УДК 519.68

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ МЕТОДОВ САМОНАСТРОЙКИ АЛГОРИТМА ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ ЗАДАЧИ СИМВОЛЬНОЙ РЕГРЕССИИ

Становов В.В.
Научный руководитель проф. Е.С. Семенкин
Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева

Рассматривается эффективность работы самонастраивающегося алгоритма генетического программирования для задачи символьной регрессии. Разработана программная среда, реализующая алгоритм. Получены результаты работы на тестовых и практических задачах.

Задача восстановления регрессионной зависимости может быть решена различными способами, в том числе с использованием метода генетического программирования для символьной регрессии. Однако, как и для генетического алгоритма, для генетического программирования существует проблема выбора тех или иных операторов: выбор пользователем неподходящих типов селекции, скрещивания и мутации может существенно снизить эффективность работы алгоритма. Данную проблему можно решить, используя различные методы самонастройки.

Для исследования эффективности работы различных методов самонастройки была разработана программная среда, реализующая алгоритм. В программе имеется возможность выбора трех типов селекции: пропорциональная, ранговая и турнирная с настраиваемым размером турнира; два типа скрещивания: стандартное и одноточечное; два типа мутации: точечная и выращиванием поддерева с настраиваемой вероятностью мутации. Были использованы два метода самонастройки: Population-Level Dynamic Probabilities (PDP) и Individual-Level Dynamic Probabilities (IDP). Данные подходы были разработаны для настройки вероятности выбора того или иного типа мутации в генетическом программировании.

Суть метода PDP состоит в том, что вероятности генетических операторов напрямую зависят от успешности применения того или иного оператора, т.е. чем более успешен оператор, тем больше он поощряется. Вероятности вычисляются следующим образом:

$$p_i = p_{all} + \left[\frac{r_i \cdot (100 - n \cdot p_{all})}{scale}\right], \text{ где}$$

$$r_i = \frac{success_i^2}{used_i}, p_{all} = \frac{20}{n}, scale = \sum_{j=1}^{n} r_j$$

Здесь $success_i$ – число успешных применений оператора i, $used_i$ – общее число применений, n – общее число операторов.

Метод IDP сопоставляет каждому из индивидов в популяции свои значения счетчиков cnt_j^i числа неудачных применений операторов. На их основании вычисляются вероятности применения операторов для каждого индивида отдельно.

$$p_i = p_{all} + \left[\frac{\left(max_{1 \le k \le n} cnt_j^k + 1 - cnt_j^i\right) \cdot (100 - n \cdot p_{all})}{n \cdot \left(max_{1 \le k \le n} cnt_j^k + 1\right) - \sum_{k=1}^{n} cnt_j^k}\right], \text{ где}$$

$$p_{all} = \frac{20}{n}, n \text{ –число операторов}$$

В качестве тестовых функций были взяты функции двух переменных с различными параметрическими структурами, приведенные в специальном репозитории. Исследование проводилось по следующей схеме: тестовые выборки объемом 400 были сгенерированы для каждой функции, в каждом независимом прогоне алгоритма устанавливалось число индивидов равным 100 и число поколений равным 1000, среднеквадратичная ошибка и

длина выражения (без учета скобок) усреднялись по 20 прогонам алгоритма. Результаты приведены в таблице ниже:

	IDP		PDP	
номер функции	ошибка	длина выражения	ошибка	длина выражения
1	0,00390794	56,75	0,00452203	60,35
2	0,0024608	61,4	0,00329423	65,95
3	0,00240595	33,8	0,00238137	38,5
4	0,00198606	25,95	0,00195891	27,5
5	0,000739306	19,20	0,000929058	15,3
6	0,00294205	51,75	0,0033497	46,55
7	0,00449295	49,1	0,00481504	36,35
8	0,002114552	28,79	0,001994	17,4
9	0,00329502	29,05	0,00298749	32
10	0,00268676	61	0,00324459	53,05
11	0,00308336	58,35	0,00416098	58,7
12	0,000486303	40,5	0,000495485	31,63
13	0,00394331	61,8	0,00404056	67
14	0,00450659	40,65	0,00658	15,33
15	0,00402964	54,15	0,00414904	52,3
16	0,00857015	10,10	0,0084215	14,10
17	0,0029538	37,3	0,00303079	33,85
18	0,002411	23,35	0,00251308	25,30
19	0,00371997	34,25	0,00396451	29,95
20	0,00198532	37,8	0,00199173	38,30
21	0,00699476	18,45	0,00704279	19
22	0,0115041	25,55	0,0108664	34,90
23	0,0124355	10,25	0,0124634	12,15
24	0,00656989	36,25	0,00671105	33,65
Среднее	0,004176045	37,73083333	0,004412822	35,79625

Кроме того, при помощи описанных алгоритмов были решены задачи банковского скоринга по базам данных Australia-1 и Germany-1, приведенным в специальном репозитории в Интернете. Данные задачи сводятся к задачам классификации, которые решались путем построения разделяющей поверхности в виде символьного выражения. Анализ результатов показывает работоспособность предложенных методов самонастройки не только на тестовых функциях, но и при решении реальных задач классификации и построения регрессии.