Разработка и исследование гибридного метода генетического программирования

авторы: Бухтояров В.В., Семенкин Е.С.

Для исследования характеристик практически любого процесса математическими методами, включая машинные, должна быть проведена формализация этого процесса, то есть построена математическая модель. Необходимость построения модели, в частности, может быть обусловлена невозможностью активного экспериментирования с объектом или чрезвычайной дороговизной постановки экспериментов на таком объекте. Исследование с помощью моделей зачастую оказывается единственно возможным способом изучения и решения важнейших практических задач. Математические модели позволяют воспроизводить реальные процессы, их структуру, свойства и поведение. С их помощью можно получить параметры и характеристики системы и ее отдельных подсистем значительно проще, быстрее и экономичнее, чем при исследовании реальной системы.

Построение модели для исследования сложных процессов и систем само по себе становится сложным процессом, требующим значительных усилий, направленных на нахождение функциональных зависимостей между входными и выходными переменными, особенно если требуется найти эти зависимости в аналитическом виде.

В большинстве численных методов идентификации (регрессионный анализ, непараметрические и нейросетевые методы) для построения зависимостей, аппроксимирующих экспериментальные данные, используют регрессионные модели. Одним из недостатков упомянутых численных методов является то, что построенная при их помощи модель по сути является моделью черного ящика. Преодолеть этот недостаток можно за счет сведения исходной задачи к задаче символьной регрессии и последующего ее решения подходящими методами. Задача символьной регрессии заключается в нахождении в символьной форме математического выражения, аппроксимирующего зависимость между входными и соответствующими им выходными переменными исследуемого процесса или системы. Полезным свойством решения задачи символьной регрессии представляется то, что полученное решение, помимо собственно вычислительной процедуры, является символьным математическим выражением – формулой, которая может быть подвергнута содержательному анализу, а затем упрощена либо уточнена. На современном этапе методы решения задачи символьной регрессии не разработаны достаточно хорошо. Метод генетического программирования – один из самых многообещающих подходов в данном направлении [1].

В настоящей статье рассматривается постановка задачи символьной регрессии, а также предлагается существенная модификация стандартного алгоритма генетического программирования, повышающая эффективность решения задач символьной регрессии методом генетического программирования за счет совершенствования процедуры синтеза структуры модели и более тонкой настройки ее параметров с помощью эволюционных алгоритмов [2].

Рассмотрим постановку задачи символьной регрессии. Пусть имеются наблюдения входных и выходных переменных исследуемого процесса V=(xi, yi), $\mathbf{i = \overline{1,n}}$, n – объем

выборки наблюдений. Известно, что существует зависимость y=f(x), связывающая входные и выходные переменные исследуемого процесса, однако вид и структура зависимости неизвестны.

Необходимо по имеющейся выборке наблюдений V найти в символьной форме математическое выражение $\hat{y} = \hat{f}(x)$, аппроксимирующее зависимость между входными и выходными переменными. Символьное математическое выражение (формула, связывающая входные и выходные параметры процесса) должно по возможности наиболее точно соответствовать реальной зависимости.

В качестве критерия оптимальности построенной регрессионной модели может использоваться, например, величина относительной средней ошибки моделирования по выборке:

$$W = \frac{1}{n \cdot (y^{max} - y^{min})} \sum_{i=1}^{n} \left| \hat{y}(x_i) - y_i \right|, \qquad (1)$$

где i – номер наблюдения в выборке, $i = \overline{1,n}$; ymax и ymin – максимальное и минимальное значения выходного параметра.

Основная идея, используемая при решении задачи символьной регрессии методом генетического программирования, состоит в том, чтобы с помощью программной реализации процедур, имитирующих эволюционные процессы, из случайно сгенерированного множества функциональных зависимостей получить зависимость, хорошо описывающую данные, представленные в исходных выборках, в идеале совпадающую с подлинной. Сам метод генетического программирования представляет собой некоторое расширение идей, заложенных в генетических алгоритмах [3]. В ряде исследований генетических алгоритмов была предложена и успешно апробирована схема их гибридизации [4]. Ввиду указанной схожести подходов, используемых в генетических алгоритмах и генетическом программировании, становится очевидным, что идеи, лежащие в основе гибридного генетического алгоритма, можно использовать и в генетическом программировании.

С этой целью была разработана специальная процедура локального поиска на структуре решения [5]. Опишем ее подробнее.

1. Отбор наиболее перспективных с точки зрения решаемой задачи индивидов текущего поколения. Обычно отбирается заданное количество индивидов, имеющих наибольшее значение функции пригодности.

2. Применение к отобранным индивидам процедуры локального поиска. Локальный поиск осуществляется на заданном числе r вершин дерева, которыми представлено исходное решение. Вершины дерева выбираются случайно с равной вероятностью $p_i = \frac{1}{s}$, где i – номер вершины, $i = \overline{1,r}$; s – число вершин в дереве.

Поиск осуществляется следующим образом:

- значение в выбранной вершине дерева заменяется новым в соответствии с правилами для локального поиска в генетическом программировании [5];

- вычисляется значение функции пригодности индивида, измененного на предыдущем шаге;

- если пригодность измененного индивида выше пригодности исходного индивида, то шаг поиска считается успешным, новое содержание вершины дерева фиксируется; иначе шаг поиска считается неудачным, вершине возвращается исходное значение;

- переходим к новому шагу локального поиска (к началу п. 2), если число шагов не превышает r, иначе поиск останавливается.

3. Возвращение индивидов в исходную популяцию для применения к ним операторов стандартного метода генетического программирования. Индивид возвращается в популяцию таким, каким он стал после применения локального поиска, и с новым значением функции пригодности.

Операторы селекции, скрещивания и мутации в гибридном алгоритме генетического программирования применяются аналогично соответствующим операторам стандартного метода генетического программирования.

Гибридизация стандартного метода генетического программирования не исключает возможности настройки параметров модели с помощью эволюционных алгоритмов оптимизации. В общем случае получаем модификацию метода генетического программирования, включающую в себя идеи и операторы стандартного метода генетического программирования, локального поиска на структуре решения (дерева) и эволюционных алгоритмов оптимизации.

Общая схема гибридного алгоритма генети- ческого программирования представлена на рисунке 1.

Для оценки эффективности предлагаемого метода гибридного генетического программирования был проведен ряд численных экспериментов. В качестве конкурирующих подходов в исследования включен стандартный метод генетического программирования (далее на рисунках и в таблицах он обозначен как Стандартный МГП).

При сравнительном исследовании эффективности использовались параметры методов, определенные на стадии предварительной оценки эффективности разработанных алгоритмов и на основе предыдущего опыта применения метода генетического программирования.

В качестве общего для всех задач критерия останова использовалось условие достижения уровня относительной ошибки моделирования либо выполнение заданного максимального числа вычислений функции пригодности.

Численные эксперименты проводились с помощью разработанной программной системы «Моделирование: гибридный метод генетического программирования», в которой реализованы исследуемые методы [6].

Для каждой задачи максимальное число вычислений функции пригодности для выполнения остановки выбиралось исходя из результатов предварительного исследования метода генетического программирования на этой задаче.

Эффективность методов оценивалась по критерию надежности, который определялся как

отношение числа запусков, в которых была достигнута заданная точность аппроксимации исходных данных, к общему числу запусков. При этом максимальное число вычислений функции пригодности не должно превышать заданное в критерии останова. Статистика для получения оценок надежности набиралась по 50 запускам каждого из рассмотренных методов. Значимость в различиях результатов алгоритмов проверялась методами ANOVA. Проверка выполнялась при уровне значимости α=0,05.

Рис. 1. Блок-схема гибридного алгоритма генетического программирования

Описание тестовых задач приведено в таблице 1. В таблице 2 представлены результаты сравнительного исследования эффективности, полученные для рассматриваемых модификаций метода генетического программирования на тестовых задачах. Жирным шрифтом выделен метод, победивший на тестовой задаче, то есть статистически значимо превосходящий по надежности конкурирующий метод.

Разработанный гибридный метод генетического программирования на шести из семи рассмотренных тестовых задачах оказался значительно эффективнее стандартного подхода.

Одной из практических задач, на которой апробировались разработанные методы и оценивалась их эффективность, являлась задача моделирования процесса рудно-

Таблица 1

Тестовые задачи

№ задачи	Моделируемая функция	Интервал варьирования переменных	Функциональное множество	Объем выборки
1	$y = \sin(x)$	$x \in [-3;4]$	$\{+,-,\times,/\}$	100
2	$y = x^2 + 2x + 3$	$x \in [-3;4]$	$\{+,-,\times,/\}$	100
3	$y = x_1^2 + x_2^2$	$x_1, x_2 \in [-4;4]$	$\{+,-,\times,/\}$	200
4	Функция Растригина: $y = 0,1x_1^2 + 0,1x_2^2 - 4\cos(0,8x_1) - 4\cos(0,8x_1) + 8$	$x_1, x_2 \in [-3;3]$	$\{+,-,\times,/,\cos,\sin,\sqrt{},\exp\}$	200
5	$y = x_1^2 \sin(x_1) + x_2^2 \sin(x_2)$	$x_1, x_2 \in [-4;4]$	$\{+,-,\times,/,\cos,\sin,\sqrt{},\exp\}$	200
6	Функция Розенброка: $y = 100\left(x_2 - x_1^2\right)^2 - \left(1 - x_1\right)^2$	$x_1, x_2 \in [-2;2]$	$\{+,-,\times,/\}$	200
7	$y = x_1^2 + x_1 x_2 + x_2^2$	$x_i \in [-2;2], i = \overline{1,3}$	$\{+,-,\times,/\}$	300

термической плавки (РТП) [7].

Исходные данные задачи следующие: имеются выборки данных, характеризующих эффективность работы печи рудно-термической плавки. В качестве управляющих параметров (входных воздействий) $\bar{x}_i$, $i = \overline{1,9}$, используются электрические параметры и загрузка шихты по отдельным составляющим, так как именно эти входные параметры влияют на процессы в печи и, кроме того, предоставляют возможность непрерывно получать достоверную информацию о них.

Таблица 2

Результаты тестирования

Тестовая задача	Стандартный МГП	Гибридный МГП
1	0,6	0,95
2	0,9	1
3	0,5	0,9
4	0,45	0,95
5	0,95	1
6	0,6	0,95
7	0,45	0,85

Известно, что существует зависимость, связывающая входные и выходные параметры процесса РТП:

$$y = f(x_1, x_2, \ldots, x_9). \quad (2)$$

Однако вид и структура этой зависимости неизвестны. Необходимо по имеющимся выборкам наблюдений восстановить в символьном виде функциональную зависимость, характеризующую процесс РТП:

$$y = \tilde{f}(x_1, x_2, \ldots, x_9). \quad (3)$$

В качестве критерия, позволяющего оценить построенную регрессионную модель, выберем критерий (1). Очевидно, что повышение качества получаемой регрессионной модели требует минимизации этого критерия:

$$W \to \min. \quad (4)$$

Значения параметров метода генетического программирования для решения предложенной задачи были определены на основе предварительного анализа эффективности работы рассматриваемого метода на тестовых задачах.

Выборка наблюдений за процессом РТП состоит из 47 элементов. При решении задачи выборка была разбита случайным образом на две части: на обучающую выборку, включающую в себя 37 элементов, и на экзаменующую, состоящую из 10 элементов.

Лучшее решение, найденное с помощью разработанного подхода, имеет следующий вид:

$$y = \left(x_1 - 8,917 + \sin(x_2) + \sin(\sin(x_4 + x_4))\right)^{1/2} \times$$

$$\times \begin{pmatrix} \sqrt{x_4} + x_4 + \sin(x_2 x_9) + 2,45\sin(x_1 x_2) + \\ + \cos(x_9 + x_9) - \sin(x_2) - 3\sin(x_4 + x_4) - \\ - \left(\sin(x_7) / \sin(\sin(\sqrt{x_2} / 700))^{1/2}\right)^{1/2} \end{pmatrix}^{-1}.$$

Для обучающей выборки относительная ошибка оценивания составила 8,8 %, а для

экзаменующей выборки относительная ошибка прогнозирования составила 10,2 %. Результаты моделирования предложенным алгоритмом генетического программирования сопоставимы с результатами нейросетевого моделирования, полученными группой аналитиков [7].

Сравнительный график прогнозируемой концентрации никеля, вычисленной на модели, и реальных данных в точках обучающей выборки, приведен на рисунке 2.

Сравнительный график концентрации никеля, спрогнозированной с помощью модели, и реальных данных в точках экзаменующей выборки приведен на рисунке 3.

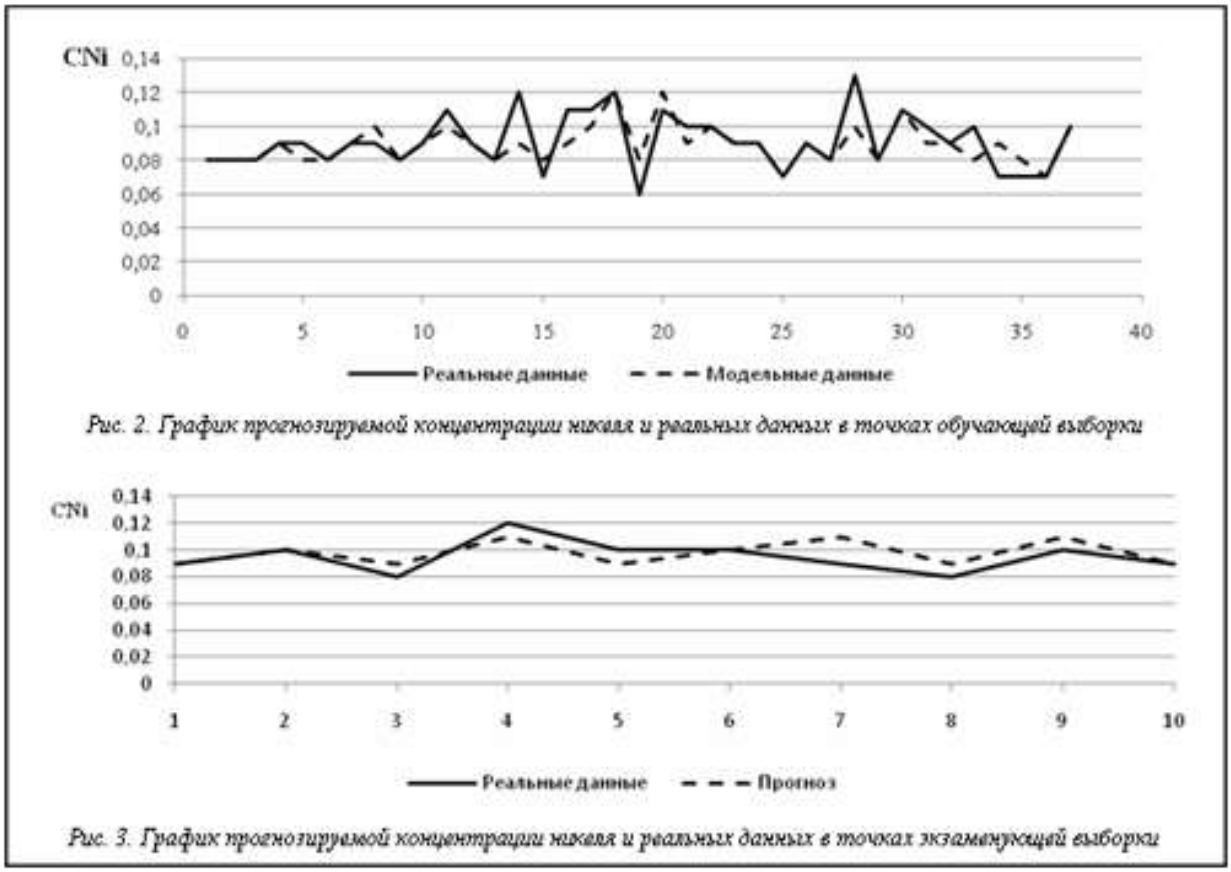

Рис. 2. График прогнозируемой концентрации никеля и реальных данных в точках обучающей выборки

Рис. 3. График прогнозируемой концентрации никеля и реальных данных в точках экзаменующей выборки

Полученные результаты решения задачи моделирования процесса РТП можно считать вполне удовлетворительными. Достигнутое значение относительной ошибки прогноза концентрации никеля в отработанном шлаке равно 10 %, что является достаточно хорошим показателем ввиду чрезвычайной ограниченности объема выборок наблюдений, отсутствия какой-либо априорной информации о структуре модели исследуемого процесса, о законах распределения помех. Все это, безусловно, затрудняет процесс построения модели РТП. Тем не менее, разработанный гибридный алгоритм генетического программирования доказал свою работоспособность в условиях ограниченности вычислительных ресурсов и может использоваться при необходимости быстро получить с приемлемой точностью аналитический вид зависимости, связывающей входные и выходные переменные, и построить прогноз состояния исследуемого процесса, значений его выходных переменных.

В целом анализ результатов решения рассмотренных задач позволяет утверждать, что совместное использование предложенного гибридного алгоритма генетического программирования и эволюционных алгоритмов настройки параметров модели позволило создать перспективный метод построения символьных регрессионных моделей,

включающий в себя средства как структурной, так и параметрической оптимизации модели.

Развитие используемого подхода к решению задачи символьной регрессии и совместное его использование с другими методами автоматического проектирования интеллектуальных технологий является перспективным направлением повышения эффективности автоматизации построения моделей объектов и систем.

Литература

1. Koza John R. The Genetic Programming Paradigm: Genetically Breeding Populations of Computer Programs to Solve Problems. Cambridge, MA: MIT Press, 1992.

2. Holland J.H. Adaptation in natural and artificial systems. Ann Arbor. MI: University of Michigan Press, 1975.

3. Goldberg D.E. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley, 1989.

4. Бежитский С.С., Семенкин Е.С., Семенкина О.Э. Гибридный эволюционный алгоритм для задач выбора эффективных вариантов систем управления // Автоматизация и современные технологии. 2005. № 11. С. 24–31.

5. Бухтояров В.В. Разработка гибридного метода генетического программирования // Решетневские чтения: матер. XII Междунар. науч. конф., посвященной памяти генерального конструктора ракетно-космических систем акад. М.Ф. Решетнева (ноябрь 2008 г., Красноярск). Красноярск: РИО СибГАУ, 2008.

6. Бухтояров В.В. Моделирование: гибридный алгоритм генетического программирования // Свид. о гос. рег. № 2010613317. М.: ФИПС, 2010.

7. Гонебная О.Е. Экспертная система рудно-термической плавки: дис. … канд. техн. наук. Красноярск: ГУЦМиЗ, 2004. 136 с.