

Монахов О.Г., Монахова Э.А. Применение метаэвристического программирования для решения задачи символьной регрессии. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XIX Междунар. научно-техн. конф. – Пенза: ПДЗ, 2019. – С. 027-033.

УДК 519.876.5 + 004.023

## ПРИМЕНЕНИЕ МЕТАЭВРИСТИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ СИМВОЛЬНОЙ РЕГРЕССИИ<sup>1</sup>

О.Г. Монахов, Э.А. Монахова

## APPLICATION OF METAHEURISTIC PROGRAMMING TO SOLVE THE PROBLEM OF SYMBOLIC REGRESSION

O.G. Monakhov, E.A. Monakhova

**Аннотация.** Рассмотрено применение алгоритма метаэвристического программирования (МР), основанного на различных биоинспирированных алгоритмах, для решения задачи символьной регрессии и синтеза нелинейных моделей и проведено сравнение данного алгоритма со стандартным алгоритмом генетического программирования (ГП). Показано, что алгоритм МР превосходит алгоритм ГП как по времени, необходимому для поиска решения (в большинстве случаев более чем на порядок), так и по вероятности нахождения заданной модели.

**Ключевые слова:** метаэвристическое программирование, генетическое программирование, нелинейные модели, биоинспирированные алгоритмы.

**Abstract.** The application of the metaheuristic programming (MR) algorithm based on various bio-inspired algorithms for solving the problem of symbolic regression and the synthesis of nonlinear models is considered, and this algorithm is compared with the standard genetic programming algorithm (GP). It is shown that the MR algorithm is superior to the GP algorithm both in the time required to find a solution (in most cases by more than an order of magnitude) and in the probability of finding a given model.

**Keywords:** metaheuristic programming, genetic programming, nonlinear models, bioinspired algorithms.

Рассмотрено решение задачи символьной регрессии и построения нелинейных моделей в виде математических выражений, функций, формул, алгоритмов и программ на основе приведенных экспериментальных данных, наборов переменных, основных функций и операций. Задача состоит в том, чтобы найти математическое выражение, которое лучше всего описывает нелинейную вычислительную модель, определяемую набором входных ( $X$ ) и выходных ( $Y$ ) экспериментальных данных на основе заданного набора базовых функций и операций и заданного набора переменных и констант, которые используются для автоматического создания аналитических выражений (формул),

---

<sup>1</sup>Работа выполнена в рамках бюджетного проекта ИВМиМГ СО РАН №0315-2016-0006.

представляющих модель, и компьютерных программ для их расчета. Предполагается, что целевая функция (фитнес-функция) вычисляет сумму квадратов отклонений выходных данных для полученного выражения от заданных опорных выходных значений  $Y$  для заданных входных данных  $X$ .

Известные подходы к этой проблеме включают в себя в основном варианты генетического программирования (ГП, GP) [1], [2], [3], которое ориентировано в первую очередь на решение задач автоматического синтеза программ на основе данных обучения посредством эволюционного поиска моделей, минимизирующих ошибку представления. При этом хромосомами имеют древовидные структуры, которые автоматически генерируются с использованием генетических операторов в GP и представляют после интерпретации некоторые выражения и компьютерные программы различных размеров и сложностей, которые реализуют выражения.

В данной работе представлен новый алгоритм метаэвристического программирования (MP) для решения задачи символьной регрессии и синтеза нелинейных моделей. Мы опишем алгоритм MP, основанный на четырех различных биоинспирированных алгоритмах для синтеза нелинейных моделей, сравним наш алгоритм со стандартным генетическим программированием и покажем, что алгоритм MP обладает более высокой эффективностью эволюционного поиска.

Метаэвристическое программирование для синтеза нелинейных моделей основано на биоинспирированных (эволюционных) вычислениях и моделировании природных процессов в совокупности агентов, каждый из которых является точкой в пространстве решений задачи оптимизации [4], [5]. Индивидуумы (агенты, муравьи, пчелы, светлячки, частицы, учащиеся и др.) являются структурами данных (хромосомами), а именно последовательностями вещественных чисел (векторов), которые кодируют математические выражения (формулы и программы). Каждая популяция представляет собой набор хромосом, и каждая хромосома в этом алгоритме определяет множество выражений (формул), возникающих из нее после декодирования. Основная идея алгоритма метаэвристического программирования – эволюционное преобразование набора хромосом (формул) в процессе естественного отбора для выживания «сильнейших». В нашем случае эти индивидуумы являются выражениями, которые имеют наименьшее значение целевой функции.

Этапы моделирования эволюционного процесса в алгоритме метаэвристического программирования на основе биоинспирированных популяционных алгоритмов описаны в следующем шаблоне.

1. Создание начальной популяции из случайно сгенерированного начального приближения решений (хромосом) в виде последовательностей вещественных чисел. Обратите внимание, что решение представлено в хромосоме в зашифрованном виде, то есть в виде генотипа.

2. Оценка популяции с помощью функции пригодности (фитнес-функции), которая показывает, насколько хорошо каждый индивидуум

решает данную проблему. При этом генотип расшифровывается в фенотип, чтобы интерпретировать хромосому как программу для расчета фитнес-функции.

3. Создание популяции следующего поколения с использованием операторов миграции, заданных для каждого биоинспирированного алгоритма, чтобы привести индивидов в этой популяции к экстремуму целевой функции.

4. Повторять пункты 2 и 3 до тех пор, пока не будет найдено решение, соответствующее заданным критериям, или пока не будет достигнуто максимальное количество поколений.

В метаэвристическом программировании предлагается унифицированный подход для декодирования основных структур данных, хромосом для различных алгоритмов, вдохновленных природой. Этот подход основан на представлении линейной структуры хромосомы в виде последовательности трехадресных инструкций и создании линейной операторной структуры для вычисления выражений (формул), закодированных в хромосоме. Для этого последовательность вещественных чисел (хромосома) делится на группы из трех элементов (триплетов)  $(h_1, h_2, h_3)$  с  $0 < h_i < 1$ . Каждая такая группа интерпретируется как трехадресная инструкция следующим образом:  $\langle oper \rangle \langle adr1 \rangle \langle adr2 \rangle$ , где операция  $oper$  применяется к операндам в инструкциях с номерами  $adr1$  и  $adr2$ , которые вычисляются по следующим формулам:  $oper = \lfloor h_1 * |F| \rfloor$ ,  $adr1 = \lfloor (h_2 * (I - 1)) \rfloor + 1$ ,  $adr2 = \lfloor (h_3 * (I - 1)) \rfloor + 1$ , *if*  $oper = 0$  *then*  $adr1 = \lfloor (h_2 * |T|) \rfloor + 1$ , где  $|F|$  – это число базовых функций,  $oper$  – это номер элемента в этом наборе, т.е. номер функции, выполняемой в текущей инструкции,  $I$  – это номер текущей инструкции,  $adr1, adr2$  – это номера предыдущих инструкций, результаты выполнения которых используются в качестве операндов в текущей инструкции, а  $|T|$  – это число элементов множества терминальных символов. Если  $oper=0$ , то инструкция интерпретируется как оператор загрузки, и загружается символ терминала с номером  $adr1$ . Решение в этом случае представляет собой набор функций, где каждая функция есть последовательность операторов от первого оператора до каждого текущего. Это позволяет, в отличие от стандартного ГП, одновременно оценивать множество выражений в виде последовательностей операторов и сокращать время поиска оптимального решения.

Преимущество этого подхода заключается в простом алгоритме декодирования действительной хромосомы (вектора) в набор команд, который позволяет использовать эти векторы в едином подходе для поиска оптимальных моделей с использованием различных биоинспирированных популяционных алгоритмов [4], [5]. В этой работе в рамках подхода метаэвристического программирования (mp) использовались четыре различных алгоритма, основанных на природных процессах, для поиска оптимальных моделей: модифицированный алгоритм оптимизации на

основе преподавания и обучения [6] (mpMTLB), однофазный мульти-групповой алгоритм преподавания и обучения [7] (mpSPM), эволюционная стратегия с адаптацией ковариационной матрицы [8] (mpCMA), алгоритм поиска на основе теплообмена [9] (mpSHT).

Для сравнения эффективности метаэвристического программирования с использованием четырех различных биоинспирированных популяционных алгоритмов и стандартного алгоритма генетического программирования для задачи поиска аналитического описания модели на основе приведенных экспериментальных данных, наборов переменных, базовых функций и операций, мы выбрали шесть тестовых функций:

$$\text{Test1: } \sin(x^3) + e^x, \text{Test2: } \sin(2x) + \frac{1}{x} - x^3, \text{Test3: } \frac{x+a}{\sin(2x-4)}, \\ \text{Test4: } 2\sin(x)\exp(a), \text{Test5: } \sin(x) + \sin(a^2), \text{Test6: } x^5 - 2x^3 + x.$$

Значения каждой функции в 20 случайных точках в диапазоне (0,2) были использованы в экспериментах. Набор базовых функций для Test1:  $F_1 = \{+, *, \sin, \exp\}$ , для Test2-Test6  $-F_2 = \{+, -, *, /, \sin, \exp\}$ , терминальные символы для Test1 и Test6  $-T_1 = \{x\}$ , для Test2 -Test5  $-T_2 = \{x, a, 2, 5\}$ .

Применяются следующие параметры сравниваемых алгоритмов: вероятность кроссовера 0.80, вероятность мутации 0.15, размер популяции 100 и максимальное число поколений 250. Длина хромосом равна 30; в GP начальная глубина дерева выражений установлена равной 6. Программа была выполнена 100 раз для каждого алгоритма и каждой тестовой функции, и результаты были усреднены. В описанных экспериментах мы использовали рекомендованные значения параметров алгоритмов их авторов в цитированных статьях для сравнения эффективности. Отдельная оптимизация параметров алгоритма не проводилась. Мы сравниваем MP на основе вдохновленных природой алгоритмов, реализованных в MATLAB, с реализацией алгоритма генетического программирования (GP) [1, 2] в MATLAB – библиотеки GPLAB v.4.02 (<http://gplab.sourceforge.net>).

Одним из основных показателей, используемых для измерения эффективности алгоритмов эволюционного синтеза, является вероятность (частота) успеха, то есть вероятность того, что алгоритм обнаружил (синтезировал) выражение, совпадающее именно с заданной функцией. Это отношение количества успешных экспериментов, когда алгоритм нашел правильное выражение, к общему количеству экспериментов с заданными параметрами. В таблице 1 представлены вероятности (частоты) успеха для разных алгоритмов, где  $\bar{p}$  – средняя частота успеха для каждого алгоритма на множестве тестовых функций. Из таблицы видно, что вероятность успеха в MP выше, чем в GP во всех случаях. В таблице 2 представлено время выполнения (в сек.) для различных алгоритмов, т. е. среднее время выполнения алгоритма до того, как алгоритм обнаружил (синтезировал) указанное выражение или выполнил требуемое число

поколений, где  $\bar{t}$  – среднее время выполнения для каждого алгоритма для всего множества тестовых функций. Эта таблица показывает, что у МР меньше времени на поиск решения (в большинстве случаев более чем на два порядка), чем у GP.

Таблица 1

*Частота успешного поиска тестовых функций*

|           | GP   | mpCMA | mpSHT | mpSPM | mpMTLB |
|-----------|------|-------|-------|-------|--------|
| Test1     | 0.67 | 1     | 1     | 1     | 1      |
| Test2     | 0    | 0     | 0.01  | 0.01  | 0.04   |
| Test3     | 0    | 0.02  | 0.01  | 0.01  | 0      |
| Test4     | 0.03 | 0.89  | 0.96  | 0.97  | 1      |
| Test5     | 0.05 | 0.99  | 1     | 1     | 1      |
| Test6     | 0.08 | 0.86  | 0.92  | 0.96  | 1      |
| $\bar{p}$ | 0.19 | 0.67  | 0.72  | 0.72  | 0.76   |

Таблица 2

*Среднее время поиска тестовых функций*

|           | GP     | mpCMA | mpSHT | mpSPM | mpMTLB |
|-----------|--------|-------|-------|-------|--------|
| Test1     | 118.6  | 1.4   | 0.9   | 1.3   | 0.66   |
| Test2     | 33.2   | 81.8  | 15.4  | 21.3  | 23.0   |
| Test3     | 11265  | 76.7  | 16.9  | 21.4  | 27.4   |
| Test4     | 4425   | 12.8  | 3.62  | 5.84  | 3.2    |
| Test5     | 4746   | 6.9   | 2.55  | 3.74  | 1.94   |
| Test6     | 5540   | 28.6  | 4.75  | 7.6   | 2.95   |
| $\bar{t}$ | 4280.8 | 30.5  | 6.04  | 8.28  | 8.05   |

Таким образом, эксперименты показывают, что предлагаемый алгоритм метаэвристического программирования превосходит известный алгоритм генетического программирования как по времени поиска решения (меньше более чем на порядок), так и по величине вероятности нахождения функции (модели).

Библиографический список

1. J. Koza, Genetic Programming II: Automatic Discovery of Reuseable Programs, MIT Press, 1996.
2. W.B. Langdon, R. Poli, Foundations of Genetic Programming. Springer-Verlag, 2002.
3. Монахов О.Г., Монахова Э.А. Синтез новых семейств оптимальных регулярных сетей на основе эволюционных вычислений и шаблонов функций //Автометрия. 2004. Т. 40. № 4. С. 106-116.
4. Ke-Lin Du, M.N. S. Swamy, Search and Optimization by Metaheuristics. Techniques and Algorithms Inspired by Nature, Basel: Birkhauser, 2016.

5. X.S. Yang, Nature-Inspired Metaheuristic Algorithms. Luniver Press. 2008.

6. M. Crepinsek, S-H Liu, L. Mernik, "A note on teaching - learning-based optimization algorithm", Information Sciences, Vol.212, 2012, pp. 79-93.

7. R. Kommadath, C. Sivadurgaprasad, P. Kotecha, "Single phase multi-group teaching learning algorithm for single objective real-parameter numerical optimization", IEEE Congress on Evolutionary Computation (CEC), 2016, pp. 1165-1172.

8. N. Hansen, A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies", Evolutionary Computation. Vol. 9, N 2, 2001, pp. 159-195.

9. D. Maharana, P. Kotecha, "Simultaneous Heat Transfer Search for Computationally Expensive Numerical Optimization", IEEE Congress on Evolutionary Computation, (CEC), 2016, pp. 2982-2988.

**Монахов Олег Геннадьевич**

Институт вычислительной  
математики и математической  
геофизики СОРАН,  
г. Новосибирск, Россия

**Monakhov O.G.**

Institute of Computational  
Mathematics and Mathematical  
Geophysics SB RAS,  
Novosibirsk, Russia

**Монахова Эмилия Анатольевна**

Институт вычислительной  
математики и математической  
геофизики СОРАН,  
г. Новосибирск, Россия

**Monakhova E.A.**

Institute of Computational  
Mathematics and Mathematical  
Geophysics SB RAS,  
Novosibirsk, Russia