УДК 004.94

ОПИС МЕТОДУ ПОБУДОВИ ЛІНІЙНОЇ МОДЕЛІ НА ОСНОВІ АНАЛІЗУ ПАРНИХ КОРЕЛЯЦІЙ ТА РЕЙТИНГУ РЕГРЕСОРІВ

Г.А. Піднебесна

Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, м. Київ

pidnebesna@ukr.net

У цій роботі пропонується алгоритм, який будує лінійну модель з оптимальною кількістю регресорів, заснований на кореляційному аналізі, з використанням зовнішнього критерію вибору оптимальної моделі, що базується на поділі вхідної вибірки даних. Це відносить запропонований метод до класу алгоритмів МГУА. Описано основні етапи роботи алгоритму, показані результати роботи на тестових експериментах.
Ключові слова: МГУА, індуктивне моделювання, перебірні алгоритми, кореляційний аналіз.

In this paper, the method of building of a linear model with the optimal number of regressors is proposed. It is based on a correlation analysis, using an external criterion for choosing the optimal model. The external criterion is based on the division of the input data sample. That is why the proposed method relates to the class of GMDH algorithms. The main stages of the algorithm work are described. The core feature of the algorithm is the use of frequency analysis of regressor ratings. The results of testing its effectiveness on test experiments are shown.
Keywords: GMDH, searching-out algorithms, inductive modeling, correlation analysis.

В работе предлагается алгоритм, который строит линейную модель с оптимальным количеством регрессоров, основанный на корреляционном анализе, с использованием внешнего критерия выбора оптимальной модели. Это относит предлагаемый метод к классу алгоритмов МГУА. Описаны основные этапы работы алгоритма, показаны результаты проверки его работы на тестовых экспериментах.
Ключевые слова: МГУА, индуктивное моделирование, переборные алгоритмы, корреляционный анализ.

Вступ

У сучасному світі аналіз даних став є однією з найважливіших задач як для комерційних, так і для наукових досліджень. Розроблено велику кількість різноманітних методів, однак підхід, заснований на використанні лінійної регресії, залишається одним з найпотужніших інструментів для роботи з даними. Проте, залишається актуальною проблема вибору оптимального числа інформативних регресорів.

Серед ефективних методів моделювання за експериментальними даними широко відомий метод групового урахування аргументів (МГУА) [1–3]. Він

ґрунтується на різних процедурах автоматичної побудови множини моделей-кандидатів та вибору кращих за зовнішніми критеріями селекції.

Запропонований в цій роботі метод відбору інформаційних аргументів за допомогою кореляційного аналізу відноситься до класу алгоритмів МГУА з неповним перебором моделей. Вони визначаються послідовним генеруванням дедалі складніших моделей з вибором оптимальної за заданим критерієм. Традиційно для МГУА використовується зовнішній критерій оцінки моделі, який ґрунтується на поділі вибірки на навчальну та перевірну.

Перебірні алгоритми на основі кореляційного ранжування послідовності аргументів розглядалась у роботах [4, 5]. В цій роботі пропонується дещо інший спосіб застосування парної кореляції та ранжування з метою селекції інформативних регресорів. При цьому використовується принцип побудови оптимальної (в сенсі мінімального значення критерію) моделі на основі аналізу рейтингу потраплянь регресорів у кращі часткові моделі. На відміну від частотного аналізу регресорів, запропонованого в [6], в цій роботі застосовується процедура кластеризації для вибору регресорів з великим рейтингом. Відповідні регресори входять в остаточну модель, коефіцієнти якої розраховуються за МНК на всій вибірці.

Загальна ідея

Запропонований спосіб відноситься до класу методів індуктивного моделювання, які визначаються послідовним генеруванням дедалі складніших моделей з вибором оптимальної за заданим критерієм. Традиційно для МГУА використовується зовнішній критерій оцінки моделі, який ґрунтується на поділі вибірки на навчальну та перевірну.

Для вибраної навчальної підвибірки на кожному кроці певним чином визначається, чи повинен регресор включатися в модель. Для цього аналізується кореляція регресорів з поточними залишковими моделями (на першому кроці – з заданим значенням вихідної змінної). За МНК будується множина моделей. Найкраща модель вибирається з використанням критерія регулярності, розрахованого на перевірній підвибірці вхідних аргументів (регресорів). Якщо регрессор вибирається на цій стадії, ми говоримо, що він має рейтинговий бал. Отримана на поточному розбитті найкраща за значенням критерія модель зберігається для подальшого аналізу.

Описана процедура багаторазово повторюється для різних способів розбиття вибірки, щоб підрахувати рейтинг (рейтингові бали) регресорів. Після завершення описаної процедури вибирається остаточна модель на основі аналізу за допомогою процедури кластеризації отриманого рейтингу

регресорів. Кінцевою моделлю вважається модель, до складу якої входять регресори з великою кількістю рейтингових голосів.

Опис числових експериментів

Випадковим чином формується:

- матриця вхідних аргументів X розмірності [n x m], де n – кількість регресорів (n = 50), а m (m = 1000) – кількість вимірювань.
- структурний вектор [1 x n], елементами якого є нулі (якщо відповідний регресор не входить до моделі) та одиниці (на місцях тих регресорів, які включаються до складу моделі),
- вектор коефіцієнтів A [1 x n].

Розраховується вектор вихідних значень Y [1 x n]: $Y=AX + \varepsilon$, де ε - білий шум (10%),.

Всі регресори не мають рейтингових балів на початку: ($V(x_1) = 0,\ldots, V(x_n) = 0$).

1. *Поділ вибірки вхідних даних.* Вибірка вхідних даних ділиться на дві підвибірки: навчальну та перевірну. В описаному експерименті вибірка ділиться випадковим чином навпіл. Параметри поділу вибірки можуть змінюватись.

2. *Побудова набору моделей-кандидатів на навчальній підвибірці.* Розраховуються значення парної кореляції кожного з регресорів вихідною змінною. Одночасно обчислюється ймовірність того, що ця величина кореляції є статистично значущою (*p-Val*).

З множини всіх регресорів вибирається підмножина «перспективних» (інформативних). «Неперспективні» (неінформативні) вилучаються з подальшого розгляду.

Рішення про включення перспективних регресорів в модель приймається відповідно до значення кореляції та значення *p-Val*. Сортуються регресори від найбільшого абсолютного значення кореляції до менших, при $p\text{-}Val \leq 0{,}5$.

На першому кроці розглядається кореляція з вихідною змінною Y. В подальшому при послідовному додаванні до моделі одного регресора розраховується кореляція регресорів з поточними залишками Y_r (різницею табличного значення змінної Y та модельного значення Y_{mod}: $Y_r = Y - Y_{mod}$). Додається до моделі найбільш корельований регресор x_{i_1} (з найменшим значенням *p-Val*). Процедура повторюється, доки є перспективні регрессори.

3. *Вибір найкращої моделі*. Розглянемо набір вибраних перспективних регресорів $bestIndX = \{ x_{i_1}, \ldots, x_{i_k} \}$. Вони утворюють множину k ($k \leq n$) моделей:

$$y_1 = a_{11} x_{i_1},$$

...

$$y_k = a_{1k} x_{i_1} + ... + a_{kk} x_{i_k}.$$

Для кожної моделі на перевірній підвибірці розраховується значення критерію. Традиційно для алгоритмів МГУА застосовується критерій регулярності. Вибирається краща модель у сенсі мінімуму критерію. Модель включає регресори $\{x_{i_1}, ..., x_{i_q}\}$, $q \le k$, $q \in \{i_1, ..., i_k\}$ з множини перспективних регресорів *bestIndX*. Кожен з тих регресорів, що увійшли до кращої моделі, отримує рейтинговий бал (додається одиниця до його рейтингу).

На рис. 1 для прикладу показано розраховані значення критерію для 50-ти моделей, побудованих шляхом послідовного додавання одного регресора відповідно описаній вище процедурі. Вертикальною лінією позначено мінімум критерію. Він відповідає побудованій моделі, до складу якої вибрано 28 регресорів, що відповідає істиній моделі наведеного числового експеримента.

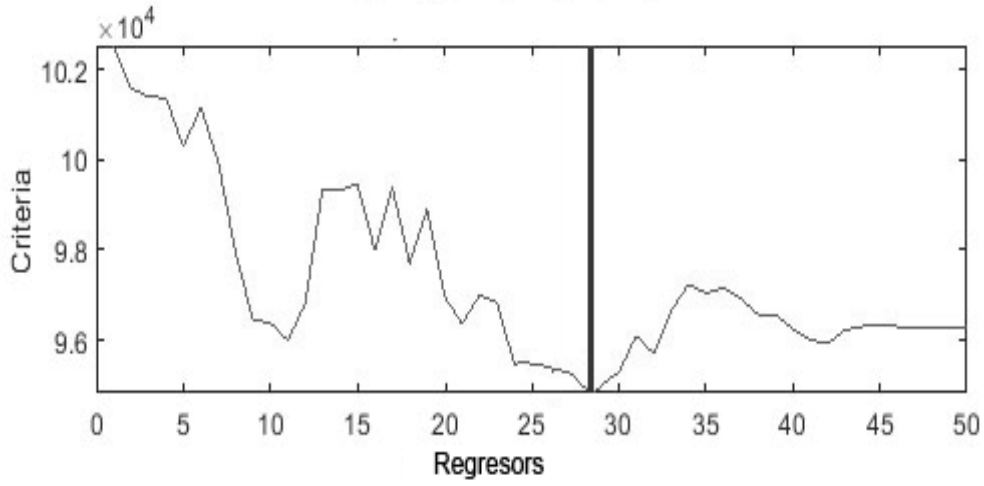


Рис. 1 Вибір моделі за мінімальним значенням критерію

4. *Підрахунок рейтингу регресорів.* Процедура моделювання (кроки 1–3) повторюється задану кількість разів. В наведеному числовому експерименті було 10 повторів. Підраховується рейтинг тих регресорів, які входять до кращої моделі на поточному розбитті вибірки. Сірим кольором на рис. 2 позначено випадки, коли регресор потрапив до моделі під час 10 розбиттів вибірки. Чорним – якщо регресор не потрапив до моделі.

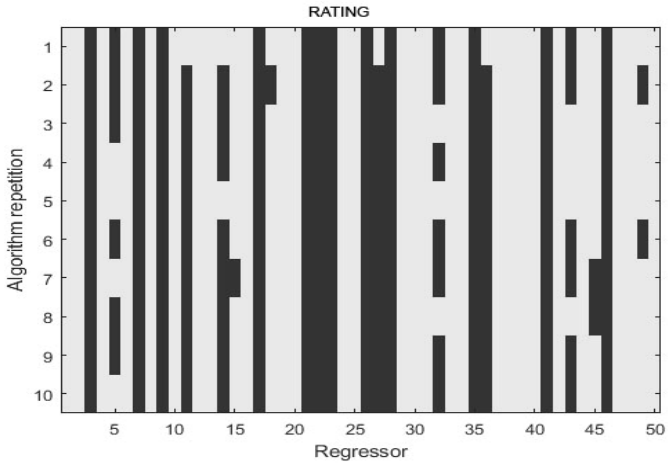


Рис. 2 Рейтинг регресорів

5. *Вибір регресорів для остаточної моделі*. Регресори, які будуть включені до остаточної моделі, вибираються на підставі аналізу отриманого рейтингу. Зроблено це може бути різними шляхами.

В пропонованому алгоритмі відбір робиться за допомогою процедури кластеризації, яка розділяє регресори на групи з низьким та високим числом рейтингових балів (рис. 3). Використовується алгоритм кластеризації методом *k*-середніх за допомогою стандартної функції *k-means* в *MatLab*. Критерієм кластеризації є мінімум внутрікластерної суми відстаней точок кластера до його центроїда. За відстань точок кластера до його центроїда береться квадрат евклідової відстані. До остаточної моделі вибираються регресори з найбільшим рейтингом (тобто, з кластера, для якого середнє значення є більшим).

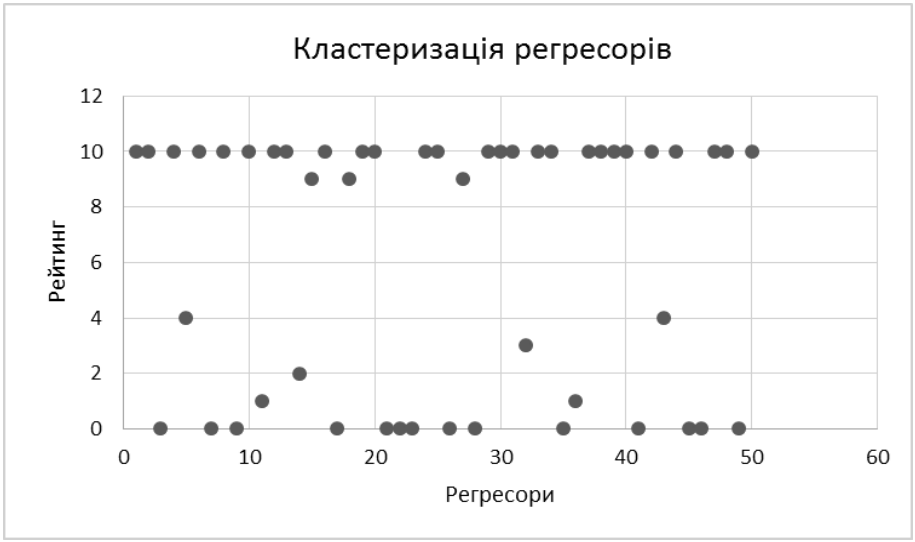


Рис. 3 Кластеризація регресорів за рейтингом

6. *Побудова остаточної моделі.* Для вибраної остаточної структури моделі розраховуються коефіцієнти та значення вихідної змінної.

В таблиці наведено значення істинних (заданих) коефіцієнтів a_i та модельних $\hat{a}_i$, отриманих за допомогою описаного алгоритму, $i = \{1, \ldots, 47\}$ – це індекси тих регресорів, що увійшли в модель.

Таблиця. Істині та модельні коефіцієнти

a_i	1.2	1.8	1.4	0.8	1	1.2	1.8	2	0.2	0.8	0.2	1
$\hat{a}_i$	1.14	1.69	1.29	0.90	1.00	1.25	1.83	1.99	0.20	0.80	0.14	0.92
i	1	2	4	6	8	10	12	13	15	16	18	20

a_i	2	1.2	0.4	0.2	1.6	0.6	0.4	1.4	1.2	1.8	1	0.8
$\hat{a}_i$	2.09	1.23	0.53	0.22	1.59	0.55	0.38	1.44	1.33	1.78	0.95	0.79
i	29	30	31	33	34	37	38	39	40	42	44	47

В наведеному прикладі з 28 істинних регресорів за допомогою описаного алгоритму було визначено 28, які увійшли в остаточну модель.

Характеристики результатів роботи запропонованого алгоритму

Для того, щоб оцінити, наскільки ефективно запропонований алгоритм знаходить істині аргументи при побудові моделі, було застосовано Коефіцієнт Жаккара [7], а також такі статистичні характеристики бінарної класифікації, як *чутливість* (sensitivity) і *специфічність* (specificity) [8] (рис. 4).

Коефіцієнт (міра подібності) Джаккара (*the Jaccard index*) - бінарна міра подібності, яка використовується для оцінки подібності скінченних множин і визначається як міра спільної частини, поділена на міру об'єднання множин:

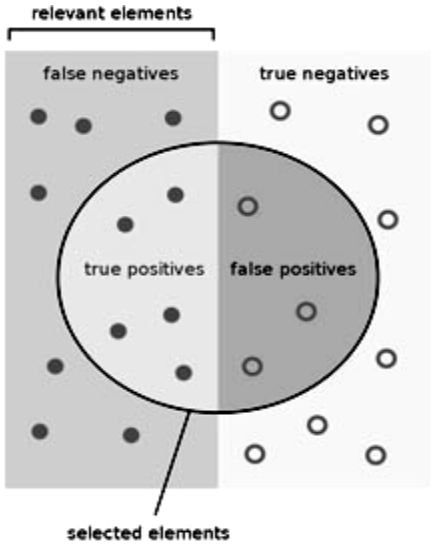
$$J(O,E) = \frac{|O \cap E|}{|O \cup E|} = \frac{|O \cap E|}{|O| + |E| - |O \cap E|}, \; 0 \le J \le 1$$

З означення видно, що індекс Джаккара належить проміжку від нуля до одиниці: $0 < J < 1$. Подібність множин тим краща, чим ближче показник індексу до одиниці.

Чутливість (яка в деяких галузях також називається справжньою позитивною швидкістю (*true positive rate*) або ймовірність виявлення (*probability of detection*)) вимірює частку фактичних позитивів, які правильно визначені як такі. Тобто, кількість винайдених істиних аргументів, які увійшли до моделі.

Специфічність (яка також називається справжнім негативним показником - *true negative rate*) вимірює частку фактичних негативів, які правильно ідентифіковані як такі (тобто, тих регресорів, які не повинні входити в модель, «неінформативних»).

Зауважимо, що терміни «позитиви» і «негативи» означають присутність або відсутність екземпляра (аргумента) у множині.



а)

б) в)

Рис. 4 Графічне зображення понять бінарної класифікації *індекс Джаккара* (а), *чутливість* (sensitivity) (б) та *специфічність* (в)

В описаному випадку порівнюються множини істиних табличних регресорів (O) та тих, що увійшли в модель (E), вибраних алгоритмом.
Запропонований алгоритм має індекс Жаккара $J \approx 0{,}88$, чутливість $sns \approx 0{,}86$ та специфічність $spc = 1$, тобто добре знаходить інформативні регресори та не включає до моделі зайвих.

Висновки

Запропоновано метод побудови по статистичній вибірці даних оптимальної (у сенсі мінімуму зовнішнього критерію) лінійної моделі з найменшою можливою кількістю регресорів. Алгоритм базується на застосуванні парної кореляції, підрахунку й аналізі рейтингу регресорів, та процедури кластеризації для відбору інформативних регресорів.

Результати числових експериментів дають підстави вважати перспективними подальші дослідження методу з застосуванням різних процедур розбиття вибірки, різних критеріїв відбору кращої моделі. Для усунення небезпеки втрачання інформативних регресорів варто розглянути свободу вибору (вибір декількох кращих моделей на кожному кроці). Перспективним може бути застосування рекурентної процедури оцінювання параметрів при побудові моделі.

Література

[1] H.R. Madala, and A.G. Ivakhnenko, Inductive Learning Algorithms for Complex Systems Modeling. New York: CRC Press, 1994, 384 p

[2] V. Stepashko, “Developments and Prospects of GMDH-Based Inductive Modeling”, In: Advances in Intelligent Systems and Computing II. CSIT 2017 / Shakhovska N., Stepashko V. (eds), AISC series, vol. 689, Cham: Springer, 2018, pp. 474-491.

[3] O. Moroz, V. Stepashko, "Hybrid Sorting-Out Algorithm COMBI-GA with Evolutionary Growth of Model Complexity," Advances in Intelligent Systems and Computing II / N. Shakhovska, V. Stepashko, Editors, AISC book series, Berlin: Springer Verlag, vol. 689, pp. 346-360, 2017.

[4] 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.

[5] O. Koshulko, A. Koshulko, "Multistage Combinatorial GMDH algorithm for parallel processing of high-dimensional data," Proc. of the 3rd Intern. Workshop on Inductive Modelling (IWIM2009), Sep. 14–19th, Krynica, Rzeszow, Poland, 2009, pp. 114–116.

[6] O. Samoilenko, V. Stepashko, “A Method of Successive Elimination of Spurious Arguments for Effective Solution of the Search-Based Modelling Tasks,” Proc. of the II Int. Conf. on Inductive Modelling, Sept. 2008, pp. 36-39

[7] Gower, J.C. “Measures of Similarity, Dissimilarity and Distance.” John Wiley & Sons, Inc. 2004 (https://doi.org/10.1002/0471667196.ess1595).

[8] P. Willett, J.M. Barnard, and G.M. Downs, “Chemical similarity searching.” Journal of Chemical Information and Computer Sciences, 38(6): pp. 983–996, 1998.