Компьютерная Математика

удк 519.217.268

А.А. ВАГИС

ГЕНЕТИЧЕСКИЕ АЛГОРИТМЫ ОПТИМИЗАЦИИ

Для построения оптимальных помехоустойчивых кодов используются генетические алгоритмы. В генетических алгоритмах в качестве «организмов» популяции используются таблицы полярности, а также три вида операций: генерирование случайных кодов для первого поколения, скрещивание кодов, мутация кодов.

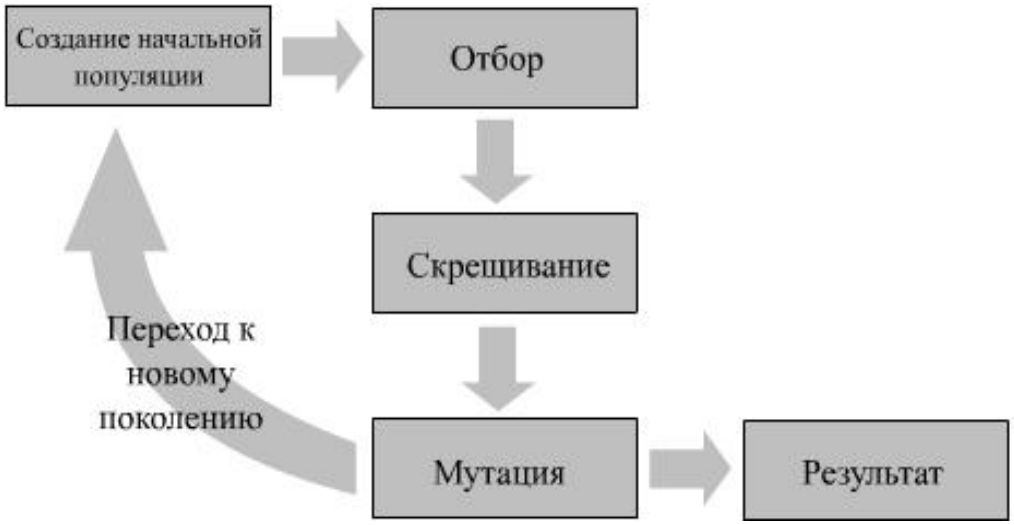
Введение. Генетический алгоритм (ГА) позволяет найти удовлетворительное решение труднорешаемых задач путем последовательного подбора и комбинирования искомых параметров с использованием механизмов, напоминающих биологическую эволюцию. Источником возникновения этих методов стали несколько открытий в биологии. Чарльз Дарвин опубликовал в 1859 году свою знаменитую работу «Происхождение видов», где были провозглашены основные принципы эволюционной теории [1]: наследственность (потомки сохраняют свойства родителей); изменчивость (потомки почти всегда не идентичны); естественный отбор (выживают наиболее приспособленные). Тем самым было показано, какие принципы приводят к эволюционному развитию флоры и фауны под влиянием окружающей среды. Однако вопрос о том, как генетическая информация передается от родителей потомкам, долгое время оставался открытым. В 1953 году в номере журнала Nature вышла статья Уотсона и Крика, которые впервые предложили модель двухцепочной спирали ДНК [2]. Таким образом, стали известны все необходимые компоненты для реализации эволюционной модели.

Принцип работы ГА. Генетические алгоритмы оперируют совокупностью особей (популяцией), которые представляют собой строки, кодирующие одно из решений задачи. Этим ГА отличается от большинства других алгоритмов оптимизации, которые оперируют лишь с одним решением, улучшая его, как это происходит, например, в градиентных методах. С помощью целевой

функции приспособленности среди всех особей популяции выделяют: наиболее приспособленные (более подходящие решения), которые получают возможность скрещиваться и давать потомство; наихудшие (плохие решения), которые удаляются из популяции и не дают потомства. Таким образом, приспособленность нового поколения в среднем выше предыдущего. В классическом ГА начальная популяция формируется случайным образом; размер популяции (количество особей L) фиксируется и не изменяется в течении работы всего алгоритма.

ГА дает преимущества при решении практических задач. Одно из них – это адаптация к изменяющейся окружающей среде. В реальной жизни поставленная задача может претерпевать большие изменения в процессе своего решения. При использовании традиционных методов все вычисления приходится начинать заново, что приводит к значительным затратам машинного времени. При эволюционном подходе популяцию можно анализировать, дополнять и видоизменять применительно к изменяющимся условиям, для этого не требуется полный перебор.

На рисунке показана схема работы любого генетического алгоритма. Проанализируем работу ГА на примере определения оптимальных помехоустойчивых генетических кодов.



РИСУНОК

Помехоустойчивость генетических кодов. В работе [3] исследована помехоусточивость стандартного генетического кода относительно полярности аминокислот при точечных мутациях нуклеотидов в кодоне. Если в результате мутации полярный остаток в белке сменится на неполярный (или наоборот),

то форма молекулы может измениться настолько, что белок не сможет выполнять свою функцию. В табл.1 неполярные аминокислоты выделены жирным шрифтом. Каждый триплет допускает 9 однократных замен, число кодирующих аминокислоты триплетов равно 61. Для стандартного кода возможны 526 замещений, в которых кодон до и после мутации не является одним из трех стоп-кодонов, из них 364 замещения не меняют полярность, т. е. его помехоустойчивость составляет 364/526 = 69,20 %. Помехоустойчивость кода – целевая функция в генетических алгоритмах при построении оптимальних кодов.

ТАБЛИЦА 1. Стандартный генетический код организмов

Первое основание	Второе основание				Третье основание
	T	C	A	G	
T	**фенилаланин**	серин	тирозин	**цистеин**	T
	фенилаланин	серин	тирозин	**цистеин**	C
	лейцин	серин	стоп	**стоп**	A
	лейцин	серин	стоп	**триптофан**	G
C	**лейцин**	**пролин**	гистидин	агринин	T
	лейцин	**пролин**	гистидин	агринин	C
	лейцин	**пролин**	глутамин	агринин	A
	лейцин	**пролин**	глутамин	агринин	G
A	**изолейцин**	треонин	аспарагин	серин	T
	изолейцин	треонин	аспарагин	серин	C
	изолейцин	треонин	лизин	агринин	A
	метионин	треонин	лизин	агринин	G
G	**валин**	**аланин**	асп. к-та	**глицин**	T
	валин	**аланин**	асп. к-та	**глицин**	C
	валин	**аланин**	глут. к-та	**глицин**	A
	валин	**аланин**	глут. к-та	**глицин**	G

Для генетических алгоритмов нужно выполнять три вида операций: генерировавать случайные коды для первого поколения; скрещивать коды; подвергать коды мутациям.

Генерация. Пусть St – стандартный генетический код, S = {0,1, stop}, где 0 соответствует неполярным аминокислотам, 1 – полярным, $P_{st}(s)$, $s \in S$ – вероятности встречи в нем соответствующих состояний:

$$P_{st}(s) = \frac{1}{64} | St^{-1}(s) | .$$

Начальное поколение состоит из случайных «перестановок» стандартного генетического кода, поскольку он уже обладает довольно высокой помехоустойчивостью, а наша цель состоит в построении оптимальных кодов. Тогда случайный код *code* можно генерировать как

$$code(c) = \begin{cases} x\ , & \text{с вероятностью } P_{st}(x), \\ \vdots & \\ stop, & \text{с вероятностью } P_{st}(stop). \end{cases}$$

Скрещивание. Пусть заданы два кода $code_1$ и $code_2$; тогда результат их скрещивания $code(c) = \{ code_1(c),$ с вероятностью 1/2; $code_2(c)$, с вероятностью 1/2$\}$.

Мутации. Пусть задан код *code* и вероятность мутации *p*. Результат мутации –

$$code'(c) = \begin{cases} code(c), & \text{с вероятностью } (1-p), \\ x\ , & \text{с вероятностью } pP_{st}(x) / (1 - P_{st}(code(c))), \\ \vdots & \\ stop, & \text{с вероятностью } pP_{st}(stop) / (1 - P_{st}(code(c))). \end{cases}$$

Оптимальные генетические коды. В ГА в качестве «организмов» популяции использовались не сами генетические коды, а таблицы полярности, которые ставят в соответствие каждому кодону один из трех типов информации: полярную аминокислоту, неполярную аминокислоту или стоп-кодон. Таким образом, значения клеток таблицы принадлежат множеству $S = \{0,1, \text{stop}\}$.

С помощью таблиц полярности можно избавиться от избыточной информации. Количество вариантов генетических кодов оценивается примерно величиной 21^{64} (это число превосходит 10^{80}), а таблиц полярности – всего 3^{64} (это число не превосходит 10^{30}). Поэтому с помощью генетических алгоритмов удалось достаточно экономно и быстро построить оптимальные варианты помехоустойчивых генетических кодов.

При выполнении операции мутации значение кода, соответствующее каждому из 64 кодонов, может с определенной вероятностью p измениться на другое возможное значение, причем распределение вероятностей перехода соответствует распределению значений в стандартном генетическом коде.

Использовались следующие параметры алгоритма: количество поколений T = 50000; размер начального поколения N = 50; максимальный размер поколения L = 250; число скрещиваний и мутаций для каждого кода $N_c = 4$ и $N_m = 2$ соответственно; вероятность мутации в отдельном кодоне p = 0,1.

В отличие от методов оптимизации операции скрещивания и мутации применяются для каждого кода в текущем поколении. Наихудшие помехоустойчивые коды, которые например не удовлетворяют характеристикам стандартного кода или близким кодам, удаляются из множества кодов поколения. Как известно, в градиентных процедурах оптимизации поиск оптимального решения проводится на каждой итерации для одного или нескольких векторов состояний.

Один из полученных оптимальных симметричных генетических кодов по отношению к полярности приведен в табл. 2. Этот код можно получить из стандартного кода с помощью восьми парных перестановок. В симметричном коде кодон (ijk) кодирует полярную аминокислоту, а антикодон $(\widetilde{kji})$ – неполярную, либо наоборот. С помощью генетических алгоритмов построены различные варианты наиболее помехоустойчивых кодов (78,30 %), отличающиеся от стандартного кода числом стоп-кодонов и количеством триплетов, определяющих полярные и неполярные аминокислоты (например, количество стоп кодонов в кодах этого класса изменялось от 2 до 4, количество кодонов полярных и неполярных аминокислот – от 27 до 34).

ТАБЛИЦА 2. Оптимальный симметричный код

Первое основание	Второе основание				Третье основание
	T	C	A	G	
T	**фенилаланин**	стоп	тирозин	серин	T
	фенилаланин	**триптофан**	тирозин	серин	C
	лейцин	**цистеин**	стоп	серин	A
	лейцин	**цистеин**	стоп	серин	G
C	**лейцин**	**пролин**	гистидин	аргинин	T
	лейцин	**пролин**	гистидин	аргинин	C
	лейцин	**пролин**	глутамин	аргинин	A
	лейцин	**пролин**	глутамин	аргинин	G
A	**изолейцин**	**глицин**	аспарагин	серин	T
	изолейцин	**глицин**	аспарагин	серин	C
	изолейцин	**глицин**	лизин	аргинин	A
	метионин	**глицин**	лизин	аргинин	G
G	**валин**	**аланин**	асп. к-та	треонин	T
	валин	**аланин**	асп. к-та	треонин	C
	валин	**аланин**	глут. к-та	треонин	A
	валин	**аланин**	глут. к-та	треонин	G

Заметим, что оптимальных помехоустойчивых кодов существует несколько. Симметрия кода, понимаемая как переход от полярной аминокислоты к неполярной при переходе от кодона к антикодону, не является необходимым условием оптимальности генетического кода.

Если в коде, приведенном в табл. 2, поменять местами аминокислоты столбцов C и A, получим код, для которого симметрия не выполняется вообще: кодон и антикодон кодируют аминокислоту одного типа, поскольку неполярные аминокислоты содержатся в столбцах T и A, а полярные – в столбцах C и G. Поэтому помехоустойчивость кода такая же, как у симметричного кода – 77,86 %.

Если во всех кодонах кода в табл. 2 переставить местами второй и третий нуклеотиды, т. е. в каждой строке кодоны «повернуть на 90 градусов», получим код с такой же помехоустойчивостью, что и симметричный код, в котором симметрия выполняется примерно в половине случаев. В этом коде неполярные аминокислоты определяются триплетами с третьим нуклеотидом T и C, а у полярных аминокислот третий нуклеотид – A и G.

Выводы. ГА – новая фундаментальная область исследований, основанная Д. Холландом, впервые примененная для решения задач оптимизации и распознавания образов. ГА моделируют в определенной степени механизмы развития органического мира на Земле. В настоящее время ГА – это мощный адаптивный поисковый метод, основанный на эволюционных факторах конструирования популяции, естественного отбора, селекции, мутации, которые реализованы в работе для построения оптимальных генетических кодов.

Как уже отмечалось выше, число вариантов генетических кодов превосходит астрономическую величину 10^{80}. Если ежесекундно перебирать генетические коды, то за 10 миллиардов лет количество просмотров не превосходит величину 10^{18}. Такую мизерную долю как 10^{-62} невозможно зафиксировать даже в квантовой механике. Поэтому весьма проблематично путем естественного отбора живых организмов построить стандартный генетический код с достаточно высоким уровнем помехоустойчивости.

В ГА в качестве «организмов» популяции использовались не сами генетические коды, а таблицы полярности, которые ставят в соответствие каждому кодону один из трех типов информации: полярную аминокислоту, неполярную аминокислоту или стоп-кодон. С помощью таблиц полярности удалось избавиться от огромной избыточной информации.

С помощью генетических алгоритмов удалось построить оптимальные варианты помехоустойчивых кодов. Найти такие варианты генетических кодов на основе оптимизационных градиентных процедур невозможно.

О.А. Вагіс

ГЕНЕТИЧНІ АЛГОРИТМИ ОПТИМІЗАЦІЇ

Для побудови оптимальних завадостійких кодів виконувались генетичні алгоритми. У генетичних алгоритмах як «організми» популяції використовувалися таблиці полярності, а також три види операцій: вибір випадкових кодів для першого покоління; схрещування кодів; мутації кодів.

A.A. Vagis

GENETIC ALGORITHMS OF OPTIMIZATION

We apply genetic algorithms to construct noise-free optimal codes. In genetic algorithms, the polarity tables are used as the “organisms” of a population, as well as three types of operations: generation of random codes for the first generation, crossing and mutation of codes.

Список литературы

1. Дарвин Ч. Происхождение видов путем естественного отбора или сохранение благоприятных рас в борьбе за жизнь. Санкт-Петербург: Наука, 1991. 544 с.
2. Watson J.D., Crick F.H. A structure for Deoxyribose Nucleic Acid. *Nature*. 1953. **171**. P. 737 – 738.
3. Гупал А.М., Сергиенко И.В. Симметрия в ДНК. Методы распознавания дискретных последовательностей. К.: Наукова думка, 2016. 228 с.

Получено 07.08.2019

Об авторе:

Вагис Александра Анатольевна,
доктор физико-математических наук, ведущий научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.