

Сравнительный анализ современных методов символьной регрессии в задачах идентификации динамических систем по данным наблюдений

Л. Чжан

Московский государственный технический университет им. Баумана

Аннотация: В современных исследованиях символьная регрессия представляет собой мощный инструмент построения математических моделей различных систем. В данной статье применяются и сравниваются три метода символьной регрессии: генетическое программирование, разреженная идентификация нелинейных динамик и гибридный метод. Эффективность каждого метода оценивается по способности находить точные модели с высокой точностью и низкой сложностью при наличии различных уровней шума в данных наблюдений. По результатам проведенных экспериментов сделан вывод, что наилучшим методом идентификации динамических систем является гибридный метод, который сочетает генетическое программирование и разреженную идентификацию.

Ключевые слова: символьная регрессия, идентификация динамических систем, генетическое программирование, разреженная идентификация нелинейных динамик, гибридный метод.

Введение

Проблемы идентификации, связанные с разработкой методов и технологий математического моделирования систем, процессов и явлений, занимают одно из центральных мест в современной теории управления [1]. Традиционные методы параметрической идентификации требуют, чтобы структура модели была задана исследователем заранее, а затем для поиска оптимальной модели используются методы оценки параметров [2]. Однако во многих современных системах динамика системы обычно неизвестна, что требует построения модели непосредственно из наблюдаемых данных, другими словами, идентификация не является параметрической [3]. Символьная регрессия (SR, Symbolic Regression) представляет собой мощный метод решения подобных задач.

SR — это класс методов, позволяющих находить математические модели в символьном виде, не требуя заранее определенной структуры модели [4]. Интерпретируемость символьных форм делает эти методы особенно важными в области моделирования с помощью машинного

обучения [5]. В последние несколько лет интерпретируемость моделей вызвала значительный интерес в различных исследовательских областях. В общем случае модель считается интерпретируемой, если связь между ее входами и выходами может быть кратко описана логическим или математическим способом. Традиционно SR реализуется с помощью генетического программирования (GP, Genetic Programming), эволюционный алгоритм, вдохновленный принципами естественного отбора и выживания сильнейших в дарвиновской теории эволюции [6,7].

В последнее время SR привлекает все большее внимание в качестве инструмента для построения моделей на основе данных [8]. Эволюция методологии SR связана с интеграцией GP, методов глубокого обучения и других инновационных подходов [9-11]. В основе нашей предыдущей работы лежала задача преодоления компромисса между точностью и сложностью моделей, отмеченного в [12]. Для решения этой проблемы был предложен новый гибридный метод под названием генетическое программирование с разреженной идентификацией (GP-SINDy, genetic programming with sparse identification), который позволяет эффективно строить математические модели динамических систем [13]. В указанной работе, однако, не учитывалось влияние шумов на процесс идентификации, в связи с этим в данной статье проводится сравнительный анализ трех методов SR, направленный на решение задачи идентификации динамических систем в условиях зашумленных данных наблюдений.

Постановка задачи идентификации динамических систем

Задача идентификации динамических систем заключается в построении математической модели поведения системы по экспериментальным данным [14]. Описание с помощью дифференциальных уравнений – это широко распространенный способ представления таких моделей. В этой связи данная

работа посвящена идентификации динамических процессов в виде дифференциальных уравнений.

Задача идентификации динамических систем сформулируем следующим образом. Допустим, их динамика в переменных состояния описывается посредством системы обыкновенных дифференциальных уравнений:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}), \quad (1)$$

где $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_n(t)]$ - вектор состояния системы в момент времени t ; $\mathbf{f}(\square)$ - вектор-функция, определяющая как состояния $\mathbf{x}$ эволюционируют с течением времени; $\boldsymbol{\theta}$ - вектор неизвестных параметров.

Исходными данными являются результаты наблюдений за поведением системы, а именно матрица $\mathbf{X} = (\mathbf{x}(t_1), \mathbf{x}(t_2), \dots, \mathbf{x}(t_m))^T$, строки которой представляют вектор состояния $\mathbf{x}(t)$ в моменты времени $t = t_1, t_2, \dots, t_m$ в виде матрицы $\mathbf{X} = (\mathbf{x}(t_1), \mathbf{x}(t_2), \dots, \mathbf{x}(t_m))^T$. Ставится задача идентификации по данным наблюдений $\mathbf{X}$ математической модели системы в виде уравнения (1), т.е. найти неизвестную функцию системы $\mathbf{f}(\square)$. Качество идентификации может оцениваться различными подходами, такими как средняя абсолютная ошибка и др.

Современный методы символьной регрессии в задачах идентификации

В настоящее время в задачах идентификации динамических систем по данным наблюдений широкое распространение получили следующие три метода регрессии: GP, метод разреженной идентификации нелинейных динамик (SINDy, Sparse Identification of Nonlinear Dynamic) и метод GP-SINDy, который сочетает преимущества обоих подходов, структура модели формируется с помощью GP, после чего ее параметры уточняются методом

SINDy. Рассмотрим особенности и алгоритмическую реализацию всех трех методов SR применительно к задачам идентификации.

GP – эволюционный алгоритм, который строит и тестирует функции-кандидаты из простых строительных блоков. Функции-кандидаты обновляются от поколения к поколению генетическими операторами, пока не будет достигнут заданный критерий завершения. Стандартная реализация GP включает четыре основных этапа: инициализацию, оценку приспособленности, создание нового поколения и завершение эволюции (Алгоритм 1) [6].

Алгоритм 1. Алгоритм GP

Input: Data collection X and $\dot{X}$

Output: Best-so-far individual

- 1: Randomly create an *initial population* of programs from the available primitives
 - 2: **repeat**
 - 3: *Execute* each program and ascertain its fitness
 - 4: *Select* one or two program(s) from the population with a probability based on fitness to participate in genetic operations
 - 5: Create new individual program(s) by applying *genetic operations* with specified probabilities
 - 6: **until** an acceptable solution is found or some other stopping condition is met (e.g., a maximum number of generations is reached)
 - 7: **return** the best-so-far individual
-

SINDy, разработанный S. Brunton и др., предназначен для поиска управляющих уравнений нелинейных динамических систем на основе данных временных рядов. В основе данного метода лежит применение разреженной регрессии к большой библиотеке функций-кандидатов с целью выделения минимального набора значимых членов, описывающих динамику системы. Его основной принцип заключается в том, что большинство физических систем управляются лишь несколькими доминирующими понятиями.

Чтобы определить функцию $f(\square)$ по данным, формируется библиотека $\Theta(\mathbf{X}) = \begin{bmatrix} 1 & \mathbf{X} & \mathbf{X}^2 & \dots & \mathbf{X}^d \end{bmatrix}$, включающая все потенциальные члены. И динамическая система может быть представлена в виде как:

$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi, \quad (2)$$

где $\dot{\mathbf{X}}$ - производный $\mathbf{X}$ по времени; $\Xi = \begin{bmatrix} \xi_1 & \xi_2 & \dots & \xi_n \end{bmatrix}$ - матрица коэффициентов, определяющих активные члены в n -й строке.

Алгоритм SINDy представлен в алгоритме 2 [9].

Algorithm 2. Алгоритм SINDy

Input: Time derivative $\dot{\mathbf{X}}$, library of candidate functions $\theta^T(\mathbf{X}, \mathbf{U})$, thresholding parameter ε

Output: Matrix of sparse coefficient vectors Ξ

```

1:  function STLS_REGRESSION ( $\dot{\mathbf{X}}, \theta^T(\mathbf{X}, \mathbf{U}), \varepsilon, N$ )
2:       $\hat{\Xi}^0 \leftarrow (\theta^T)^\dagger \dot{\mathbf{X}}$ 
3:      while not converged do
4:           $k \leftarrow k + 1$ 
5:           $\mathbf{I}_{small} \leftarrow (abs(\hat{\Xi}) < \varepsilon)$ 
6:           $\hat{\Xi}^k(\mathbf{I}_{small}) \leftarrow 0$ 
7:          for all variables do
8:               $\mathbf{I}_{big} \leftarrow \neg \mathbf{I}_{small}(:, ii)$ 
9:               $\hat{\Xi}^k(\mathbf{I}_{big}, ii) \leftarrow (\theta^T(:, \mathbf{I}_{big}))^\dagger \dot{\mathbf{X}}(:, ii)$ 
10:         end for
11:     end while
12: end function

```

Гибридный метод GP-SINDy представляет собой комбинацию GP и SINDy. В этом методе GP используется для построения структуры модели, а затем параметры оцениваются с помощью SINDy.

Метод включает следующие этапы:

1) Сбор данных из исходной системы. Исходные данные $\mathbf{X}$ и соответствующие производные $\dot{\mathbf{X}}$ (если недоступны, вычисляются методом центральных разностей) собираются.

2) Символьная регрессия методом GP. Процесс регрессии выполняется с заранее определенными параметрами эволюционный вычислений, включая функциональное и терминальное множество и т.д.

3) Построение Парето-фронта. На основе результатов регрессии GP (этап 2) осуществляется построения Парето-фронта для каждой переменной системы.

4) Формирование функций-кандидатов из Парето-фронта. Из Парето-фронта извлекаются ключевые структурные части $(\theta_1, \theta_2, \dots, \theta_k)$ модели системы.

5) Разреженная регрессия методом SINDy. Метод SINDy используется для анализа библиотеки функций-кандидатов $\Theta(\mathbf{X}) = [\theta_1 \theta_2 \dots \theta_k]$, полученной на этапе 4, с целью определения соответствующих параметров функций.

6) Построение итоговой модели. Итоговым результатом является символьная модель в виде дифференциальных уравнений, описывающая динамику системы.

Алгоритм гибридного метода проиллюстрирован на рис. 1.

Данный подход применяется к идентификации моделей, в которых структурные уравнения, линейны по параметрам, и одновременно содержат нелинейные функции:

$$\frac{d}{dt}\mathbf{x} = \sum_{i=1}^r c_i F_i(\mathbf{x}), \quad (3)$$

здесь $F_i(\mathbf{x})$ - функции, которые включают в себе основные части (линейные и нелинейные) модели, или структурные части; c_i - соответствующие параметры функций, или параметрические части; r - количество функциональных частей модели.



Рис. 1. – Алгоритм гибридного метода GP-SINDy

Для сравнительной оценки эффективности методов идентификации динамических систем используются показатели точности и сложности модели.

Точность оценивается с помощью средней абсолютной ошибки, вычисляемой по формуле:

$$Q_{\text{точ}} = MAE = \frac{1}{k} \sum_{i=1}^k |y_i - \tilde{y}_i|, \quad (4)$$

где y_i - экспериментальные данные, $\tilde{y}_i$ - смоделированные данные, k - объем выборки.

Сложность для простоты определяется как структурная сложность математического выражения, описываемая следующим образом:

$$Q_{\text{слож}} = \text{len}(\tilde{\mathbf{f}}), \quad (5)$$

где $\tilde{\mathbf{f}}$ - функция, полученная в результате идентификации исследуемой системы, которой с неизвестной функцией $\mathbf{f}(\square)$, $\text{len}()$ - символ сложности модели.

Компьютерная апробация и сравнительный анализ методов

В исследовании рассматриваются три классические динамические системы (система Lorenz, математический маятник и осциллятор Ван дер Поля). Исходные данные были зашумлены добавлением шумов различного уровня (0, 0.001, 0.003, 0.005), распределенных по гауссовскому закону. Данные были разделены на две части: первая половина выступила в качестве обучающей для построения моделей, вторая - для верификации их эффективности.

В таблице № 1-3 представлены результаты моделирования трех систем методом GP, SINDy и GP-SINDy. Результаты включают сложность модели, точность обучения и точность тестирования. Результаты сравнительного анализа различных методов при разных уровнях шума представлены в рис 2-3. В итоге можно заключить, что шум оказывает существенное влияние на точность построенной модели. С увеличением шума точность модели снижается. В отличие от этого, влияние шума на сложность модели не является значительным.

GP: по своей природе, метод GP является стохастическим, поэтому точность и сложность построенных моделей также обладают степенью случайности. Однако общий тренд точности демонстрирует снижение по мере увеличения уровня шума (Рис. 2).

SINDy: при низком уровне шума метод SINDy позволяет получить простые и точные модели. Однако с увеличением уровня шума данный метод, стремясь сохранить высокую точность, жертвует простотой модели, что приводит к ее переобучению (Рис. 2).

GP-SINDy: метод GP-SINDy демонстрирует отличные показатели как по точности модели, так и по ее сложности: обеспечивая простоту модели, он одновременно достигает высокой точности, не проявляя склонности к переобучению (Рис. 3).

Таблица № 1

Результаты моделирования системы Lorenz тремя методом

Шум 0	Метод		GP	SINDy	GP-SINDy
	Сложность	x	10	7	7
		y	12	9	9
		z	11	7	7
	Точность обучения	x	1.0763600	0.0000000	0.0000000
		y	0.7220940	0.0000000	0.0000000
		z	0.2793570	0.0000000	0.0000000
	Точность тестирования	x	1.0798099	0.0000000	0.0000000
		y	0.9266881	0.0000000	0.0000000
		z	0.2793570	0.0000000	0.0000000
Шум 0.001	Сложность	x	15	7	7
		y	17	9	9
		z	7	7	7
	Точность обучения	x	1.2795200	0.5122000	0.4836000
		y	1.7032800	0.5600000	0.5365000
		z	0.9358410	0.5089000	0.4989000
	Точность тестирования	x	1.1309737	0.4840000	0.4836000
		y	1.7790151	0.5373000	0.5365000
		z	0.9466814	0.4999000	0.4989000
Шум 0.003	Сложность	x	10	7	7
		y	5	16	9
		z	11	7	7
	Точность обучения	x	1.7622300	1.3381000	1.4065000
		y	5.1915200	1.5985000	1.6335000
		z	1.4464000	1.4168000	1.4058000
	Точность тестирования	x	1.7411442	1.4063000	1.4065000
		y	4.9736916	1.6361000	1.6335000
		z	1.4290837	1.4065000	1.4058000
Шум 0.005	Сложность	x	13	11	7
		y	12	9	9
		z	12	7	7
	Точность обучения	x	3.7960800	2.3186000	2.1983000
		y	2.5629000	2.4574000	2.5512000
		z	2.6291300	2.4719000	2.3676000
	Точность тестирования	x	3.4953936	2.2006000	2.1983000
		y	2.7188455	2.5532000	2.5512000
		z	2.5713532	2.3705000	2.3676000

Таблица № 2

Результаты моделирования системы маятника тремя методом

Шум 0	Метод		GP	SINDy	GP-SINDy
	Сложность	θ	1	1	1
Шум 0.001	Сложность	w	4	4	4
		θ	0.0000000	0.0000000	0.0000000
	Точность обучения	w	0.0191783	0.0000000	0.0000000
		θ	0.0000000	0.0000000	0.0000000
	Точность тестирования	w	0.0188728	0.0000000	0.0000000
		θ	0.0000000	0.0000000	0.0000000
Шум 0.003	Сложность	θ	1	3	3
		w	4	4	4
	Точность обучения	θ	0.0642710	0.0643000	0.0631000
		w	0.2390750	0.1681000	0.1631000
	Точность тестирования	θ	0.0631081	0.0631000	0.0631000
		w	0.2375925	0.1631000	0.1631000
Шум 0.005	Сложность	θ	1	3	3
		w	5	48	4
	Точность обучения	θ	0.3223270	0.3223000	0.3199000
		w	0.8903350	0.8897000	0.8598000
	Точность тестирования	θ	0.3199386	0.3199000	0.3199000
		w	0.8597484	0.8608000	0.8598000

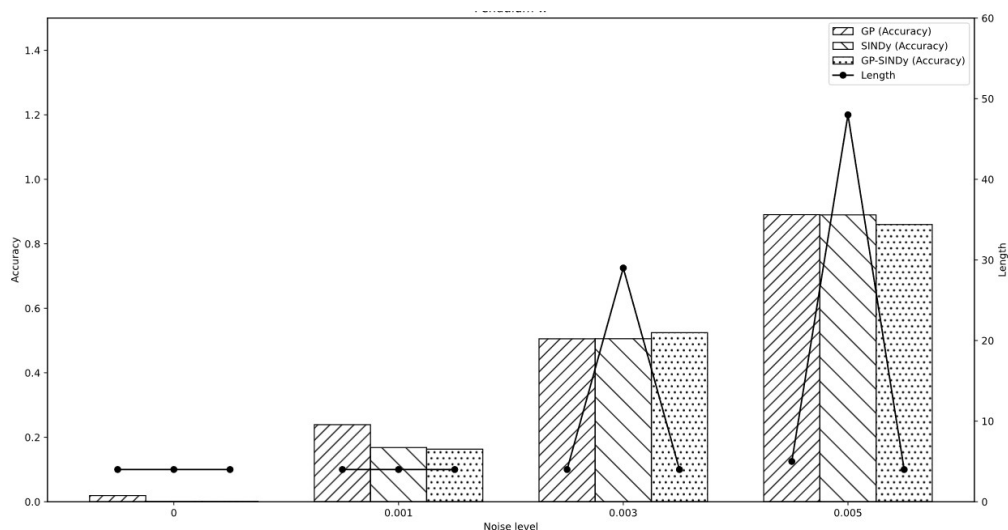


Рис. 2. – Сравнение сложности и точности обучения в системы маятника

Таблица № 3

Результаты моделирования системы Ван дер Поля тремя методом

Шум 0	Метод		GP	SINDy	GP-SINDy
	Сложность	x	1	1	1
Шум 0.001	Сложность	y	21	15	15
		x	0.0000000	0.0000000	0.0000000
	Точность обучения	y	0.6634650	0.0000000	0.0000000
		x	0.0000000	0.0000000	0.0000000
	Точность тестирования	y	0.7945399	0.0000000	0.0000000
		x	0.0869932	0.0870000	0.0865000
Шум 0.003	Сложность	y	21	15	15
		x	0.4251640	0.0890000	0.0905000
	Точность обучения	y	0.0864654	0.0865000	0.0865000
		x	0.5226946	0.0904000	0.0905000
	Точность тестирования	y	1	3	3
		x	25	15	15
Шум 0.005	Сложность	y	0.2555420	0.2555000	0.2697000
		x	0.2726710	0.2725000	0.2599000
	Точность обучения	y	0.2697349	0.2697000	0.2697000
		x	0.2604278	0.2601000	0.2599000
	Точность тестирования	y	1	103	3
		x	23	125	15
Шум 0.005	Сложность	y	0.4537600	0.4535000	0.4516000
		x	0.4728660	0.4688000	0.4734000
	Точность обучения	y	0.4516332	0.5323000	0.4516000
		x	0.4746546	0.6602000	0.4734000
	Точность тестирования	y			
		x			

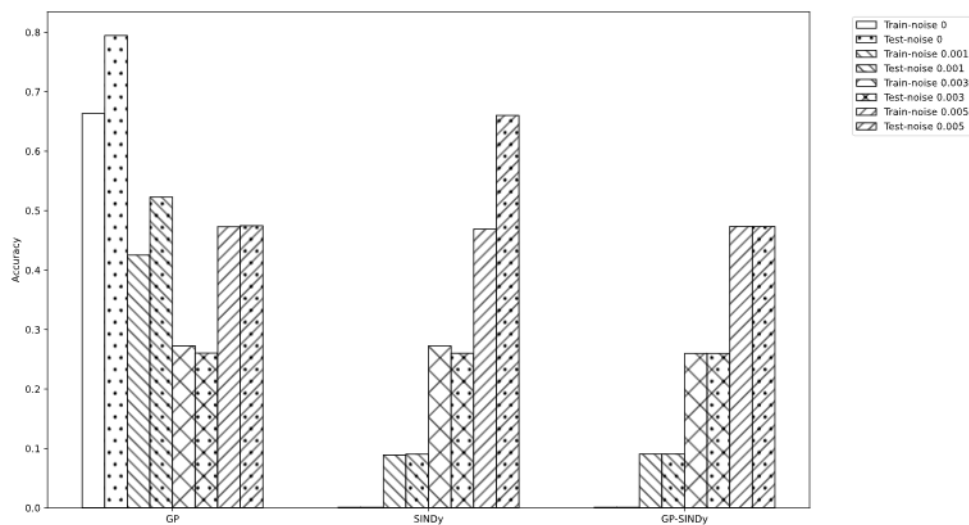


Рис. 3. – Сравнение результатов переменной y системы Ван дер Поля

Заключение

В данной работе проводится сравнение трех методов SR на задачах идентификации динамических систем. Было обнаружено, что для всех трех задач GP-SINDy превосходит как по точности, так и по сложности. Модель, полученная SINDy, оказалась почти такой же точной, но не совсем такой простой, когда в данных присутствуют шумы.

При моделировании почти все переменные имели ошибки, но особенно велика была ошибка комплексных переменных. Проведенное сравнение трех методов SR показано, что все они пригодны для построения математических моделей динамических систем по данным наблюдений, но именно GP-SINDy демонстрирует превосходящую эффективность. Полученные результаты позволяют считать GP-SINDy перспективным методом в данной области.

Литература

1. Семенов А. Д., Артамонов Д. В., Брюхачев А. В. Идентификация объектов управления: Учебное пособие. Пенза: Изд-во Пенз. гос. ун-та, 2003. 211 с.
2. Corriou JP. Models and Methods for Parametric Identification. In: Process Control. Springer, London. 2004. URL: researchgate.net/publication/319191179_Models_and_Methods_for_Parametric_Identification
3. Костюк В.И., Плотников А.А., Соколова А.Е. Алгоритмы идентификации нелинейных систем на основе данных. Информационные технологии и прикладная математика. – Арзамас: Арзамасский филиал ННГУ, 2023. – 104 с.
4. Kronberger G., Burlacu B., Kommenda M., Winkler S.M., & Affenzeller M. Symbolic Regression (1st ed.). Chapman and Hall/CRC. 2024. URL: taylorfrancis.com/books/mono/10.1201/9781315166407/symbolicregressionmicha

el-affenzeller-bogdan-burlacu-michael-kommenda-gabriel-kronberger-stephan-winkler

5. Makke N., Chawla S. Interpretable scientific discovery with symbolic regression: a review. *Artif Intell Rev* 57, 2, 2024. URL: link.springer.com/article/10.1007/s10462-023-10622-0

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

7. Чжан Л., Филимонов Н.Б. Разработка и исследование алгоритма генетического программирования для структурно-параметрической идентификации динамических систем // Высокопроизводительные вычислительные системы и технологии. 2024. Т. 8, № 1. С. 196-201.

8. Sun S, Ouyang R, Zhang B, et al. Data-driven discovery of formulas by symbolic regression. *MRS Bulletin*, 2019, 44(7): pp. 559-564.

9. Brunton S., Proctor J., & Kutz J. Discovering governing equations from data by sparse identification of nonlinear dynamical systems // *Proceedings of the National Academy of Sciences*. – 2016. – Vol. 113(15). – pp. 3932-3937.

10. Petersen B. et al. Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients. 2019. URL: semanticscholar.org/paper/Deepsymbolicregression%3AREcoveringmathematicalPetersenLandajuela/3ee45877f7f14c8ee4872a7249f74eef7dc2255f

11. Udrescu S. M., & Tegmark M. AI Feynman: A Physics-Inspired Method for Symbolic Regression. *Science Advances*, 6(16), eaay2631, 2020. URL: arxiv.org/pdf/1905.11481

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

13. Чжан Л. Гибридный метод структурно-параметрической идентификации динамических систем // Системы анализа и обработки данных. 2025. № 2 (98). С. 111–122.

14. Ljung L. System Identification. In: Procházka, A., Uhlíř, J., Rayner, P.W.J., Kingsbury, N.G. (eds) Signal Analysis and Prediction. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, MA. 1998. URL: books.google.co.in/books?id=qk2LBkKg5zcC&source=gbp_navlinks_s

References

1. Semenov A. D., Artamonov D. V., Bryukhachev A. V. Identifikatsiya ob"ektov upravleniya [Identification of control objects]: Uchebnoe posobie. Penza: Izd-vo Penz. gos. un-ta, 2003. P. 211.

2. Corriou JP. Models and Methods for Parametric Identification. In: Process Control. Springer, London. 2004. URL: researchgate.net/publication/319191179_Models_and_Methods_for_Parametric_Identification

3. Kostyuk V.I., Plotnikov A.A., Sokolova A.E. Algoritmy` identifikacii nelinejny`x sistem na osnove danny`x. Informacionny`e texnologii i prikladnaya matematika. [Algorithms for Identifying Nonlinear Systems Based on Data. Information Technology and Applied Mathematics]. Arzamas: Arzamasskij filial. NNGU, 2023. 104 p.

4. Kronberger G., Burlacu B., Kommenda M., Winkler S.M., & Affenzeller M. Symbolic Regression (1st ed.). Chapman and Hall/CRC. 2024. URL: taylorfrancis.com/books/mono/10.1201/9781315166407/symbolicregressionmichaelfenzeller-bogdan-burlacu-michael-kommenda-gabrielkronbergerstephanwinkler

5. Makke N., Chawla S. Interpretable scientific discovery with symbolic regression: a review. Artif Intell Rev 57, 2, 2024. URL: link.springer.com/article/10.1007/s10462-023-10622-0

6. Koza J.R. Genetic programming: On the programming of computers by means of natural selection. The MIT Press, 1992. 836 p.
7. Zhang L., Filimonov N.B. Vysokoproizvoditel'nye vychislitel'nye sistemy i tekhnologii. 2024. T. 8, № 1. pp. 196-201.
8. Sun S, Ouyang R, Zhang B, et al. Data-driven discovery of formulas by symbolic regression. MRS Bulletin, 2019, 44(7): pp. 559-564.
9. Brunton S., Proctor J., & Kutz J. Proceedings of the National Academy of Sciences. 2016. Vol. 113(15). pp. 3932-3937.
10. Petersen B. et al. Deep symbolic regression: recovering mathematical expressions from data via risk-seeking policy gradients. 2019. URL: semanticscholar.org/paper/Deepsymbolicregression%3AREcoveringmathematicalPetersenLandajuela/3ee45877f7f14c8ee4872a7249f74eef7dc2255f
11. Udrescu S. M., & Tegmark M. AI Feynman: A Physics-Inspired Method for Symbolic Regression. Science Advances, 6(16), eaay2631, 2020. URL: arxiv.org/pdf/1905.11481
12. Zhang L. Sistemy analiza i obrabotki dannykh. 2024. №2, pp. 85-93.
13. Zhang L. Sistemy analiza i obrabotki dannykh. 2025. №2, pp. 111–122.
14. Ljung L. System Identification. In: Procházka, A., Uhlíř, J., Rayner, P.W.J., Kingsbury, N.G. (eds) Signal Analysis and Prediction. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, MA. 1998. URL: books.google.co.in/books?id=qk2LBkKg5zcC&source=gbp_navlinks_s

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 17.09.2025

Дата публикации: 29.10.2025