УДК 681.5.015

ГІБРИДНІ АЛГОРИТМИ САМООРГАНІЗАЦІЇ МОДЕЛЕЙ ДЛЯ ПРОГНОЗУВАННЯ СКЛАДНИХ ПРОЦЕСІВ

В.С. Степашко, О.С. Булгакова, В.В. Зосімов

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України , 03680 Київ, просп. Академіка Глушкова, 40
stepashko@irtc.org.ua, sashabulgakova1@gmail.com, zosimovvv@bk.ru

Запропоновано новий різновид алгоритму МГУА гібридного типу, що поєднує в собі як багаторядні, так і комбінаторні схеми самоорганізації моделей і містить такі основні нововведення: використання початкових аргументів на кожному ряді, щоб запобігти втраті істотних аргументів; оптимізація структури кожної частинної моделі за комбінаторним алгоритмом для уникнення переускладнення.

Ключові слова: індуктивне моделювання, МГУА, багаторядний алгоритм, комбінаторний алгоритм, гібридний алгоритм.

A new kind of GMDH algorithm of hybrid type is proposed combining both multilayered and combinatorial schemes for model self-organization and includes the following main novelties: using initial arguments in each layer to avoid loosing the relevant ones; optimization of every partial model structure by combinatorial algorithm to avoid overfitting.

Keywords: inductive modeling, GMDH, multilayered algorithm, combinatorial algorithm, hybrid algorithm.

Предложена новая разновидность алгоритма МГУА гибридного типа, сочетающая в себе как многорядные, так и комбинаторные схемы самоорганизации моделей и включающая следующие основные нововведения: использование начальных аргументов на каждом ряде, чтобы предотвратить потерю существенных аргументов; оптимизация каждой частной модели по комбинаторному алгоритму, чтобы избежать переусложнения.

Ключевые слова: индуктивное моделирование, МГУА, многорядный алгоритм, комбинаторный алгоритм, гибридный алгоритм.

Вступ

Метод групового урахування аргументів (МГУА) застосовується в найрізноманітніших галузях для аналізу даних та пошуку знань, моделювання систем та прогнозування процесів. Основними алгоритмами МГУА є комбінаторний і багаторядний, які існують в різних модифікаціях [1]. В комбінаторних алгоритмах організується повний перебір усіх можливих структур моделей, тому вони обмежені в своєму застосуванні часом обчислень та розмірністю задач. Багаторядні алгоритми, на відміну від комбінаторних, дозволяють розв’язувати задачі великої розмірності за прийнятний час завдяки впорядкованому перебору моделей.

Проте багаторядний алгоритм має свої недоліки, зокрема, можливість втрати інформативних аргументів. Класичний алгоритм побудований так, що втрата аргументу на якійсь із початкових стадій процесу моделювання є остаточною, і якщо цей аргумент інформативний, кінцева модель буде

неефективною. Крім того, при застосуванні нелінійного (квадратичного) частинного опису відбувається стрімке зростання ступеня нелінійності моделі.

Для усунення вказаних недоліків пропонується кілька істотних модифікацій багаторядної структури алгоритму МГУА за рахунок поєднання (гібридизації) кращих особливостей і переваг як багаторядних, так і комбінаторних схем перебору моделей.

1. Постановка задачі ідентифікації моделей.

Нехай маємо вибірку $W=[Xy]$, що містить n точок спостережень, які утворюють матрицю вимірювань m незалежних вхідних змінних (аргументів) $X=\{x_{ij},\ i=1,...,n;\ j=1,...,m\}$ і вектор однієї вихідної (залежної) величини $y = (y_1, ..., y_m)^T$, причому $n > m$.

Загалом задача ідентифікації полягає у формуванні за даними вибірки деякої множини F моделей різної структури вигляду

$$\hat{y}_f = f(X, \hat{\theta}_f) \tag{1}$$

і відшуканні оптимальної моделі за умовою:

$$f^* = \arg\min_{f \in F} CR(y, f(X, \hat{\theta}_f)), \tag{2}$$

причому оцінка параметрів в (1) для кожної моделі $f \in F$ є розв'язком ще однієї екстремальної задачі

$$\hat{\theta}_f = \arg\min_{f \in R^{s_f}} QR(y, X, \theta_f), \tag{3}$$

де s_f - складність моделі f, що дорівнює числу ненульових компонентів у моделі (2), а QR – критерій якості розв'язку задачі параметричної ідентифікації кожної частинної моделі, що генерується в задачі структурної ідентифікації.

2. Модифікації багаторядного алгоритму МГУА

У класичному багаторядному алгоритмі кожна частинна модель формується як комбінація $f(x_i, x_j)$ пари аргументів x_i, x_j $(i, j = 1,2,...,m, i < j)$ на початковому ряді та комбінація $f_r(y_i^{r-1}, y_j^{r-1}$ пари кращих моделей, отриманих на попередньому ряді, починаючи з другого. В алгоритмі можуть бути використані лінійні або нелінійні (повний квадрат) частинні описи:

$$y_l^r = a_0 + a_1 y_i^{r-1} + a_2 y_j^{r-1} \tag{4}$$

$$y_l^r = a_0 + a_1 y_i^{r-1} + a_2 y_j^{r-1} + a_3 \left(y_i^{r-1}\right)^2 + a_4 y_i^{r-1} y_j^{r-1} + a_5 \left(y_j^{r-1}\right)^2 \tag{4a}$$

Коефіцієнти $a_0, a_1, \ldots, a_5$ оцінюються за МНК на навчальній частині вибірки. Кращі F моделей за критерієм селекції беруть участь у формуванні моделей наступного ряду, рис.1. Процес закінчується на ряді, після якого

мінімальне значення критерію почне зростати. При цьому існує можливість втрати істотних аргументів у процесі селекції.

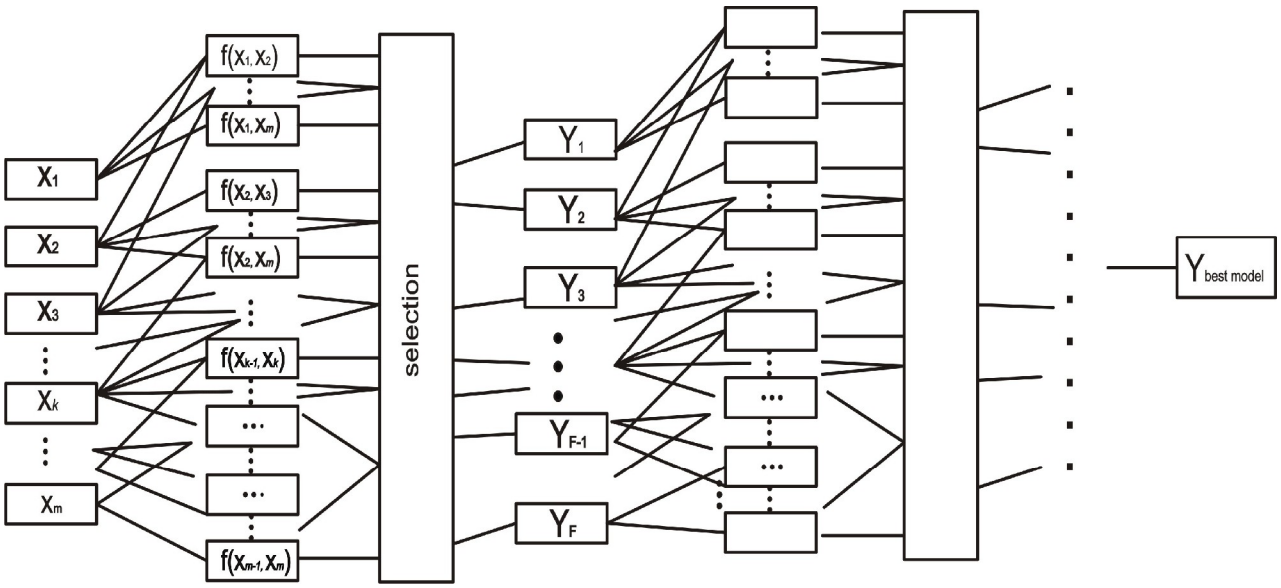

Рис.1. Класичний багаторядний алгоритм МГУА

Ідею пропонованих модифікацій класичного багаторядного алгоритму видно з порівняння Рис. 1 та Рис. 2 – 4: починаючи з другого кроку, при генеруванні моделей наступного ряду можуть використовуватися знову початкові аргументи, які були втрачені на попередньому ряді, при цьому формуються моделі виду

$$y_l^r = a_0 + a_1 y_i^{r-1} + a_2 x_j , \qquad (5)$$

причому функція $f(x_i, x_j)$ першого ряду має вигляд:

$$f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_5 x_1 x_2 + a_4 x_2^2 , \qquad (6)$$

і тим самим отримуємо можливість ускладнювати структури моделей. Для вибору кращих моделей структури (6) можна використовувати комбінаторну оптимізацію складності частинних моделей, яка може бути застосована на будь-якому ряді.

Таким чином, у цьому типі алгоритмів виключається можливість втрати істотних аргументів на початкових рядах та додається можливість комбінаторної оптимізації складності частинних моделей. Очевидно, що при цьому забезпечуються такі можливості: побудова лінійних моделей за використання нелінійного частинного опису (нелінійні члени можуть відкидатись), а також поступове підвищення степеня нелінійності.

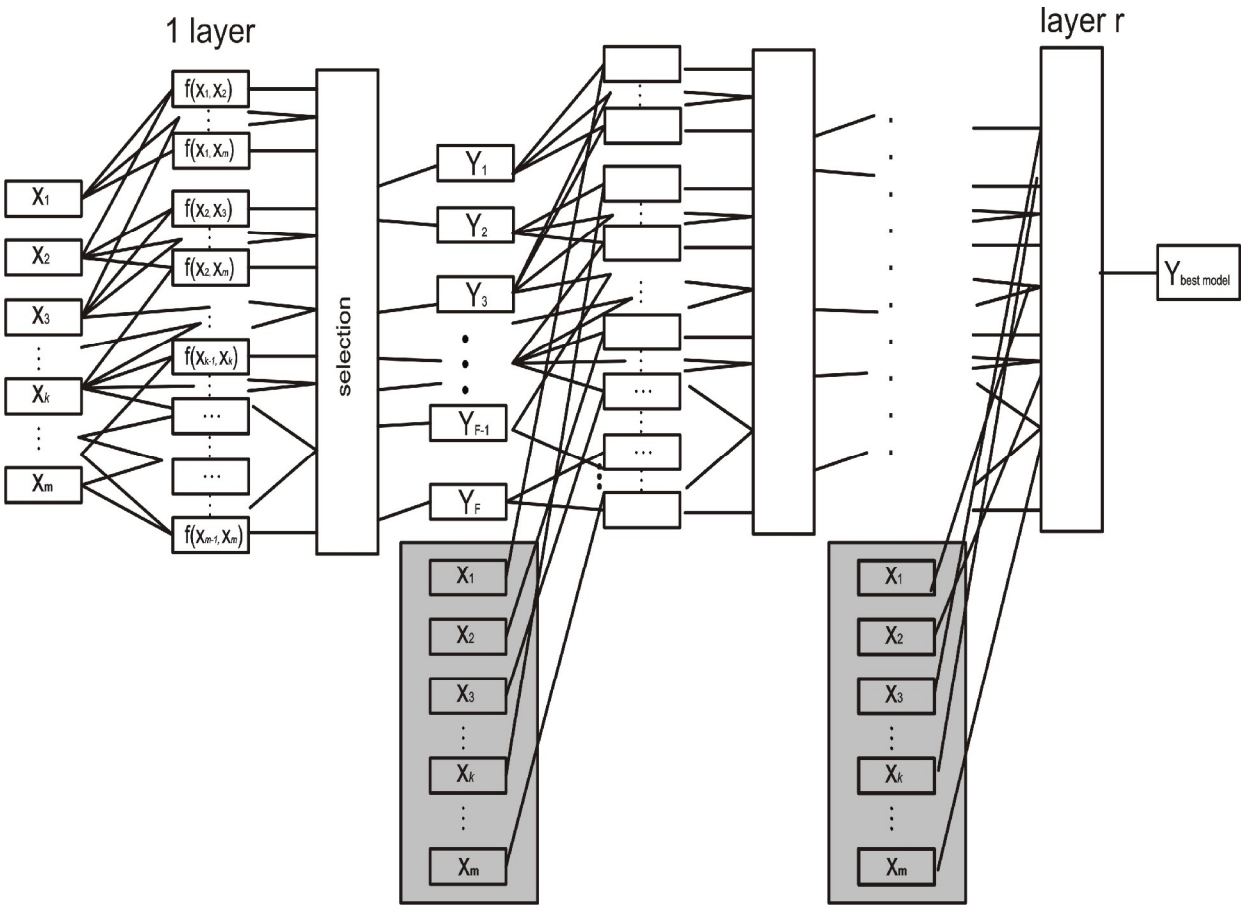

Рис. 2. Модифікований МГУА з додаванням початкових аргументів, втрачених на попередніх рядах

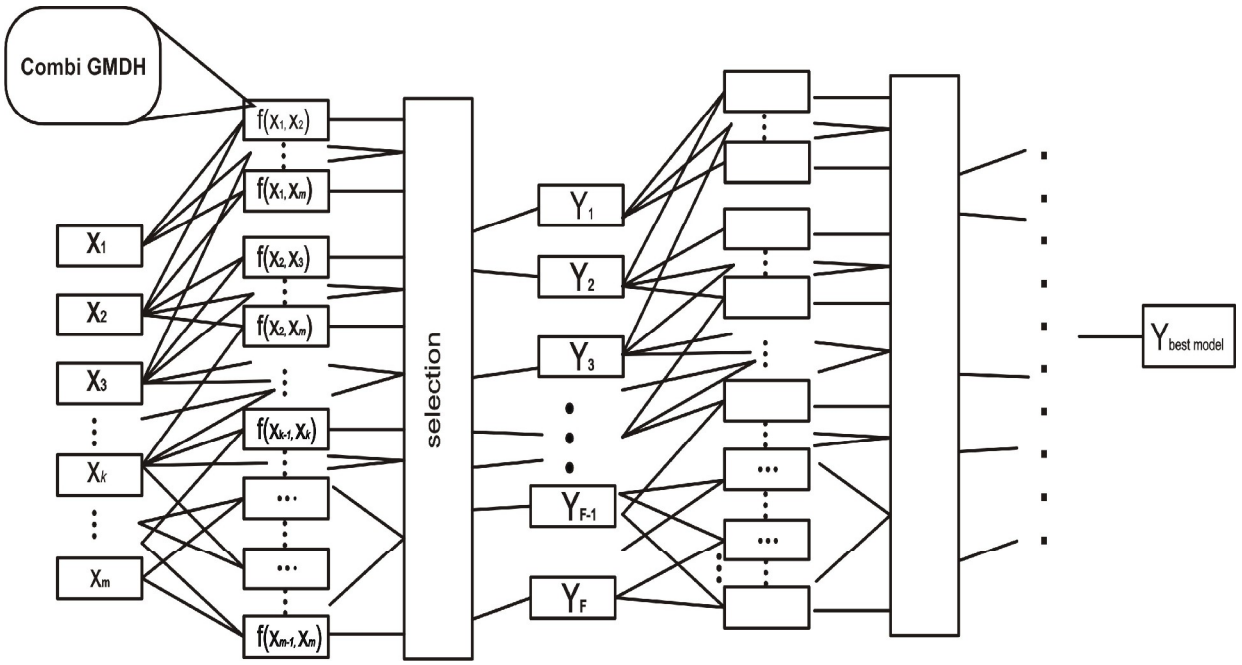

Рис. 3. Гібридний алгоритм МГУА з комбінаторною оптимізацією частинних моделей, де $f(x_i, x_j)$ визначається за (6)

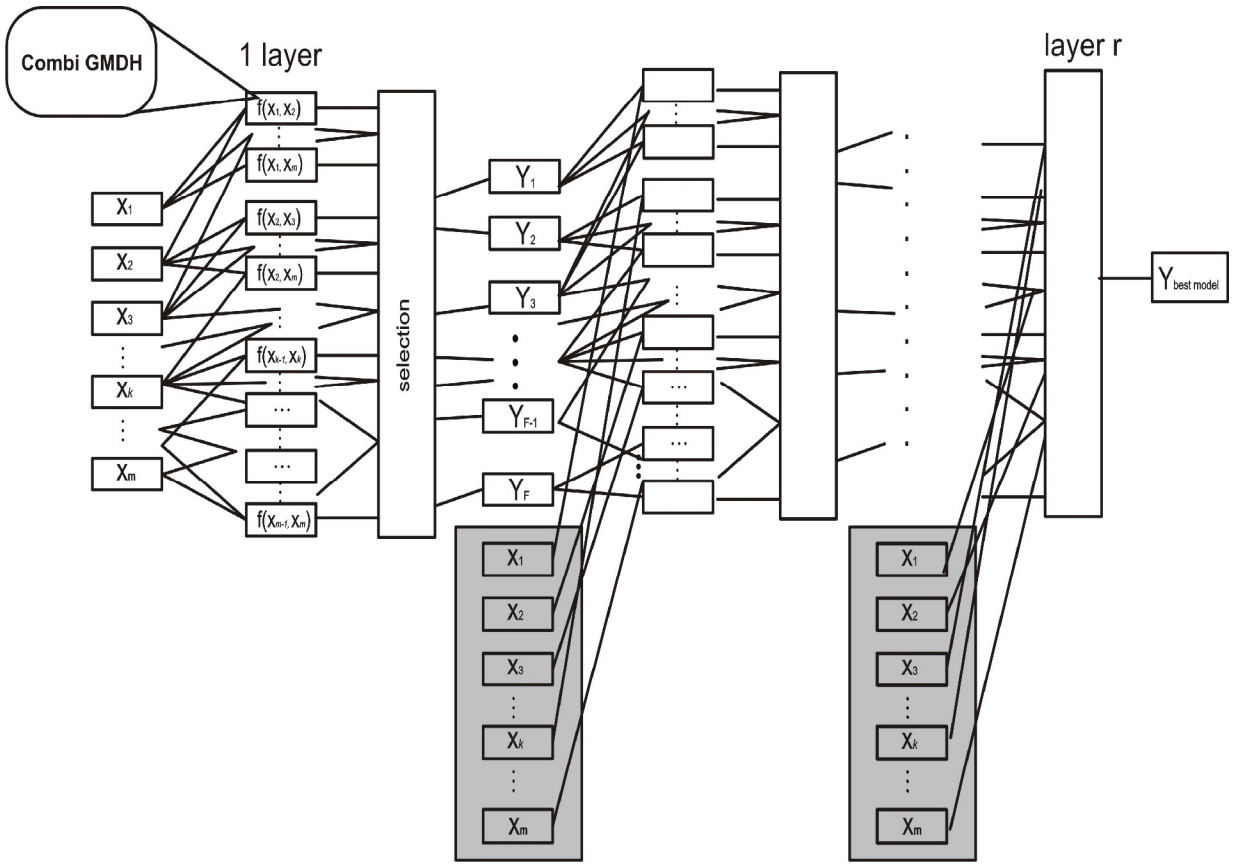

Рис. 4. Гібридний МГУА з комбінаторною оптимізацією частинних моделей та додаванням початкових аргументів, втрачених на попередніх рядах

Представлені модифікації класичного багаторядного алгоритму МГУА дозволяють отримати такі основні варіанти багаторядних алгоритмів:

1. Класичний алгоритм МГУА з частинним описом виду (4);
2. Модифікований з описом виду (5);
3. Комбінований, коли рівноправно використовуються проміжні та початкові аргументи;
4. у кожному варіанті можна використовувати чи не використовувати оптимізацію складності кожної частинної моделі із застосуванням повного перебору варіантів (комбінаторного алгоритму).

Запропоновані модифікації багаторядного алгоритму з включенням на кожному кроці початкових аргументів та з комбінаторною оптимізацією дозволяють удосконалити генератор структур класичного багаторядного алгоритму МГУА та отримати нові результати, в яких відсутня можливість втрачати інформативні аргументи, що можуть бути відсіяні на попередніх етапах моделювання квадратичних частинних моделей. Застосування комбінаторного перебору варіантів дає можливість оптимізувати складність кожної частинної моделі, що генерується в процесі роботи алгоритму. Метою цього є запобігти переускладненню моделі.

3. Результати досліджень

Дослідження були проведені на штучно згенерованих даних. Якість моделі була обчислена вибірці *B* як значення критерію регулярності *AR*:

$$AR = \left\| y_B - X_B \hat{\theta}_A \right\|^2 \tag{7}$$

Експеримент 1. Пошук моделі за даними без шуму

Вибірка згенерована випадковим чином і містить 200 змінних і 200 точок (рядків таблиці даних). Вибірка ділиться на дві частини: перша (135 рядків) – робоча вибірка, яка використовується для оцінки коефіцієнтів і розрахунку критеріїв, друга (65 рядків – кожний третій рядок) – тестова вибірка.

Вихідна величина, що не містить шуму, представлена такою істинною лінійною залежністю від перших 10 аргументів з усіх 200:

$$y = -3x_1 - 3x_2 + 5x_3 - x_4 - x_5 + 3x_6 + x_7 - 2x_8 + x_9 + x_{10}.$$

Істинна квадратична модель від 4 аргументів з усіх 200 має вигляд:

$$y = -3x_1 - 2x_2 x_3 + x_{25}^2 + x_{25} x_1 + 12.$$

Розглянемо чотири варіанти пошуку оптимальної моделі з використанням таких алгоритмів:

1) класичний багаторядний алгоритм;
2) модифікований МГУА з додаванням початкових аргументів, утрачених на попередніх рядах;
3) модифікований МГУА з комбінаторною оптимізацією частинних;
4) модифікований МГУА з комбінаторною оптимізацією частинних моделей та з додаванням втрачених початкових аргументів;
5) модифікований комбінаторний алгоритм [4].

В результаті застосування вказаних п’яти варіантів розробленого гібридного алгоритму отримали порівняльні результати точності побудованих моделей, подані в таблицях 1 та 2 (де n/a означає відсутність даних з огляду на неможливість виконати перебір варіантів).

Таблиця 1

Порівняння значень критерію регулярності для різних модифікацій багаторядного алгоритму для лінійної моделі

№ варіанта пошуку моделі	Лінійна задача	
	AR	Модель
1	3,045	$y = -3{,}001x_1 + 5{,}000x_3 - 1{,}002x_4 - 1{,}001x_5 + 3{,}000x_6 + 0{,}999x_7 - 1{,}998x_8 + 1{,}001x_9 + 1{,}000x_{10} - 3{,}035x_{27}$

2	$8*10^{-5}$	$y = -3{,}000x_1 - 3{,}00025x_2 + 5x_3 - 1{,}0001x_4 - 1{,}010x_5 + + 3{,}000003x_6 + x_7 - 2x_8 + x_9 + 0{,}999x_{10}$
3	2,015	$y = -3{,}000x_1 + 5{,}000x_3 - 1{,}002x_4 - 1{,}001x_5 + 3{,}000x_6 + + 0{,}999x_7 - 1{,}998x_8 + 1{,}001x_9 + 1{,}000x_{10} - 2{,}996x_{73}^2$
4	$2*10^{-4}$	$y = -3{,}000x_1 - 3{,}00025x_2 + 5x_3 - 1{,}0001x_4 - 1{,}010x_5 + + 3{,}000003x_6 + x_7 - 2x_8 + x_9 + 0{,}999x_{10}$
5	0,380	$y = -2{,}998x_1 - 2{,}990x_2 + 5{,}011x_3 - 0{,}992x_4 - 0{,}991x_5 + + 3{,}008x_6 + 0{,}990x_7 - 2{,}019x_8 + 1{,}005x_9 + 0{,}982x_{10}$

Таблиця 2

Порівняння значень критерію регулярності для різних модифікацій багаторядного алгоритму для квадратичної моделі

№ варіанта пошуку моделі	Нелінійна задача	
	AR	Модель
1	11,309	$y = -3x_1 - 1{,}231x_2 + 2{,}711x_{25} + 0{,}998x_{32} + x_{37} - - 2{,}001x_{45} + 12$
2	3.011	$y = -3x_1 - 1{,}231x_2 + 6{,}711x_{25} - 2{,}001x_{45} + 12$
3	0,462	$y = -2{,}0001x_2x_3 + x_{25}^2 + x_{25}x_1 - 2{,}999x_{45} + 12$
4	$3*10^{-8}$	$y = -3{,}000x_1 - 2{,}000x_2x_3 + x_{25}^2 + x_{25}x_1 + 12{,}000$
5	n/a	n/a

Порівнюючи отриманні значення, можна зробити такі висновки:

- Для лінійної моделі найкраще рішення отримали за допомогою використання модифікованого МГУА з додаванням початкових аргументів, втрачених на попередніх рядах. Тим самим в класичному алгоритмі інформативні аргументи були відсіяні на початкових рядах, що значно погіршило отриману модель.
- Для квадратичної моделі найкраще рішення отримали за допомогою модифікованого МГУА з комбінаторною оптимізацією частинних моделей та з додаванням початкових аргументів, втрачених на попередніх рядах.

Експеримент 2. *Пошук моделі при додаванні шуму до вихідної змінної.*

Вибірка згенерована випадковим чином і містить 40 змінних і 60 точок (рядків). Вибірка ділиться на дві частини: перші 40 рядків – робоча вибірка, яка

використовується для оцінки коефіцієнтів і розрахунку критеріїв, друга (20 рядків – кожен третій рядок) – тестова вибірка.

Вихідна величина, що не містить шуму, представлена такими варіантами залежності від змінних:

лінійна залежність: $\overset{\circ}{y} = 0{,}5 - 1{,}2x_2 + 5x_{10} - 3{,}4x_{25}$;

квадратична залежність: $\overset{\circ}{y} = 7 - x_1 + 2x_{31} - 1{,}2x_1x_{31} + 2{,}7x_{12}^2$.

Шум генерується за такою формулою:

$$y = \overset{\circ}{y} + \alpha(2\gamma - 1)\frac{\overset{\circ}{y}_{\max} - \overset{\circ}{y}_{\min}}{200}, \qquad (8)$$

де $\overset{\circ}{y}$ - точний сигнал,

α - рівень шуму у відсотках,

γ - випадкова величина, рівномірно розподілена на інтервалі $[0;1]$,

$\overset{\circ}{y}_{\max}$, $\overset{\circ}{y}_{\min}$ - відповідно максимальне та мінімальне значення істинної вихідної величини.

Згенеровані таким чином вибірки шуму відповідної довжини додавались до точних значень вихідної змінної.

В цьому експерименті було застосовано описані далі варіанти пошуку оптимальної моделі в умовах шуму різної інтенсивності.

Варіант 1. Пошук лінійної моделі при доданні до вихідної змінної шуму за допомогою різних варіантів пошуку.

В таблицях 3-5 наведено значення критерію *AR* та обрані моделі без додавання до вихідної змінної шуму та з шумом 10% і 30%.

Таблиця 3

Результати побудови лінійної моделі (без шуму)

№ варіанта пошуку моделі	*AR*	Модель
1	62,768	$y = 0{,}457 - 0{,}027x_4 - 0{,}015x_5 + 4{,}968x_{10} - 0{,}013x_{13} + 0{,}002x_{14} - 3{,}474x_{25} - 0{,}749x_{26} - 0{,}158x_{39}$
2	$4*10^{-5}$	$y = 0{,}499 - 1{,}200x_2 + 5{,}000x_{10} - 3{,}399x_{25}$
3	34,256	$y = 0{,}89 + 4{,}044x_{10} - 3{,}037x_{25} - 2{,}930x_{35} - 0{,}799x_2^2$
4	$5*10^{-5}$	$y = 0{,}497 - 1{,}201x_2 + 5{,}000x_{10} - 3{,}399x_{25}$

Таблиця 4

Результати побудови лінійної моделі при доданні шуму 10%

№ вар. пошуку моделі	*AR*	Модель
1	77,98	$y = 0{,}595 - 0{,}024x_4 - 0{,}007x_5 + 5{,}508x_{10} - 0{,}016x_{13} + 0{,}001x_{14} - 3{,}857x_{25} + 0{,}002x_{27} - 0{,}028x_{37} - 0{,}17x_{39}$
2	$4{,}5*10^{-5}$	$y = 0{,}540 - 1{,}320x_2 + 5{,}500x_{10} - 3{,}540x_{25}$
3	38,27	$y = 0{,}92 + 5{,}023x_{10} - 3{,}137x_{25} - 3{,}030x_{35} - 0{,}990x_2^2$
4	$5{,}4*10^{-5}$	$y = 0{,}542 - 1{,}306x_2 + 5{,}601x_{10} - 3{,}399x_{25}$

Таблиця 5

Результати побудови лінійної моделі при доданні шуму 30%

№ вар. пошуку моделі	*AR*	Модель
1	114,07	$y = 0{,}720 - 0{,}029x_4 - 0{,}009x_5 + 6{,}660x_{10} - 0{,}0206x_{13} + 0{,}006x_{14} - 4{,}663x_{25} + 0{,}003x_{27} - 0{,}003x_{37} - 0{,}216x_{39}$
2	$2{,}21*10^{-4}$	$y = 0{,}664 - 1{,}596x_2 + 6{,}649x_{10} - 4{,}522x_{25}$
3	47,12	$y = 0{,}92 + 6{,}033x_{10} - 2{,}132x_{25} - 2{,}002x_{35} - 1{,}010x_2^2$
4	$3{,}21*10^{-4}$	$y = 0{,}673 - 1{,}601x_2 + 7{,}001x_{10} - 3{,}509x_{25}$

Варіант 2. Пошук квадратичної моделі при доданні до вихідної змінної шуму за допомогою різних варіантів пошуку.

В таблицях 6-8 наведено значення критерію *AR* та обрані моделі без додавання до вихідної змінної шуму та з шумом 10% і 30%.

Таблиця 6

Результати побудови квадратичної моделі (без шуму)

№ вар. пошуку моделі	*AR*	Модель
1	$11{,}201*10^3$	$y = -44{,}725 - 5{,}589x_1 + 0{,}536x_5 + 0{,}380x_{10} + 30{,}378x_{12} - 0{,}052x_{13} - 0{,}405x_{18} + 0{,}006x_{25} - 2{,}429x_{31} + 0{,}936x_{37} + 0{,}120x_{38}$
2	$5{,}969*10^3$	$y = -32{,}277 - 8{,}074x_1 - 0{,}179x_4 + 1{,}266x_5 - 1{,}441x_6 + 0{,}842x_7 - 0{,}884x_8 + 30{,}475x_{12} - 0{,}717x_{21} - 4{,}934x_{31} - 2{,}878x_{33}$
3	$1{,}383*10^3$	$y = 6{,}89 - 0{,}044x_1 + 2{,}047x_{31} - 2{,}030x_1x_{31} + 2{,}799x_{12}^2$
4	$0{,}684*10^3$	$y = 7{,}001 - 0{,}004x_1 + 2{,}000x_{31} - 1{,}130x_1x_{31} + 2{,}700x_{12}^2$

Таблиця 7

Результати побудови квадратичної моделі, шум 10%

№ вар. пошуку моделі	*AR*	Модель
1	$12{,}861*10^3$	$y = -46{,}302 - 5{,}997x_1 + 0{,}917x_5 + 0{,}400x_{10} + 33{,}379x_{12} - 0{,}281x_{13} - 0{,}593x_{18} + 0{,}007x_{25} - 2{,}810x_{31} - 0{,}210x_{37} + 0{,}094x_{38}$
2	$7{,}240*10^3$	$y = -35{,}977 - 8{,}838x_1 - 0{,}120x_4 + 1{,}470x_5 - 1{,}468x_6 + 0{,}987x_7 - 0{,}958x_8 + 33{,}491x_{12} - 0{,}824x_{21} - 5{,}473x_{31} + 3{,}110x_{33}$
3	$1{,}672*10^3$	$y = 7{,}332 - 0{,}114x_1 + 2{,}107x_{31} - 2{,}018x_1x_{31} + 3{,}009x_{12}^2$
4	$0{,}720*10^3$	$y = 7{,}011 - 0{,}007x_1 + 2{,}031x_{31} - 1{,}176x_1x_{31} + 2{,}712x_{12}^2$

Таблиця 8

Результати побудови квадратичної моделі, шум 30%

№ вар. пошуку моделі	*AR*	Модель
1	$18{,}593*10^3$	$y = -57{,}354 - 7{,}080x_1 + 0{,}688x_5 + 0{,}534x_{10} + 38{,}977x_{12} - 0{,}0557x_{13} - 0{,}489x_{18} + 0{,}019x_{25} - 3{,}168x_{31} + 7{,}474x_{31} + 0{,}158x_{37} + 0{,}172x_{38}$
2	$10{,}170*10^3$	$y = -39{,}887 - 10{,}507x_1 - 0{,}414x_4 + 1{,}483x_5 - 1{,}827x_6 + 1{,}111x_7 - 1{,}109x_8 + 39{,}558x_{12} - 0{,}906x_{21} - 6{,}389x_{31} + 3{,}730x_{33}$
3	$2{,}152*10^3$	$y = 7{,}432 - 0{,}214x_1 + 2{,}211x_{31} - 2{,}108x_1x_{31} + 3{,}100x_{12}^2$
4	$1{,}220*10^3$	$y = 7{,}031 - 0{,}015x_1 + 2{,}041x_{31} - 1{,}101x_1x_{31} + 2{,}900x_{12}^2$

Як бачимо з наведених результатів експериментів, структура оптимальних моделей, побудованих кожним з алгоритмів, при додаванні шуму до вихідної змінної не змінюється.

Приклади показали, що для лінійної моделі найдоцільнішим буде використання модифікованого МГУА з додаванням початкових аргументів, втрачених на попередніх рядах. Такий самий результат дав модифікований МГУА з комбінаторною оптимізацією частинних моделей та з додаванням початкових аргументів, втрачених на попередніх рядах, але процес знаходження оптимальної моделі зайняв більш часу.

Для знаходження квадратичної моделі гібридний алгоритм МГУА з комбінаторною оптимізацією частинних моделей та з додаванням початкових аргументів, втрачених на попередніх рядах, виявився оптимальним. В обох випадках класичний багаторядний алгоритм виявився найменш ефективним.

Висновки

В роботі подано доцільні модифікації багаторядного алгоритму МГУА. Ці зміни дозволяють отримати кілька нових варіантів багаторядних алгоритмів, які представлені в статті.

Запропоновані варіанти гібридного багаторядного алгоритму з включенням на кожному кроці початкових аргументів та з комбінаторною оптимізацією дозволяють удосконалити генератор структур класичного багаторядного алгоритму МГУА та отримати нові результати, в яких відсутня можливість втрати інформативних аргументів, що можуть бути відсіяні в процесі переходу від ряду до ряду алгоритму як з лінійними, так і з квадратичними частинними моделями. Застосування комбінаторного перебору варіантів дає можливість оптимізувати кожну частинну модель, що генерується в процесі роботи алгоритму. Це дозволяє уникнути переускладнення моделі.

Завдяки цим особливостям розробленого гібридного алгоритму при його застосуванні з квадратичними частинними моделями може бути побудовані лінійні, білінійні та нелінійні моделей складних систем і процесів.

Ефективність запропонованих модифікацій багаторядного алгоритму МГУА продемонстровано на кількох тестових задачах побудови лінійних та нелінійних моделей.

Література

1. Ivakhnenko A.G.: Group method of data handling - competitor for the method of stochastic approximation, Soviet Automatic Control, No. 3, pp. 58-72, 1968.
2. Madala H.R, Ivakhnenko A.G.: Inductive learning algorithms for complex systems modeling. Boca Raton, London, Tokyo: CRC Press Inc., 1994.
3. Stepashko V.S.: Combinatorial GMDH algorithm with the optimal scheme of models sorting-out, Soviet Automatic Control, No. 3, pp. 31-36, 1981.
4. Samoilenko O.A., Stepashko V.S.: Combinatorial GMDH algorithm with successive selection of arguments. Proceedings of IWIM 2007 in Prague, pp. 139-140, 2007.
5. Bulgakova O., Kordik P. Methods of true data mining model selection - with experimental results. Proceedings of IWIM 2009 in Krynica, Poland, pp. 23-27, 2009.