ISSN 2782-2001
Системы анализа
и обработки данных
том 94, № 2, 2024, с. 85–93

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 94, No. 2, 2024, pp. 85–93

ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ
И ТЕЛЕКОММУНИКАЦИИ

INFORMATION
TECHNOLOGIES
AND TELECOMMUNICATIONS

УДК 004.94

DOI: 10.17212/2782-2001-2024-2-85-93

Анализ компромисса между точностью и сложностью идентифицированных моделей динамических систем*

Л. ЧЖАН

105005, РФ, г. Москва, ул. 2-я Бауманская, 5, Московский государственный технический университет

chzhanl2@student.bmstu.ru

Одним из интенсивно развивающихся направлений современной теории управления является идентификация систем, связанная с построением математических моделей систем в виде совокупности математических соотношений, адекватно отражающих основные свойства системы. Всё большую популярность в задачах структурно-параметрической идентификации систем на основе доступных наблюдений и экспериментальных данных находят методы символьной регрессии, позволяющие строить регрессионные модели в виде кодов математических выражений в символьной форме. Среди известных численных эволюционных методов символьной регрессии общепризнанным «фаворитом» является метод генетического программирования, применение которого позволяет описать поиск решения задачи как построение регрессионной модели путем перебора различных произвольных суперпозиций функций из некоторого заранее заданного набора. При этом важными показателями, определяющими качество идентификации математической модели системы, является точность и сложность идентифицированной модели. Нередко полученные в результате решения задачи идентификации модели системы недостаточно точны или избыточно сложны. В результате решение задачи идентификации неразрывно связано с обеспечением достаточной точности и простоты идентифицированной модели. В связи с этим естественно придерживаться принципа сбалансированной идентификации, который указывает на поиск компромисса между точностью воспроизведения и мерой сложности идентифицированной модели. Целью настоящей работы, развивающей концепцию сбалансированной идентификации, является анализ компромисса между точностью и сложностью моделей динамических систем, идентифицированных методом генетического программирования. В работе вводится в рассмотрение функционал «точность – сложность», позволяющий при решении задачи идентификации вычислять баланс компромисса между данными ключевыми показателями идентифицированных моделей. Эффективность предложенного функционала демонстрируется на примере компьютерной идентификации методом генетического программирования динамической системы Лоренца.

Ключевые слова: идентификация динамических систем, построение математической модели, символьная регрессия, генетическое программирование, компромисс между точностью и сложностью, фитнес-функция, функционал «точности – сложности», сбалансированная идентификация

* *Статья получена 19 февраля 2024 г.*

ВВЕДЕНИЕ

При решении задач системного анализа, управления и обработки информации в различных областях одной из ключевых является проблема идентификации, связанная с получением или уточнением по экспериментальным данным математической модели реальных динамических систем (объектов, явлений, процессов) [1]. Результаты решения задачи идентификации являются исходными данными для проектирования систем управления, оптимизации, исследования, прогноза, анализа параметров систем и т. д.

Поскольку дифференциальное уравнение играет важную роль в описании систем в области технических исследований, задача идентификации может быть сведена к нахождению правых частей этих уравнений, описывающих поведение реальной системы [2]. До последнего времени большинство методов идентификации, таких как линейная регрессия и полиномиальный метод, не позволяли найти структуру функций исследуемой системы. Структура функций обычно задается исследователем, затем оптимальные величины параметров получают численными методами по разными критериям соответствия между моделью и экспериментальными данными. С появлением метода символьной регрессии, основанного на генетическом программировании, стало возможным в задачах идентификации искать одновременно и параметры функций, и их структуры [3].

Символьная регрессия – это мощная техника для нахождения регрессионного уравнения в символьной форме путем перебора суперпозиций заранее заданного набора функций с помощью эволюционных вычислений. Полученные данные аппроксимируются подходящим математическим уравнением [4]. В настоящее время известные методы символьной регрессии основаны на применении эволюционных алгоритмов и машинного обучения. Главное различие между этими методами заключается в способе представления результатов и используемом алгоритме оптимизации [5, 6].

Идею решения различных задач с помощью символьной регрессии посредством генетического программирования (ГП, genetic programming) впервые представил Джон Коза [7, 8]. Генетическое программирование является общепризнанным «фаворитом» среди численных эволюционных методов символьной регрессии для решения задачи построения математической модели динамических систем, особенно когда уравнения системы полностью неизвестны и доступны только временные ряды, отражающие процесс эволюции системы [9, 10].

В более ранней работе над решением проблемы идентификации динамических систем доказана эффективность метода ГП и обнаружено, что по мере увеличения поколения эволюций получаемая модель становится более точной и в то же время более сложной [10]. Одновременно результат Парето-фронт показывает, что точность модели быстро возрастает при некоторой минимальной сложности, затем улучшается лишь незначительно с более сложными уравнениями, т. е. между точностью и сложностью существует компромисс. Метод, который ищет компромисс между точностью и сложностью, называется также методом сбалансированной идентификации [11].

Анализ компромисса между точностью и сложностью уже использовался для идентификации моделей на основе обработки набора измерений характеристик моделируемого явления и при идентификации моделей Винера – Гаммерштейна [12]. Многокритериальный генетический алгоритм, который позволяет несколькими критериями влиять на выбор структур – кандидатов

модели, применяется при идентификации полиномиальных моделей для реальной нелинейной системы и позволяет лучше выбирать окончательную идентифицированную модель [13]. В статье ставится цель провести анализ компромисса между точностью и сложностью идентифицированных моделей при идентификации методом ГП.

Статья состоит из трех разделов. В первом разделе рассматривается постановка задачи идентификации и применение ГП для решения задачи построения математической моделей. Во втором разделе представлен функционал «точности – сложности», используемый для оценивания моделей во время эволюции. В третьем разделе представлены результаты моделирования предлагаемого функционала.

1. ИДЕНТИФИКАЦИЯ ДИНАМИЧЕСКИХ СИСТЕМ МЕТОДОМ ГП

Задача идентификации также называется задачей построения математической модели. Она заключается в извлечении внутренней базовой модели системы из полученных экспериментальных или наблюдаемых данных. При исследовании динамической системы совокупность дифференциальных уравнений часто используется для описания физических явлений, происходящих в системе. Проблему идентификации можно сформулировать как извлечение неизвестной функции $\mathbf{f}$ по временным данным векторов состояния системы $\mathbf{X} = (\mathbf{x}(t_1), \mathbf{x}(t_2), ...\mathbf{x}(t_m))^T$, подробно см. [10]. В данном случае кроме точности сложность модели также является важным элементом оценивания эффективности полученной модели при идентификации, поскольку эффективность управления будет в значительной степени зависеть от полученной модели.

Алгоритм ГП, предложенный John Koza в 1992 г. [7], является разновидностью эволюционных вычислений. Он основан на дарвиновском естественном отборе, в результате которого получаются новаторские решения к задаче. Данный алгоритм представлен на рис. 1 [8].

Рис. 1. Общая структура алгоритма генетического программирования

Fig. 1. General structure of the genetic programming algorithm

Общая структура алгоритма ГП состоит из следующих шагов: для начала требуется временный ряд данных состояния системы $\mathbf{X}$ и его производная $\dot{\mathbf{X}}$, затем по заранее заданным параметрам алгоритма необходимо создать начальную популяцию особей – случайные потенциальные решения (каждая особь – возможное решение задачи), оценить приспособленность каждого решения по значению фитнес-функции, выбрать наилучшие решения и создать из них популяцию нового поколения особей путем генетических операций (репродук-

ции, скрещи-вания и мутации), проверить условия завершения эволюции (до максимального числа поколений или ошибка по фитнес-функции менее заданного значения), принять решение об окончании либо повторении итерации поиска [10]. В результате получается наилучшее решение задач.

2. ФУНКЦИОНАЛ «ТОЧНОСТИ – СЛОЖНОСТИ»

Идентификация системы может рассматриваться как задача оптимизации, и мерой, по которой определить, насколько хорошо модели-кандидаты подходят к идентифицируемой системе, является точность модели. Единый стандарт оценки модели точностью часто приводит к особенно сложной модели, при практических применениях такая модель не может отвечать требованиям анализа и управления системой, а среди множества показателей эффективности модели точность и сложность являются самыми важными [13]. В связи с этим вместо первоначального метода оценки модели точностью предлагается функционал Q для оценивания эффективности модели:

$$Q = Q_{\text{точ}} + \lambda Q_{\text{слож}}, \tag{1}$$

где Q – новый показатель эффективности модели; $Q_{\text{точ}}$ – показатель точности модели; $Q_{\text{слож}}$ – сложность модели; λ – параметр, обеспечивающий баланс между точностью и сложностью модели.

Точность относится к способности модели достоверно представлять моделируемую систему, другими словами, разницу между реакциями реальной системы и модели [9]. В настоящей работе в качестве меры точности модели $Q_{\text{точ}}$ выберем среднюю абсолютную ошибку. Чем ближе модель к системе, тем выше ее точность (тем меньше показатель):

$$Q_{\text{точ}} = MAE = \frac{1}{m}\sum_{i=1}^{m}\left|x_i - \tilde{x}_i\right|, \tag{2}$$

где x_i – измерение (экспериментальные данные); $\tilde{x}_i$ – расчетные данные по полученной модели; m – количество экспериментальных данных.

Сложность модели $Q_{\text{слож}}$ выражается в сложности полученных уравнений, поскольку результат работы алгоритма ГП обычно представлен в виде дерева. Существует несколько интуитивных мер для определения сложности полученных моделей: количество листьев или слоев дерева [14]. Выберем длину уравнений (сумма узлов и листьев) для определения сложности модели. Чем больше значение показателя, тем сложнее модель:

$$Q_{\text{слож}} = len(\tilde{\mathbf{f}}), \tag{3}$$

где $\tilde{\mathbf{f}}$ – полученная функция модели для описания исследуемой системы с неизвестной функцией $\mathbf{f}$; *len* – символ сложности модели.

Из формулы (1) видно, что при $\lambda = 0$ сложность модели не влияет на результат моделирования, по мере увеличения λ влияние сложности модели на результат возрастает. В таком случае задача идентификации динамических систем формулируется следующим образом: для заданного сбора данных $\mathbf{X}$ и $\dot{\mathbf{X}}$ найти функцию $\tilde{\mathbf{f}}$, обеспечивающую экстремум функционала Q, баланс

между (2) и (3). Предложенный функционал учитывает принцип сбалансированной идентификации, достигает уменьшения сложности модели при обеспечении точности модели с помощью параметра регуляризации.

3. РЕЗУЛЬТАТ И ОБСУЖДЕНИЕ

В качестве тестирования выбрана типичная динамическая система Лоренца [15]. Параметры алгоритма ГП представлены в табл. 1.

Таблица 1

Table 1

Параметры настройки алгоритма ГП

GP algorithm settings

Количество поколений	20
Размер популяции	5000 шт.
Терминальное множество	$T = \{x, y, x, R\}$
Функциональное множество	$F = \{+, -, *, /, \sin, \cos\}$
Метод инициализации	Комбинированный метод (halfandhalf)
Вероятность кроссинговера	0.7
Вероятность мутации	0.1
Фитнес-функция	Функционал точности-сложности

База данных временных рядов состояний системы X и соответствующих производных $\dot{X}$ получена путем моделирования методом Рунге – Кутты в среде Matlab. Для нахождения модели системы в форме дифференциальных уравнений используется алгоритм ГП [7, 8], который реализован на основе пакета DEAP (англ. Distributed Evolutionary Algorithms in Python) в среде Python. При оценке приспособленности полученных моделей применен функционал Q, для того чтобы исследовать баланс между точностью и сложностью модели. Полученная наилучшая модель (или субоптимальная) имеет наименьшую ошибку идентификации с учетом простоты структуры модели.

В табл. 2 представлен результат моделирования параметра x системы Лоренца.

Таблица 2

Table 2

Параметры наилучшей модели при разных значениях λ

Parameters of the best model for different values λ

λ	N (номер поколения)	Ошибка	Длина решения
0.1	9	1.30368e-15	41
0.2	8	2.30882e-15	45
0.3	8	2.55232e-15	42
0.4	10	2.63871e-15	42
0.5	10	2.88931e-15	40
0.6	16	1.75503e-15	39
0.7	19	13.1784	6
0.8	19	10.5308	6
0.9	19	26.5903	3
1.0	19	26.5903	3

Результаты показывают, что с увеличением параметра коэффициента λ для оценивания приспособляемости полученных моделей в процессе идентификации методом ГП полученная наилучшая модель в ходе одного поколения действительно становится проще (длина решения сокращалась), и чем выше значение параметра λ, тем проще полученное уравнение. Однако простота модели достигается ценой снижения точности модели, увеличение значения λ также приводит к снижению точности модели (ошибка модели увеличилась). В данном исследовании сложность полученной модели быстро снижается при $\lambda = 0,7$.

Точность и сложность полученных моделей во время эволюции при разных значениях параметра λ представлены на рис. 2, где видно, что общая тенденция линий не изменяется: по мере увеличения числа поколения эволюции модель становится сложнее и точнее. При $\lambda < 0,6$ влияние параметра почти незаметно. Кривая сильно изменилась при $\lambda > 0,6$: диапазон значений сложности модели сильно сокращен, и изменение точности замедляется.

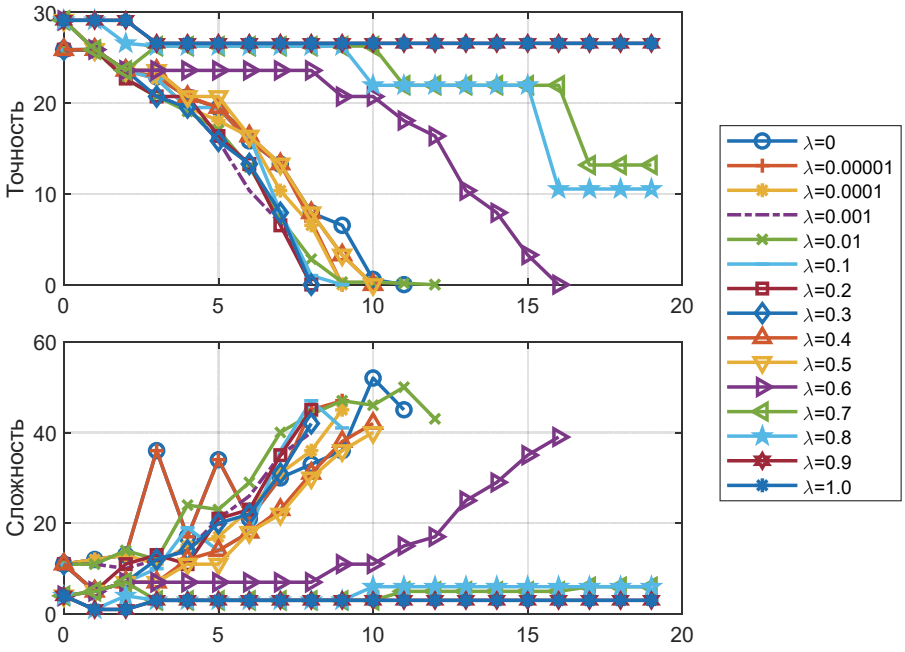

Рис. 2. Результат точности и сложности моделей во время эволюционного поиска при разных значениях параметра λ

Fig. 2. Result of model accuracy and complexity during evolutionary search for different values of the parameter λ

Стоит отметить, что при небольших значениях наилучшая модель часто обнаруживается при меньшем числе поколений (примерно в 10-м поколении), при больших значениях удовлетворительные результаты часто не получаются до конца цикла (т. е. после достижения максимального числа поколений).

ЗАКЛЮЧЕНИЕ

В работе исследован компромисс между точностью и сложностью при решении задачи построения математических моделей динамических систем на основе наблюдаемых данных алгоритмом ГП. Предложен функционал «точности – сложности» для балансирования точности и сложности идентифицированных моделей при решении задач идентификации. Важно подчеркнуть, что подход, используемый в настоящей работе, направлен на получение интерпретируемой модели для понимания свойства систем и дальнейшего управления исследуемой системой. Другими словами, полученная модель должна иметь форму, дающую базовую информацию о структуре системы, и одновременно быть не излишне сложной. Результат показывает, что точность и сложность являются двумя важными принципами, характеризующими сбалансированную идентификацию. При решении различных задач необходимо учитывать баланс между ними.

СПИСОК ЛИТЕРАТУРЫ

1. *Esfandiari R.S., Lu B.* Modeling and analysis of dynamic systems. – Boca Raton, FL: CRC Press, 2014. – 558 p.

2. *Ljung L.* System identification: theory for the user. – 2nd ed. – Upper Saddle River, NJ: Prentice-Hall, 1999. – 631 p. – (Prentice Hall information and system sciences series).

3. *Данг Т.Ф., Дивеев А.И., Софронова Е.А.* Решение задач идентификации математических моделей объектов и процессов методом символьной регрессии // Cloud of Science. – 2018. – Т. 5, № 1. – С. 147–162.

4. Prediction of dynamical systems by symbolic regression / M. Quade, M. Abel, K. Shafi, R.K. Niven, B.R. Noack // Physical Review E. – 2016. – Vol. 94 (1). – P. 012214.

5. *Дивеев А.И., Шмалько Е.Ю.* Современные методы символьной регрессии и их модификации (обзор) // Вопросы теории безопасности и устойчивости систем. – М., 2018. – № 20. – С. 133–158.

6. *Карасева Т.С.* Эволюционные алгоритмы решения задач символьной регрессии для идентификации динамических систем: дис. … канд. техн. наук. – Красноярск, 2023. – 128 с.

7. *Koza J.R.* Genetic programming: On the programming of computers by means of natural selection. – The MIT Press, 1992. – 836 p.

8. *Poli R., Langdon W.B., McPhee N.F.* A field guide to genetic programming / with contributions by J.R. Koza. – GPBiB, 2008. – 252 p.

9. *Дивеев А.И., Софронова Е.А.* Метод генетического программирования с сетевым оператором для идентификации систем управления // Вестник Донского государственного технического университета. – 2010. – Т. 10, № 5 (48). – С. 624–634.

10. *Чжан Л., Филимонов Н.Б.* Идентификация динамических систем на основе обработки экспериментальных данных методом генетического программирования // Journal of Advanced Research in Natural Science. – 2023. – № 18. – С. 4–12.

11. *Соколов А.В., Волошинов В.В.* Выбор математической модели: баланс между сложностью и близостью к измерениям // International Journal of Open Information Technologies. – 2018. – Т. 6, № 9. – С. 33–41.

12. WH-MOEA: A multi-objective evolutionary algorithm for Wiener-Hammerstein system identification. A novel approach for trade-off analysis between complexity and accuracy / J. Zambrano, J. Sanchis, J.M. Herrero, M. Martinez // IEEE Access. – 2020. – Vol. 8. – P. 228655–228674.

13. *Fonseca C.M., Fleming J.P.* Non-linear system identification with multiobjective genetic algorithms // IFAC Proceedings. – 1996. – Vol. 29 (1). – P. 1169–1174.

14. *Vladislavleva E.J., Smits G.F., Den Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // IEEE Transactions on Evolutionary Computation. – 2008. – Vol. 13 (2). – P. 333–349.

15. *Lorenz E.N.* Deterministic nonperiodic flow // Journal of the Atmospheric Sciences. – 1963. – Vol. 20. – P. 130–141.

DOI: 10.17212/2782-2001-2024-2-85-93

Analysis of the trade-off between accuracy and complexity of identified models of dynamic systems *

L. ZHANG

Bauman Moscow State Technical University, 5, 2-ya Baumanskaya Street, Moscow, 105005, Russian Federation

chzhanl2@student.bmstu.ru

Abstract

One of the rapidly developing areas of the modern control theory is the identification of systems associated with the construction of mathematical models of systems in the form of a set of mathematical relationships that adequately reflect the basic properties of the system. Symbolic regression methods are becoming more and more popular in the problems of structural-parametric identification of systems based on available observations and experimental data, which allow building regression models in the form of codes of mathematical expressions in a symbolic form. Among the known numerical evolutionary methods of symbolic regression, the universally acknowledged "favorite" is the method of genetic programming", the application of which allows us to describe the search for solving a problem as the construction of a regression model by enumerating various arbitrary superpositions of functions from some predetermined set. In this case, the important indicators determining the quality of identification of the mathematical model of the system are the accuracy and complexity of the identified model. Often, the system model obtained as a result of solving identification problem is not accurate enough or excessively complex. As a result, the solution to the identification problem is inextricably linked to ensuring sufficient accuracy and simplicity of the identified model. In this connection, it is natural to adhere to the principle of balanced identification, which indicates the search for a compromise between the accuracy of reproduction and the measure of complexity of the identified model. The purpose of this paper, which develops the concept of balanced identification, is to analyze the trade-off between the accuracy and complexity of models of dynamic systems identified by genetic programming. In this paper we introduce a functional "accuracy-complexity", which allows us to balance the trade-off between these key indicators of the identified models when solving the identification problem. The effectiveness of the proposed functional is demonstrated on the example of computer identification by genetic programming of a Lorentz dynamical system.

Keywords: identification of dynamic systems, mathematical model building, symbolic regression, genetic programming, trade-off between accuracy and complexity, fitness function, accuracy-complexity functional, balanced identification

REFERENCES

1. Esfandiari R.S., Lu B. *Modeling and analysis of dynamic systems*. Boca Raton, FL, CRC Press, 2014. 558 p.

2. Ljung L. *System identification: theory for the user*. 2nd ed. Upper Saddle River, NJ, Prentice-Hall, 1999. 631 p.

3. Dang T.Ph., Diveev A., Sofronova E. Reshenie zadach identifikatsii matematicheskikh modelei ob"ektov i protsessov metodom simvol'noi regressii [Mathematical models identification of objects and processes by symbolic regression]. *Cloud of Science*, 2018, vol. 5, no. 1, pp. 147–162. (In Russian).

4. Quade M., Abel M., Shafi K., Niven R.K., Noack B.R. Prediction of dynamical systems by symbolic regression. *Physical Review E*, 2016, vol. 94 (1), p. 012214.

5. Diveev A.I., Shmalko E.Yu. Sovremennye metody simvol'noi regressii i ikh modifikatsii (obzor) [Modern methods of symbolic regression and their modifications (review)]. *Voprosy teorii bezopasnosti i ustoichivosti sistem* [Issues in the theory of safety and stability of systems]. Moscow, 2018, no. 20, pp. 133–158.

* *Received 19 February 2024.*

6. Karaseva T.S. *Evolyutsionnye algoritmy resheniya zadach simvol'noi regressii dlya identifikatsii dinamicheskikh sistem*. Diss. kand. tekhn. nauk [Evolutionary algorithms for solving symbolic regression problems for identification of dynamical systems. Dr. eng. sci. diss.]. Krasnoyarsk, 2023. 128 p.

7. Koza J.R. *Genetic programming: On the programming of computers by means of natural selection.* The MIT Press, 1992. 836 p.

8. Poli R., Langdon W.B., McPhee N.F. *A field guide to genetic programming*. With contributions by J.R. Koza. GPBiB, 2008. 252 p.

9. Diveyev A.I., Sofronova E.A. Metod geneticheskogo programmirovaniya s setevym operatorom dlya identifikatsii sistem upravleniya [Genetic programming method for control systems identification]. *Vestnik Donskogo gosudarstvennogo tekhnicheskogo universiteta = Vestnik of Don State Technical University* 2010, vol. 10, no. 5 (48), pp. 624–634.

10. Zhang L., Filimonov N.B. Identifikatsiya dinamicheskikh sistem na osnove obrabotki eksperimental'nykh dannykh metodom geneticheskogo programmirovaniya [Identification of dynamic systems based on experimental data processing using genetic programming]. *Journal of Advanced Research in Natural Science*, 2023, no. 18, pp. 4–12. (In Russian).

11. Sokolov A.V., Voloshinov V.V. Vybor matematicheskoi modeli: balans mezhdu slozhnost'yu i blizost'yu k izmereniyam [Choice of mathematical model: balance between complexity and proximity to measurements]. *International Journal of Open Information Technologies*, 2018, vol. 6, no. 9, pp. 33–41. (In Russian).

12. Zambrano J., Sanchis J., Herrero J.M., Martinez M. WH-MOEA: A multi-objective evolutionary algorithm for Wiener-Hammerstein system identification. A novel approach for trade-off analysis between complexity and accuracy. *IEEE Access*, 2020, vol. 8, pp. 228655–228674.

13. Fonseca C.M., Fleming J.P. Non-linear system identification with multiobjective genetic algorithms. *IFAC Proceedings*, 1996, vol. 29 (1), pp. 1169–1174.

14. Vladislavleva E.J., Smits G.F., Den Hertog D. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*, 2008, vol. 13 (2), pp. 333–349.

15. Lorenz E.N. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 1963, vol. 20, pp. 130–141.

Для цитирования:

Чжан Л. Анализ компромисса между точностью и сложностью идентифицированных моделей динамических систем // Системы анализа и обработки данных. – 2024. – № 2 (94). – С. 85–93. – DOI: 10.17212/2782-2001-2024-2-85-93.

For citation:

Zhang L. Analiz kompromissa mezhdu tochnost'yu i slozhnost'yu identifitsirovannykh modelei dinamicheskikh system [Analysis of the trade-off between accuracy and complexity of identified models of dynamic systems]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2024, no. 2 (94), pp. 85–93. DOI: 10.17212/2782-2001-2024-2-85-93.