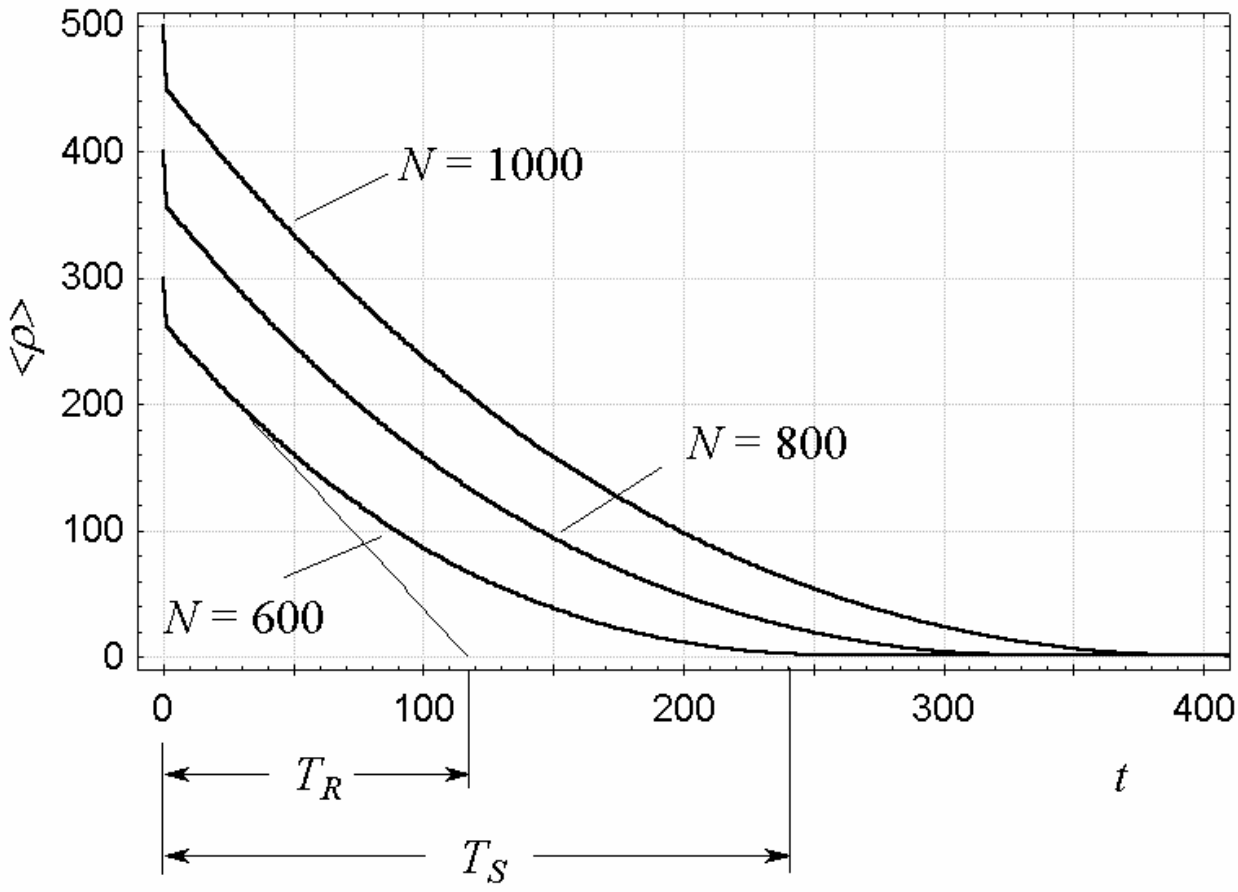
Полученные оценки демонстрируют, что эволюционный процесс, как алгоритм оптимизации «субоптимален»: он не обеспечивает максимальную скорость поиска (для конкретной задачи возможны более эффективные алгоритмы, такие как последовательный поиск для рассматриваемого случая), тем не менее, он намного эффективнее случайного перебора. А так как эволюционный метод поиска прост и универсален, то он может рассматриваться как хороший эвристический метод оптимизации для широкого класса задач.

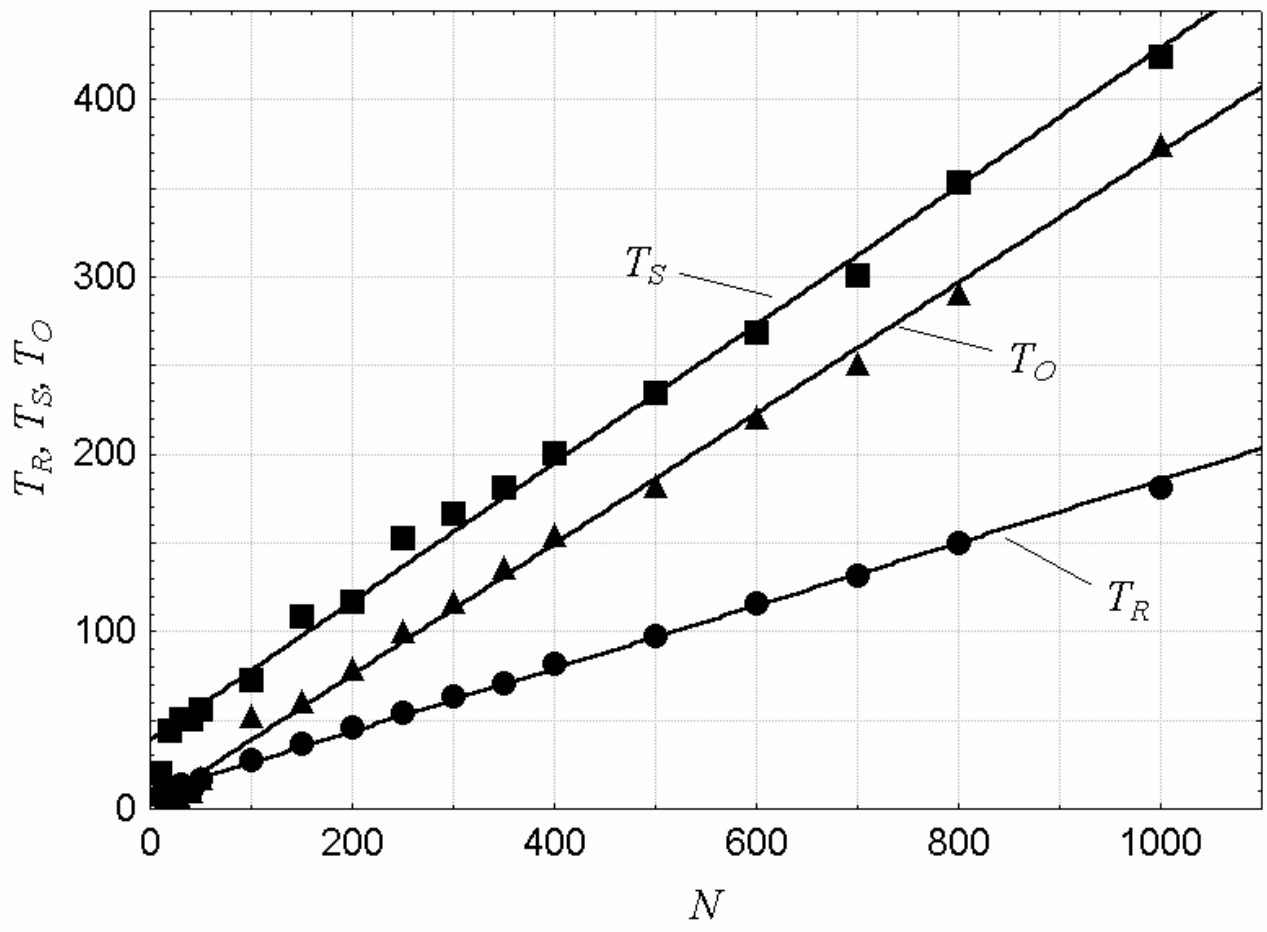
Работа выполнена при финансовой поддержке РФФИ (проект № 04-01-00179) и РАН (Программа "Интеллектуальные компьютерные системы", проект 2-45).

Список литературы

1. Эйген М. Самоорганизация материи и эволюция биологических макромолекул. – М.: Мир, 1973, 216 с. (Eigen M. Selforganization of Matter and the Evolution of Biological Macromolecules // Die Naturwissenschaften, 1971, vol. 58, No 10, pp. 465-523).
2. Эйген М., Шустер П. Гиперцикл. Принципы самоорганизации макромолекул. – М.: Мир, 1982, 270 с. (Eigen M., Schuster P. The Hypercycle: A Principle of Natural Self-Organization, Springer-Verlag: Berlin, Heidelberg, New York, 1979).
3. Holland J.H. Adaptation in Natural and Artificial Systems – Boston, MA: MIT Press, 1992, 211 pp.
4. Редько В.Г. Эволюционная кибернетика. – М.: Наука, 2001, 156 с.
5. Кимура М. Молекулярная эволюция: теория нейтральности. – М.: Мир, 1985, 400 с. (Kimura M. The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, 1983).
6. Редько В.Г. Биофизика.1986. Т. 31. N.3. С. 511-516.
7. Редько В.Г. Биофизика. 1990. Т.35. Вып.5. С.831-834.


t = 500
t = 150
t = 100
t = 50
t = 1
t = 0
0,4
0,3
0,2
0,1
0,0
P(ρ)
0
50
100
150
200
250
ρ


500
400
300
200
100
0
<ρ>
N = 1000
N = 800
N = 600
0
100
200
300
400
t
T_R
T_S


400
300
200
100
0
0
200
400
600
800
1000
T_R, T_S, T_O
N
T_S
T_O
T_R

Подписи к рисункам

Рис. 1. Эволюция распределения особей, N = 500, n = N, β=1, P = 0, 002. По оси ординат отложена доля особей, имеющих рассматриваемое значение ρ . t – номер поколения.

Рис. 2. Зависимости $<\rho>(t)$ для разных значений N . Показана схема оценки значений T_R и T_S для случая N = 600. Зависимости усреднены по 50 расчетам.

Рис. 3. Зависимости времени релаксации T_R, времени выхода на стационар T_S и времени нахождения оптимального решения T_O от длины последовательностей N. Зависимости усреднены по 50 расчетам.

УДК 32.81

В.Г. Редько, Ю.Р. Цой

ОЦЕНКА ЭФФЕКТИВНОСТИ ЭВОЛЮЦИОННЫХ АЛГОРИТМОВ

Реферат

Исследована модель эволюции популяции информационных последовательностей. Сделана грубая аналитическая оценка эффективности эволюционных алгоритмов; оценка проверена путем компьютерного моделирования. Показано, что для случая поиска экстремума унимодальной функции N бинарных переменных эволюционный алгоритм обеспечивает нахождение оптимума при анализе порядка N^2 вариантов.

Институт оптико-нейронных технологий РАН

119333, Москва, ул. Вавилова 44 корп. 2.

Редько Владимир Георгиевич Телефон: (095) 1356331; Факс: (095) 1351351;

E-mail: redko@iont.ru

Томский политехнический университет

634034, Томск, пр. Ленина, 30

Цой Юрий Робертович Телефон: (3822) 418912; Факс: (3822) 419149; E-mail: qai@mail.ru

Контактное лицо – Редько Владимир Георгиевич

Телефоны: (095) 135-63-31 (рабочий)

(095) 538-88-32 (домашний)

(В.Г. Редько)

(Ю.Р. Цой)