УДК 681.513

АНАЛІЗ ЕФЕКТИВНОСТІ ЗАСТОСУВАННЯ ЧАСТОТНОГО КРИТЕРІЮ В АЛГОРИТМІ ПОСЛІДОВНОГО ВІДСІЮВАННЯ НЕІНФОРМАТИВНИХ АРГУМЕНТІВ

О.А. Самойленко В.С. Степашко

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України, Київ, пр-т Академіка Глушкова, 40

soa0pga@gmail.com

В статті аналізуються властивості частотного критерію в задачі оцінювання рівня інформативності аргументів. На основі виконаних експериментів обґрунтовано доцільність та ефективність застосування цього критерію в алгоритмі послідовного відсіювання неінформативних аргументів.
Ключові слова: індуктивне моделювання, МГУА, критерій інформативності аргументів.

Properties of a frequency criterion in the task of arguments informativity level estimation are analyzed in the article. Expediency and efficiency of this criterion use in the algorithm of successive sifting of uninformative arguments is justified based on the made experiments.
Key words: inductive modeling, GMDH, criterion of arguments informativity.

В статье анализируются свойства частотного критерия в задаче оценивания уровня информативности аргументов. На основании выполненных экспериментов обоснована целесообразность и эффективность применения этого критерия в алгоритме последовательного отсеивания неинформативных аргументов.
Ключевые слова: индуктивное моделирование, МГУА, критерий информативности аргументов.

Вступ

Алгоритм повного перебору моделей COMBI [1, 2], розроблений в рамках методології МГУА, ефективно розв’язує задачі моделювання за вибірками даних з відносно невеликою кількістю аргументів – за наявності понад 30 аргументів повний перебір стає неможливим.

В [3-6] для вирішення цієї проблеми запропоновано використовувати різні процедури послідовного відсіювання найменш інформативних аргументів, причому рівень інформативності оцінюється за допомогою так званого частотного критерію. Метою цієї роботи є обґрунтування доцільності та ефективності використання цього критерію в задачах великої розмірності.

1. Постановка задачі

Будемо вважати, що задано вибірку даних $W = (X, y)$, $\dim X = n \times m$, а залежність вихідної змінної y від входів X є лінійною:

$$y = X\theta + \xi = \overset{o}{y} + \xi, \quad \overset{o}{y} \overset{\Delta}{=} X\theta, \tag{1}$$

де θ – вектор, що складається з m невідомих істинних параметрів, з яких тільки $s_0 \le m$ ненульові, ξ – випадковий вектор шуму. Існує точна (істинна) модель об'єкта виду $\overset{o}{y} = \overset{o}{X}\theta_0$, де $\overset{o}{X} \in R^{n \times s_0}$, причому всі вектор-стовпці $\overset{o}{x}_j$, $j = 1,...,s_0$, містяться в матриці X, а θ_0 – невідомий істинний вектор параметрів, який складається тільки з тих компонент вектора θ, які не дорівнюють нулю і визначають інформативні змінні (від яких залежить істинний вихід $\overset{o}{y}$). Набір з s векторів $X_s \in X$, що входять в певну модель, будемо називати структурою цієї моделі, а кількість s цих векторів – складністю моделі. В такому випадку задача структурної ідентифікації має полягати у визначенні найкращого наближення до невідомих значень s_0, θ_0 при одночасному розділенні вхідних змінних матриці X на інформативні та неінформативні.

Загалом задача ідентифікації полягає в формуванні за даними вибірки W деякої множини $\Im$ моделей різної структури виду $y_f = f(X, \hat{\theta}_f)$ та відшукання оптимальної моделі за мінімумом заданого критерію $CR(\cdot)$:

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

де оцінки параметрів $\hat{\theta}_f$ для кожної $f \in \Im$, що вказують на набір ненульових компонент (структуру), визначаються за умовою

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

де $Q(\cdot) \ne CR(\cdot)$ – критерій якості розв'язку задачі параметричної ідентифікації кожної часткової моделі, згенерованої в задачі структурної ідентифікації (2).

В задачах побудови моделей використовуються різні критерії рівня інформативності аргументів для визначення тих векторів, що можуть входити до структури істинної моделі. В алгоритмах послідовного відсіювання [4, 5, 6] використовується частотний критерій, запропонований в [3]. Відповідно до цього підходу поетапно серед певної кількості моделей, побудованих алгоритмом МГУА, відбирається F найкращих моделей за заданим критерієм, які складають множину $\Im_F$. Рівень інформативності кожного аргумента визначається частотою його присутності в цих кращих моделях і визначається за формулою:

$$FC_j = \frac{q_j}{F} \tag{4}$$

Тут q_j – кількість моделей з множини $\Im_F$, які містять j -й аргумент.

Розглянемо детальніше цей критерій у поєднанні з процедурами послідовного відсіювання неінформативних аргументів та застосуємо чисельні

експерименти на тестових задачах для обґрунтування доцільності та ефективності його застосування.

2. Дослідження ефективності частотного критерію

Для проведення чисельних експериментів і досліджень, пов'язаних з частотним критерієм, розглянемо два штучно згенерованих тестових приклади побудови оптимальних моделей.

Приклад 1. Нехай $m = 20$, $n = 250$, $s_0 = 10$. Сформуємо вибірку X, $\dim X = 250 \times 20$, що складається з векторів $x_j, j = 1,...,20$. Кожному x_j відповідає певна функція $f_j(t_k)$ (табл. 1). Вектори $t_k, k = 1,...,5$ генеруються за допомогою генератора випадкових рівномірно розподілених чисел.

Таблиця 1

Залежність змінних x від t_k

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
$f(t)$:	t_1	t_2	t_3	t_1^2	t_1t_2	t_1t_3	t_1t_4	t_2^2	t_2t_3	t_3^2	t_4	t_5	t_1t_5	t_2t_4	t_2t_5	t_3t_4	t_3t_5	t_4^2	t_4t_5	t_5^2

Задамо вектор $\theta_0 = [-3; -3; 5; -1; -1; 3; 1; -2; 1; 1]^T$ параметрів істинної моделі, а її структурний вектор у вигляді такої послідовності нулів і одиниць: $d = [1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0]$ – тобто перші 10 аргументів є істинними, від яких залежить вихідна змінна, а інші – надлишкові. Використовуючи формулу (1) і шум ξ (рівень шуму в даному прикладі складає 2% від істинного виходу $\overset{o}{y}$), обчислимо вихідну величину:

$$y = \overset{o}{X}\theta_0 + \xi = -3x_1 - 3x_2 + 5x_3 - x_4 - x_5 + 3x_6 + x_7 - 2x_8 + x_9 + x_{10} + \xi. \qquad (5)$$

Тут $\dim y = \dim \xi = n \times 1$, $\dim \overset{o}{X} = n \times s_0$, $\dim \theta_0 = s_0 \times 1$, $s_0 = 10$, а x_j, $j = 1,...,s_0$ – вектор-стовпці матриці $\overset{o}{X}$.

Приклад 2. Виконаємо експерименти на вибірці, розглянутій у прикладі 1, але для ускладнення задачі пошуку істинної моделі доповненій до 200 аргументів зайвими векторами, які не входять в істинну модель. Таким чином, обидва масиви даних відповідно з 20 та 200 аргументами відповідають одній і тій же істинній моделі.

Тепер задача полягає у відновленні вектора θ_0 за допомогою пошуку оптимальної моделі $\overset{o}{y} = \overset{o}{X}\theta_0$ на всій вибірці $W = (X, y)$ за деяким критерієм $CR(\cdot)$, яким у нашому випадку буде критерій регулярності AR [2, 7].

Для знаходження кращих моделей в [5, 6] в поєднанні з частотним критерієм застосовувались наступні два багатоетапні підходи: FSS (Forward Successive Selection) і BSS (Backward Successive Selection).

Кожен етап методу FSS містить такі кроки:

Крок 1: На вибірці W з m аргументами будуються всі можливі моделі складності від 1 до максимально прийнятного (за часовими затратами) $s_{\max}$. Кількість таких моделей дорівнює $P_{s_{\max}} = \sum_{j=1}^{s_{\max}} C_m^j$. Максимальна складність моделей $s_{\max}$ визначається нерівністю: $P_{s_{\max}} \le P_{\max}$, де $P_{\max}$ – максимальна кількість моделей, які можна побудувати за прийнятний час. Для побудови цих моделей може використовуватись комбінаторний алгоритм з обмеженням складності COMBIS [1].

Крок 2: За значеннями критерію AR відбираються кращі моделі для формування множини $\Im_F$.

Крок 3: Кожен аргумент множини $\Im_F$ оцінюється за частотним критерієм FC оцінки рівня інформативності.

Крок 4: Перед початком наступного етапу за допомогою відсіювання найменш інформативних аргументів формується нова вибірка W зі зменшеною кількістю аргументів.

Вказані дії виконуються доти, поки кількість аргументів m, що залишились у вибірці W на певному етапі, не дозволить виконати повний перебір.

В цьому підході ми оцінюємо значущість аргументів на моделях малої складності та на кожному етапі, зменшуючи кількість аргументів на вибірці W, намагаємось, наскільки можливо, збільшити складність.

Метод BSS розглядає моделі не малої складності, як це було запропоновано в FSS, а великої. На відміну від методу FSS, який будує моделі на підматрицях X_d, що складаються з векторів x_j, $j = 1,...,s_{\max}$, метод BSS розглядає моделі на підматрицях $X_{\bar{d}}$ з векторами x_j, $j = m - s_{\max},...,m$, що доповнюють підматриці X_d до матриці X. Структурний вектор $\bar{d}$ є інвертованим по відношенню до вектора d, тобто кожна одиниця в $\bar{d}$ відповідає нулю в d і кожен нуль в $\bar{d}$ відповідає одиниці в d. До того ж кількість моделей, що розглядаються на кожному етапі в FSS, дорівнює кількості моделей з інвертованими структурними векторами, побудованими на

кожному етапі в BSS ($P_{s_{\max}} = P_{m-s_{\max}}$). На відміну від FSS, в методі BSS прохід по моделям виконується від моделей більшої складності до моделей з меншою кількістю вхідних змінних x (складність змінюється від m до $m - s_{\max}$).

Результати застосування методів BSS і FSS з використанням частотного критерію FC при розв'язанні раніше розглянутих задач наведені в табл. 2.

Метод FSS при розв'язанні задачі 1 з 20 аргументами знаходить ту ж модель, що і метод повного перебору COMBI, тільки значно швидше. При розв'язанні задачі з 200 аргументами на першому ж етапі втрачаються 5 істинних аргументів: $x_4, x_5, x_7, x_9, x_{10}$, що мають мінімальне значення FC = 0.05 серед усіх змінних (Рис. 1).

Таблиця 2

Порівняння ефективності роботи методів FSS і BSS з COMBI при розв'язанні задач з числом аргументів m=20 та m=200

Методи	*Кількість аргументів у вибірці (m)*	*Кількість істинних аргументів у кращій моделі*	*Кількість побудованих моделей*	*Час виконання, с.*	*AR*	*Помилка моделі на контрольних даних %*
COMBI	20	10	1 048 575	21	2.36	2.8
FSS	20	10	125 994	1	2.36	2.8
BSS	20	10	125 994	3	2.36	2.8
FSS	200	5	508 664	4	439.2	44.8
BSS	200	10	10 427	22	1.89	1.4

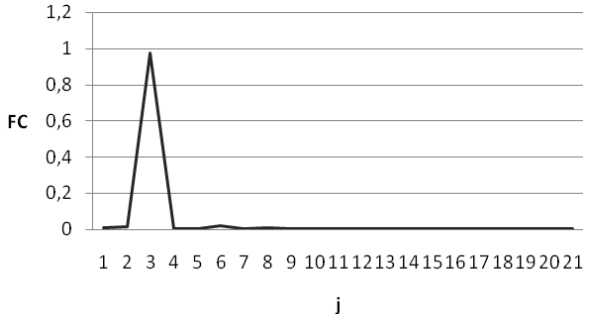

Рис. 1. Значення FC для кожного аргумента x_j, визначене на моделях складності $s = 1, 2$ при $F = m = 200$

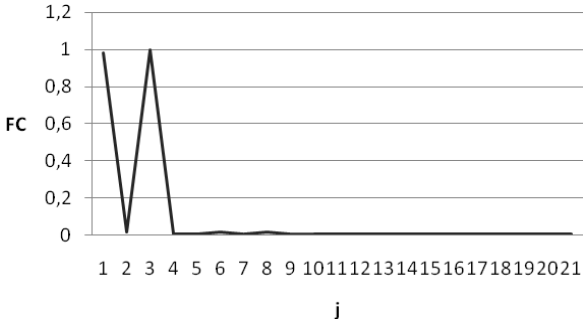

Рис. 2. Значення FC для кожного аргумента x_j, визначене на моделях складності $s = 1, 2, 3$, при $F = m = 200$

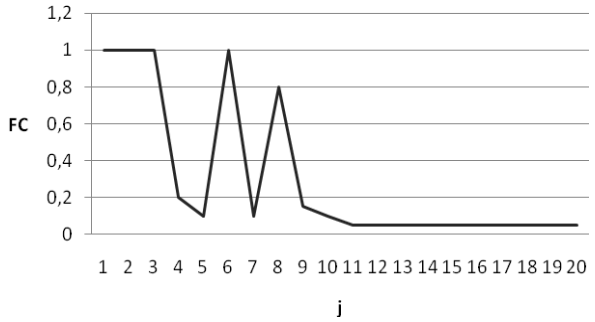

Рис. 3. Значення FC для кожного аргумента x_j, визначене на моделях складності $s = 1,...,6$, при $F = m=20$

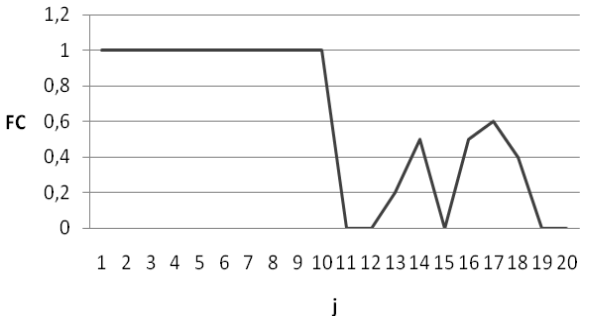

Рис. 4. Значення FC для кожного аргумента x_j, визначене на моделях складності $s = s_0,...,m$, при $F = m = 20$

Аналізуючи результати, подані на рис. 1-4, можна припустити, що значення критерію FC залежать від складності відповідних моделей. Крім того, з рис. 4 бачимо, що при розгляді моделей складності $s > s_0$ всі істинні аргументи мають максимальне значення критерію FC. Це може означати, що моделі, які містять всі істинні аргументи, є кращими за критерієм AR.

Для підтвердження цієї гіпотези розглянемо залежність значення критерію AR від присутності всіх істинних аргументів в побудованій моделі.

Розіб'ємо всю множину $\mathfrak{I}$ на підмножини $\mathfrak{I}_s, s = 1,...,m$, моделей однакової складності s. Кожну множину $\mathfrak{I}_s$ розіб'ємо ще на підмножини $\mathfrak{I}'_s$ і $\mathfrak{I}_s^{\sim}$, що складаються відповідно з моделей $f'_s \in \mathfrak{I}'_s$, які включають всі істинні аргументи, та моделей $f_s^{\sim} \in \mathfrak{I}_s^{\sim}$, які не містять хоча б одного істинного аргумента.

Використовуючи приклад 1, в якому перші 10 аргументів істинні та $m = 20$, побудуємо всі множини $\mathfrak{I}'_s$ і $\mathfrak{I}_s^{\sim}$. На кожній множині $\mathfrak{I}'_s$ знайдемо мінімальне і максимальне значення критерію регулярності AR для моделей $f'_s \in \mathfrak{I}'_s$, а на множині $\mathfrak{I}_s^{\sim}$ – мінімальне значення критерію AR для моделей $f_s^{\sim} \in \mathfrak{I}_s^{\sim}$ для кожного $s = s_0,...,m-1$ (Рис. 5).

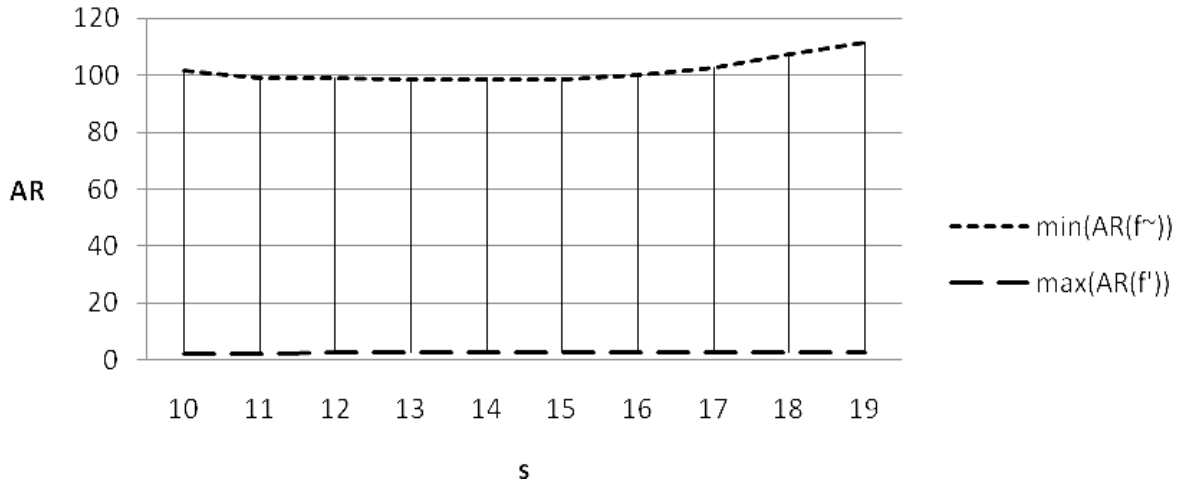

Рис. 5. Залежність критерію оцінки моделей від включення в модель всіх істинних аргументів

Наведені на рис. 5 графіки показують, що максимальні значення критерію *AR* для всіх моделей $f'_s \in \Im'_s$ значно менші від мінімальних значень критерію *AR* для всіх моделей $f_s^{\sim} \in \Im_s^{\sim}$. Це підтверджує висловлену вище гіпотезу про найбільш часте включення істинних аргументів до кращих моделей і обґрунтовує доцільність використання частотного критерію для оцінювання ступеня інформативності аргументів.

Крім того, це дає підставу вважати, що при пошуку істинної моделі необхідно враховувати не парну і не часткову кореляцію групи аргументів з вихідною величиною (рис. 1-3), а множинну кореляцію всієї групи істинних аргументів з вихідною змінною (рис. 4).

На відміну від FSS, в методі BSS частотний критерій визначається на моделях, що містять усі істинні аргументи. Тим самим враховується кореляція всієї групи істинних аргументів з вихідною величиною, що забезпечує високу надійність визначення рівня інформативності аргументів.

Висновки

У процесі досліджень виконано низку експериментів, які показують доцільність використання частотного критерію FC для оцінювання рівня інформативності аргументів. Врахування множинної кореляції всієї групи істинних аргументів з вихідною величною підвищує якість визначення інформативності аргументів, що забезпечується застосуванням частотного критерію FC в поєднанні з методом BSS.

Проведені експерименти підтвердили ефективність використання частотного критерію в методах послідовного відсіювання аргументів. Метод BSS із застосуванням частотного критерію дозволяє за допомогою неповного

перебору знаходити на вибірках великої розмірності моделі, якість яких не поступається якості моделей, знайдених за допомогою повного перебору. При цьому процес знаходження оптимальних моделей радикально прискорюється.

Література

1. Степашко В.С. Комбинаторный алгоритм МГУА с оптимальной схемой перебора моделей // Автоматика. – 1981. – №3. – С. 31 – 36.
2. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – Киев: Наук. думка, 1985. – 216 с.
3. Степашко В.С., Коппа Ю.В. Опыт применения системы АСТРИД для моделирования экономических процессов по статистическим данным // Кибернетика и выч. техника, 1998. - Вып.117. – С. 24-31.
4. Samoilenko O.A., Stepashko V.S. Combinatorial GMDH algorithm with successive selection of arguments. – Proceedings of the II International Workshop on Inductive Modelling IWIM-2007, 19-23 September 2007, Prague, Czech Republic. – Prague: Czech Technical University, 2007. – P. 139-143.
5. Самойленко О.А., Степашко В.С. Оптимізація структури моделей методом послідовного відсіювання аргументів. – Матеріали міжнародної конференції „Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту”, Євпаторія, 19-23 травня 2008 р.: – Херсон: Вид-во ХНТУ, 2008. – Том 3, Частина 2. – С. 83-86.
6. Samoilenko O., and Stepashko V. A method of Successive Elimination of Spurious Arguments for Effective Solution of the Search-Based Modelling Tasks. – Proceedings of the II International Conference on Inductive Modelling ICIM-2008, 15-19 September 2008, Kyiv, Ukraine. – Kyiv: IRTC ITS NANU, 2008. – P. 36-39.
7. Ivakhnenko A.G., Ivakhnenko G.A., Savchenko E.A., and Wunsch D. Problems of Further Development of GMDH Algorithms: Part 2 // Pattern Recognition and Image Analysis , Vol. 12, № 1, 2002, pp. 6-18.