УДК 004.94

КІР'ЯНОВ О.В.

ПОКРАЩЕННЯ РЕЗУЛЬТАТІВ ОТРИМАНИХ ПОЛІНОМІАЛЬНИМ БАГАТОРЯДНИМ МГУА ДЛЯ ЗАДАЧ З ВЕЛИКОЮ КІЛЬКІСТЮ ЗМІННИХ ЗА ДОПОМОГОЮ ДОДАВАННЯ ЗВОРОТНИХ ЗМІННИХ

В даній статті запропоновано алгоритм, який за допомогою додавання зворотних змінних покращує результати отримані поліноміальним багаторядним МГУА для задач з великою кількістю змінних. Особливістю цього алгоритму є те, що для відбору найінформативніших змінних використовуються результати отримані за допомогою МГУА . Проведено порівняльний аналіз наведеного алгоритму з поширеним алгоритмом з використанням коефіцієнтів кореляції; експериментально доведено, що наведений алгоритм дає кращі результати.

This article describes algorithm that improves results found with the help of polynomial multilayered GMDH by introducing reverted variables. Peculiarity of this algorithm is that it uses results got by GMDH to select the most informative variables. Comparative analysis of the described algorithm and widespread algorithm that use correlation coefficients was done. It was proved experimentally that described algorithm gave better results.

Вступ

Метод групового урахування аргументів (МГУА) був запропонований наприкінці 60-х – початку 70-х років академіком О.Г. Івахненко. Цей метод використовує ідеї самоорганізації і механізми живої природи – схрещування (гібридизацію) і селекцію (добір). МГУА це не один алгоритм, а цілий спектр алгоритмів, кожен з яких вирішує задачі при певних умовах. Найбільш поширеними варіаціями є: комбінаторний, багаторядний, з послідовним виділенням трендів, нечіткий та інші [1].

МГУА показав високу ефективність при застосуванні в різних сферах: аналізу та прогнозуванні економічних систем; медичної діагностиці; аналізу та прогнозуванні екологічних систем; прогнозу погоди, тощо [2].

Індуктивні алгоритми МГУА передбачають перебір усіх можливих варіантів (моделей) і вибір найкращої моделі. Для отримання найкращих результатів додають якомога більше вхідних змінних. Однак, через перебір усіх варіантів час розрахунків залежить експоненціально від кількості вхідних змінних. Для задач з великою кількістю змінних (до 500) не можна просто додавати зворотні змінні на вхід МГУА, тому що це «значно» збільшить час розрахунків. Тому використовують різні методи для відсіювання найменш інформативних змінних. Рекомендований спосіб відсіювання це спосіб з використанням коефіцієнтів кореляції, коли визначаються коефіцієнти кореляції вхідних змінних з виходом моделі, розміщу-

ються змінні у порядку спадання коефіцієнтів кореляції та відбираються перші *n* змінних [1, 3, 4].

Існує інше перспективне рішення для комбінаторного МГУА – це комбінаторний МГУА з послідовним відбором змінних (Combinatorial GMDH algorithm with successive selection of arguments). Основна ідея полягає в тому, що відбираємо найінформативніші змінні за допомогою самого ж МГУА [3, 4]. У цій статті ця ідея була розвинена для поліноміального багаторядного алгоритму МГУА.

Постановка задачі

Припустимо, що є багатовимірна вибірка даних *W*, яка складається з *n* спостережень за багатовимірною змінною $X = \{X_1, X_2, ..., X_m\}$. Необхідно знайти залежність кожної змінної X_i від усіх інших: $X_i = f_i(X/X_i), i = 1..m$, використовуючи поліноміальний багаторядний алгоритм МГУА. Результатом роботи поліноміального багаторядного алгоритму МГУА є поліном Колмогорова-Габора [1], в нашому випадку він буде мати наступний вигляд:

$$f_i(X/X_i) = a_0 + \sum_{j=1}^{m-1} a_j x_j + \sum_{j=1}^{m-1}\sum_{k=1}^{m-1} a_{jk} x_j x_k + ... \qquad (1)$$

Якість знайденої залежності f_i повинна оцінюватись за зовнішнім критерієм регулярності $AR(f_i)$. Критерій регулярності визначається наступним чином: вибірка W_i формується з *W* видаленням спостережень за змінною X_i; далі W_i поділяється на дві підвибірки W_{iA}(навчальну) та W_{iB}(перевірочну). f_{iA} – це залежність f_i коефіцієнти $a_j, a_{jk}, ...$ якої знайдені на вибірці W_A. Тоді значення критерію регулярності для залежності f_i знаходиться за наступною формулою: $AR(f_i) = \| y_B - f_{iA}(W_B) \|$. При порівнянні двох залежностей кращою залежністю вважається та, значення критерію якої менше.

Необхідно покращити знайдені залежності за допомогою додавання зворотних змінних.

Розв‘язання задачі

Багаторядні алгоритми МГУА

Відповідно до поставленої задачі, для розв’язку скористуємося багаторядним алгоритмом МГУА, який широко застосовується для рішення некоректних чи недовизначених задач моделювання, тобто у випадку, коли число точок у таблиці дослідних даних менше числа аргументів, що

входять у синтезовану модель. Але він також успішно застосовується для задач, коли вихідних даних досить для застосування однорядного МГУА.

Багаторядні алгоритми працюють за наступною схемою:

Будуються часткові описи від усіх попарних комбінацій початкових даних $y_1 = f(x_1, x_2), y_2 = f_2(x_1, x_3), ..., y_k = f_k(x_{n-1}, x_n)$. Коефіцієнти знаходяться за допомогою МНК (метод найменших квадратів). З цих моделей вибирається деяке число кращих за зовнішнім критерієм селекції.

Змінні, отримані на попередньому кроці, разом з початковими даними формують вхідні данні для нових часткових описів: $z_1 = \varphi(x_1, y_1), z_2 = \varphi_2(x_1, y_2), ..., z_l = \varphi_k(y_{k-1}, y_k)$. Знов відбираються найкращі моделі за зовнішнім критерієм селекції і передаються на наступний крок.

Якщо значення зовнішнього критерію зменшується, то переходимо на крок 2, інакше алгоритм зупиняється.

Кожний частковий опис може бути лінійним:

$$f = a_0 + a_1 x_i + a_2 x_k$$

або нелінійним

$$f = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j .$$

Критерій селекції вибирається в залежності від «якості» вхідних даних та мети побудови поліномів (знаходження зв‘язків, побудова короткострокового прогнозу, тощо). В нашому випадку, в постановці було визначено, що необхідно використовувати зовнішній критерій регулярності.

Тепер зробимо припущення, які допоможуть нам розв‘язати поставлену задачу.

Припущення №1. Припустимо, що за допомогою поліноміального багаторядного МГУА знайдено оптимальну за зовнішнім критерієм залежність (1). Усі x_l при яких стоїть ненульовий коефіцієнт a_l будуть найінформативнішими змінними для y.

Наприклад, знайдено залежність $y = a_0 + a_1 x_2 + a_2 x_4 x_5$. Тоді для y інформативними змінними будуть x_2, x_4, x_5.

Дійсно, якщо МГУА побудував оптимальну (чи одну з оптимальних) залежність, то ці змінні будуть найінформативніші. Якщо ми візьмемо інші змінні або викинемо одну з них, то отримаємо «гіршу» залежність, тому що при переборі МГУА відібрав залежність з мінімальним значенням зовнішнього критерію.

Припущення №2. Припустимо, що за допомогою поліноміального багаторядного МГУА знайдено оптимальну за зовнішнім критерієм залеж-

ність (1). Відбираємо усі одночлени в які входить змінна x_1 і при яких стоїть ненульовий коефіцієнт a_1. Усі інші змінні, що також входять в відібрані одночлени, будуть інформативними для змінної x_1, але їх спочатку потрібно обернути ($\frac{1}{x_i}$).

Наприклад, нехай знайдено залежність $y = a_0 + a_1 x_2 x_4 x_6 + a_2 x_4 x_5$. Тоді для змінної x_4 інформативними змінними будуть $\frac{1}{x_2}, \frac{1}{x_5}, \frac{1}{x_6}$, для змінної x_2 – $\frac{1}{x_4}, \frac{1}{x_5}, \frac{1}{x_6}$ тощо.

Припустимо, що у знайденому поліномі змінна x_k присутня лише у одному одночлені: $y = a_0 + x_k f_1(X/x_k) + f_2(X/x_k) + f_3(X/x_k) + \ldots$, де $f_i(X/x_k) = a_i \prod_{j=1, j\neq k}^{k} x_j$.

Перенесемо x_k в ліву частину, а все інше у праву:

$$x_k = \frac{y - a_0 - f_2(X/x_k) - f_3(X/x_k) - \ldots}{f_1(X/x_k)}.$$

Тобто x_k залежить від $\frac{1}{f_1(X/x_k)} = \frac{1}{a_1 \prod_{j=1, j\neq k}^{k} x_j}$. З цього випливає, що x_k залежить від усіх змінних з $f_1(X/x_k)$, але взятих оберненими. Зробимо евристичне припущення, що такий самий результат ми отримуємо і у випадку коли змінна x_k присутня в декількох одночленах і її не можна виразити через інші змінні.

На основі наведених вище припущень розроблено алгоритм, який складається з трьох кроків:

Алгоритм

Знаходимо p найкращих залежностей для кожної змінної X_i використовуючи поліноміальний багаторядний МГУА:

$f_{ij}(X/X_i), i = 1..m, j = 1..p$

Для $\forall X_i$ визначаємо множину найінформативніших змінних X_{Ei}, що складається як зі звичайних вхідних змінних X_j так і зворотних $\frac{1}{X_j}$:

Всі X_j, що входять у p найкращих поліномів для X_i, потрапляють у X_{Ei} (на основі Припущення №1)

Знаходимо всі X_j, що помножуються на X_i в будь якому поліномі. $\frac{1}{X_j}$ потрапляють у X_{Ei} (на основі Припущення №2).

Використовуючи багаторядний поліноміальний алгоритм МГУА і X_{Ei} як набір вхідних змінних знов знаходимо залежності.

Недоліком цього алгоритму є те, що кількість вхідних змінних, які використовуються на кроці 3, може бути набагато більше ніж на 1-ому кроці. А так як час роботи використаного алгоритму МГУА залежить експоненціально від кількості вхідних змінних, то це може призвести до значного збільшення часу розрахунків. Тому X_{Ei} повинна містити не більше змінних ніж початкова кількість змінних m, щоб не збільшити «значно» час розрахунків за цим алгоритмом. Якщо кількість змінних у X_{Ei} перевищує m, тоді є 2 варіанти:

Зменшити кількість поліномів, з яких ми отримуємо найінформативніші змінні (рекомендований варіант)

Відсіяти змінні з найменшими коефіцієнтами кореляції.

Таким чином, час розрахунків в середньому збільшиться приблизно у 2 рази, так як при другому розрахунку кількість змінних не перевищує початкової кількості.

Експерименти

Експерименти проводилися на даних з сайту http://www.gmdh.net/GMDH_dat.htm. Мета експериментів була виявити який з двох алгоритмів кращий: алгоритм з використанням коефіцієнтів кореляції чи наведений вище алгоритм. В таблицях попарно порівнюються результати отримані за допомогою багаторядного поліноміального МГУА: без використання зворотних змінних (далі буде називатися «МГУА без звор. змін.»), з використання зворотних змінних відібраних за допомогою алгоритму з коефіцієнтами кореляції («МГУА зі звор. змін. за коеф. корел.»), з використанням зворотних змінних відібраних за допомогою наведеного вище алгоритму («МГУА зі звор. змін. за алгоритмом»). Будемо вважати, що поліном покращився, якщо значення зовнішнього критерію селекції зменшилося більше ніж на 5%. І відповідно, будемо вважати, що поліном погіршився, якщо значення зовнішнього критерію селекції збільшилося більше ніж на 5%.

Таблиці мають наступну структуру: в першому стовпчику дається назва моделі, в наступних – кількість залежностей, які покращились або погіршились.

Назва моделі	**«МГУА зі звор. змін. за коеф. корел.»**				**«МГУА зі звор. змін. за алгоритмом»**			
	Порівняння з «МГУА без звор. змін.»		**Порівняння з «МГУА зі звор. змін. за алгоритмом»**		**Порівнянні з «МГУА без звор. змін.»**		**Порівнянні з «МГУА зі звор. змін. за коеф. корел.»**	
	Покращилось	**Погіршилось**	**Покращилось залеж-**	**Погіршилось**	**Покращилось**	**Погіршилось**	**Покращилось**	**Погіршилось**

	залежно-стей	залеж-ностей	ностей	залежно-стей	залеж-ностей	залеж-ностей	залеж-ностей	залеж-ностей
Демографія	5	0	6	2	5	1	2	6
Сумарна демографія	12	2	6	9	12	2	9	6
Сумарні макроекон. показники	7	3	5	8	12	0	8	5
Показники індустрії	2	11	1	13	10	4	13	1
Кількість зайнятих в індустрії	4	5	3	4	4	5	4	3
Інфляція	7	4	0	9	12	1	9	0
Випуск продукції	3	3	3	0	3	3	0	3

Якщо порівнювати розглянуті алгоритми з «МГУА без звор. змін.», то з **табл.1** ми бачимо, що більшість поліномів покращились після використання наведеного вище алгоритму (72% залежностей покращилось). Трохи гірший результат дає алгоритм з використанням коефіцієнтів кореляції (покращилось майже 50%залежностей). Якщо ж порівняти результаті цих двох алгоритмів, то ми бачимо, що покращилось 56% поліномів проти 30%, які погіршились. Таким чином, можна зробити висновок, що наведений вище алгоритм дає кращі результати ніж алгоритм з використанням коефіцієнтів кореляції.

Необхідно додати, що поліноми знайдені на першому кроці алгоритму не повинні бути загублені. Маючи повний набір поліномів отриманих як з використанням зворотних змінних так і без них, ми повинні відібрати найкращі залежності. Таким чином в нас не буде поліномів, які погіршились після введення зворотних змінних.

Висновки

Використання алгоритму МГУА для визначення найінформативніших змінних було започатковано в комбінаторному МГУА з послідовним відбором змінних. В цій статті запропоновано метод відбору звичайних і зворотних змінних з використанням поліноміального багаторядного МГУА. Був проведений порівняльний аналіз з алгоритмом відбору змінних з використанням коефіцієнтів кореляції і показано, що запропонований алгоритм дає кращі результати, тобто для більшості змінних знаходить залежності з меншим значенням критерію регулярності.

Даний напрямок є дуже перспективним. Подальша робота може бути спрямована на поширення використання МГУА для відбору інформативних змінних як для інших алгоритмів (як пов’язаних з МГУА так і не пов‘язаних).

Список використаної літератури

1. Зайченко Ю.П. Основи проектування інтелектуальних систем – К.: Видавничий дім «Слово», 2004. – 352 с.
2. Madala H.R ., Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.
3. Samoilenko O.A., Stepashko V.S. Combinatorial GMDH algorithm with successive selection of arguments. IWIM, Prague, 2007
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, p.6-18.