УДК 519.2

СТРУКТУРНЕ МОДЕЛЮВАННЯ СТІЙКЕ ДО ВИКИДІВ У ВХІДНИХ ТА ЗАЛЕЖНИХ ЗМІННИХ

В. Шапошник[1], А. Е. П. Вілла[2], Т. Аксенова[3]

[1]*НВК «Інституту прикладного системного аналізу», НТУУ «КПІ», Київ, Україна;*
[2] *відділ нейроевристичних досліджень, Інститут систем інформації, Університет м. Лозанна, Швейцарія;*
[3] *НВК «Інституту прикладного системного аналізу», НТУУ «КПІ», Київ, Україна;RTRA, Фонд "Нейронауки на межі наноелектроніки", Гренобль, Франція*

Ця стаття описує удосконалення алгоритму для розв'язку задач нелінійного структурного моделювання та оцінки параметрів мультиваріативних нелінійних моделей за наявності у вхідних та залежних змінних викидів точок. Алгоритм, з одного боку, як і його попередник, базується на поліноміальній нейронній мережі і побудований на основі Методу Групового Урахування Аргументів, що наділяє його властивостями універсального еволюційного структурного моделювання. З іншого боку, він удосконалений повністю робастною системою оцінки параметрів моделі, що розширює сферу його застосування на дані з викидами як у залежних, так і у вхідних змінних. Запропонований алгоритм на тестових прикладах продемонстрував здатність коректно обробляти інформацію з викидами, забезпечуючи високу точність синтезу структури моделі та оцінки її параметрів.

Ключові слова: поліноміальна нейрона мережа, робастна регресія, нелінійна регресія, GM-оцінки, структурне моделювання.

This paper describes advances in the algorithm development designed to solve a task of optimal polynomial model selection on multivariate data sets in presence of outliers in both explanatory and response variables. On one side novel algorithm, as its ancestor, is based on GMDH-type PNN, which gives him an universal model structure identification abilities thanks to the evolving adaptively synthesized bounded network. And on the other side the algorithm is enhanced with GM-estimator used for parameter search which allows him achieve robustness to outliers in both explanatory and response variables. Enhanced RPNN demonstrated robustness to outliers in both explanatory and response variables and good accuracy of the automatic structure syntheses.

Keywords: polynomial neural network, robust regression, non-linear regression, GM-estimators, structure selection.

Эта статья описывает усовершенствование алгоритма разработанного для решения задач нелинейного структурного моделирования и оценки параметров мультивариативных нелинейных моделей при наличии в исходных и зависимых переменных точек-отклонений. Алгоритм, с одной стороны, как и его предшественник, базируется на полиномиальной нейронной сети, основанной на Методе Группового Учета Аргументов, которая наделяет его свойствами универсального эволюционного метода. С другой стороны, он усовершенствован полностью робастной системой оценкой параметров модели, что расширяет область его применения на данных с зашумленными входными и зависимыми переменными. Предложенный алгоритм на тестовых примерах продемонстрировал способность корректно обрабатывать зашумленную информацию, обеспечивая высокую точность синтеза структуры модели и оценки ее параметров.

Ключевые слова: полиномиальная нейронная сеть, робастная регрессия, нелинейная регрессия, GM-оценки, структурное моделирование.

1. Вступ

Найпоширенішою задачею наукових досліджень є пояснення фізичних залежностей і побудова моделей, що описують ті чи інші процеси. Найчастіше дослідник за деякою множиною експериментальних величин (вхідних змінних), якими він може маніпулювати, та величин, які він спостерігає (залежні змінні) має вивести тип їх взаємозв'язку. Часто залежні змінні містять грубі похибки вимірювання, а у вхідних змінних зустрічаються точки, які дуже відрізняються від більшості даних і мають бути опрацьовані окремо. Задача робастної регресії виявити закономірність між вхідними та залежними змінними, за умови, що дані можуть містити суттєві похибки.

Розглянемо наступну регресійну модель:

$$\mathbf{y} = f(\mathbf{x}; \beta_{\mathbf{0}}) + \xi \qquad (1)$$

де $f(\cdot)$ функція залежності, нелінійна в загальному випадку, $\mathbf{x} = \{x_1, ..., x_m\}$ $\mathbf{x} \in R^m$ вектор незалежних вхідних змінних, $\beta_{\mathbf{0}} \in R^p$ вектор параметрів моделі залежності, $\mathbf{y}$ залежна змінна, а ξ похибка вимірювань.

Припускаємо, що у вхідних змінних $\mathbf{X}$ і в залежній змінній $\mathbf{y}$ можуть бути присутніми значні похибки вимірювання або величини, розподіл яких значно відрізняються від більшості даних – надалі такі точки називаються викидами. Очевидно, що викиди можуть спричинити значні відхилення у оцінці параметрів моделі $\beta_{\mathbf{0}}$, а також, що набагато гірше, у побудові самої моделі $f(\cdot)$. Для лінійних задач розроблені методи робастної оцінки параметрів моделей (Метод Найменших Медіан [16], алгоритми S-оцінки [9,15], MM-оцінки [17]), які характеризуються *стійкістю* до значного числа викидів у даних [7,4]. Для побудови самих моделей (для нелінійного випадку) існує клас алгоритмів, які базуються на Методі Групового Урахування Аргументів (МГУА), котрий, що правда, не має засобів обробки викидів. Ця стаття описує удосконалений алгоритм, який об'єднує переваги обох підходів, що дозволяє виконувати задачу структурного пошуку поліноміальної моделі із стійкістю до викидів у початкових даних.

2. Методи

2.1. Метод групового урахування аргументів

МГУА [18,10] заснований на переборі моделей з поступовим їх ускладненням. У процесі роботи алгоритм шукає структуру моделі та ступінь впливу вхідних змінних на залежну змінну. За допомогою індуктивного методу самоорганізації для неточних, зашумлених та/або коротких вибірок може бути знайдена оптимальна нефізична модель у явному вигляді точність прогнозу часто якої вище, а структура – простіше, ніж структура повної фізичної моделі.

МГУА вирішує багатовимірну задачу мінімізації:

$$\hat{m} = \arg\min_{m \subset M} CR(m), CR(m) = f(P, S, \xi^2) \quad (2)$$

де M – множина моделей, що розглядаються, $CR(m)$ – зовнішній критерій якості моделі m з множини G, P – кількість вхідних змінних, S – складність моделі та ξ^2 – очікувана дисперсія шумів.

В основі лежить послідовна перевірка моделей з вибором найкращих кандидатів у відповідності до заданого критерію. Генерація множини моделей для перевірки здійснюється функцією-генератором, яка створює нову модель виходячи з підмножини вхідних змінних та моделей попереднього кроку. Зазвичай використовують поліноміальні генераторні функції, однією із можливих є $g_{ijk}(\bullet) = \alpha_1 x^i + \alpha_2 x^j x^k$. Докладно ітеративні алгоритми МГУА описані в [18,10].

2. Стійкий критерій оцінки моделі

Розглянемо критерії оцінки моделей у алгоритмі, що використовується замість оригінального Методу найменших квадратів. Ми зупинимось виключно на методах *лінійної* регресії, так як МГУА базується саме на лінійному регрессорі, а пошук в просторі нелінійних моделей забезпечується генераторною функцією.

Розв'язок виразу (1) у випадку лінійної моделі базується на мінімізації максимальної правдоподібності по вектору параметрів моделі β_0. Класична узагальнена оцінка максимальної правдоподібності (М-оцінка) для лінійного випадку була запропонована Хубером [8] і має вигляд:

$$min_\beta \sum_{i=1}^{n} \rho(r_i), \quad (3)$$

де r_i – залишки між побудованою моделлю та даними спостереження $r_i = y_i - f(x, \beta)$, а $\rho(\cdot)$ симетрична функція оцінювання залишків з єдиним мінімумом у точці 0. Зазвичай функція ρ розглядається як функція штрафів для викидів у залежній змінній $\mathbf{y}$. Якщо функція $\rho(\cdot)$ має першу похідну, то вираз (3) може бути записаний у такій формі:

$$\sum_{i=1}^{n} \rho'(r_i(\mathbf{x};\beta)) r_i(\mathbf{x};\beta) = \sum_{i=1}^{n} \psi(r_i(\mathbf{x};\beta)) r_i(\mathbf{x};\beta) = 0, \quad (4)$$

де $\psi(r_i(\mathbf{x};\beta)) = \rho(r_i(\mathbf{x};\beta)) / d\beta$ розглядається, як штрафна функція для точок, котрі вважаються викидами у вхідних та/або залежних змінних. Зазвичай систему легше розв'язати у продиференцьованому вигляді, як у (4).

У якості функції штрафів ψ використовується класична функція, запропонована Хубером:

$$\psi(r) = \begin{cases} r & \text{if } abs(r) < c \\ c \cdot sign(r) & \text{if } abs(r) \geq c \end{cases} \quad (5)$$

де c константа налаштування, яка регулює чутливість до викидів [8].

Вираз (4) розв'язується за допомогою ітеративних чисельних методів. Ми використовували ітеративний метод найменших зважених квадратів (Iteratively Reweighted Least Squares method – IRLS) [13]. В матричній формі рішення системи (4) записується таким чином:

$$\hat{\beta}^{\mathbf{i}+1} = (\mathbf{X}^{\mathbf{T}}\mathbf{w}^{-1}(\hat{\beta}^{i})\mathbf{X})^{-1}\ \mathbf{X}^{\mathbf{T}}\mathbf{w}^{-1}(\hat{\beta}^{\mathbf{i}})\mathbf{y} \tag{6}$$

$$\mathbf{w}^{-1} = min\{1,1/|\psi(\mathbf{r})|\} \tag{7}$$

де $\mathbf{w}^{-1}$ вектор ваги, що обчислюється базуючись на значеннях функції $\psi(\cdot)$ для залишків. Ітерації повторюються до сходження оцінки параметрів $\hat{\beta}$. Обмеження покладені на функцію $\rho(\cdot)$ забезпечують стійкість методу до викидів залежної змінної $\mathbf{y}$, проте він залишається чутливим до викидів у вхідних змінних $\mathbf{X}$.

Для покращення сходження оцінок параметрів $\hat{\beta}$ та підвищення точки перелому Йохай [17] запропонував удосконалену оцінку – MM-оцінку. Цей метод передбачає мінімізацію M-оцінки нормалізованих залишків. В цьому випадку критерій мінімізації матиме вигляд:

$$min_{\beta}\sum_{i=1}^{n}\rho(r_i/\hat{\sigma}_r). \tag{8}$$

Щоб отримати повну стійкість (до викидів у вхідних та залежних змінних) було запропоновано S- та GM-оцінки [9], які передбачала введення функції штрафів не тільки для залишків, а й для спостережень вхідних змінних $\hat{X}$. Тоді критерій мінімізації змінюється таким чином:

$$min_{\beta}\sum_{i=1}^{n}\psi(r_i(\mathbf{x};\hat{\mu};\hat{\sigma}_x))r_i(\mathbf{x};\hat{\mu};\hat{\sigma}_x)). \tag{9}$$

На жаль, вид функції $\psi(\cdot)$ запропонований авторами GM-оцінки вимагає надзвичайно великих обчислювальних витрат при розв'язанні системи. Для спрощення розв'язку, зазвичай, використовують припущення про незалежність змінних, що дозволяє переписати функцію $\psi(\cdot)$ у вигляді добутку двох функцій штрафів: окремо для вхідних змінних та залишків, таким чином, значно знижується обчислювальна складність задачі. При чисельному ітеративному розв'язку відповідно використовується вектори ваги змінних: $\mathbf{w} = \mathbf{w_r}\mathbf{w_x}$ де $\mathbf{w_r}$ функція штрафів залишків, як вказано в (7), а функція $\mathbf{w_x}$ функція штрафів викидів серед множини точок вхідних даних, яка визначається таким чином:

$$w_i = min\{1,\sqrt{C_\tau/d_m^2(\mathbf{x})}\}\ \forall i = \{1,...,n\}, \tag{10}$$

де C_τ константа налаштування, яка відповідає очікуваному рівню точок-викидів серед вхідних даних τ, а $d_m^2(x)$ – відстань Махаланобіса [11]. Ми

використовували C_τ рівну оберненій величині кумулятивної функції χ^2 розподілу з $m-1$ ступенем свободи для точки τ [6,12]. Дистанція Махаланобіса розраховується за наступною формулою:

$$d_m^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_x)^T \hat{\mathbf{S}}^{-1} (\mathbf{x} - \hat{\mu}_x) \tag{11}$$

де $\hat{\mu}_x$ та $\hat{S}$ – *стійкі* оцінки центру та коваріації множини точок $\mathbf{X}$. Тобто кожній точці ставиться у відповідність скалярна оцінка входження його у основний кластер даних, яка базується на нормованій стійкій оцінці відстані від центру кластеру – відстані Махаланобіса. В загальному випадку w_x повинні бути переобчисленні на кожній ітерації розв'язання системи рівнянь. Проте, для спрощення обчислювальної складності алгоритму, ми знаходили $\mathbf{w_x}$ тільки один раз. У якості стійкої оцінки центру даних використовувалась медіана $\mathbf{X}$, а стійка оцінка матриці коваріації обчислювалась за допомогою алгоритму попарних коваріацій Гнанадесікана-Кеттенрінга (Orthogonalized Gnanadesikan and Kettenring algorithm [12,5]).

У випадку викидів тільки в залежній змінній в IRLS та РПНМ критерій оцінки моделі – це зважена сума квадратів залишків:

$$RSS = \frac{1}{n-1} \sum_{i=1}^{n} \rho(r_i / \hat{\sigma}). \tag{12}$$

Проте в удосконаленому алгоритмі, необхідно також брати до уваги можливі викиди у вхідних змінних $\hat{\mathbf{X}}$ і тому, критерій оцінки моделей було змінено за аналогією до постановки задачі мінімізації:

$$RSS_w = \frac{1}{n-1} \sum_{i=1}^{n} \rho(r_i \cdot w_i^2 / \hat{\sigma}), \tag{13}$$

де $\mathbf{w_x} = \{w_i\}, i = \{1, ..., n\}$ обчислюється за виразом (10).

Аналогічним чином звичайний критерії Акаікі (з додатковим термом корекції на випадок малої вибірки експериментальних даних) у МГУА-частині алгоритму було замінено на його стійку модифікацію [1,3,14]:

$$\begin{aligned} AICr &= \frac{1}{n-k} \sum_{i=1}^{n} \rho(r_i \cdot w_i^2 / \hat{\sigma}) + \frac{n+k}{n-k-2} \\ &= \frac{n-1}{n-k} RSS_w + \frac{n+k}{n-k-2}, \end{aligned} \tag{14}$$

де k кількість термів в моделі, що оцінюється.

3. Удосконалена РПНМ

Удосконалена версія алгоритму базується на ітераційному алгоритмі МГУА ПНМ для генерації та вибору моделей [2] з повністю стійким до викидів "ядром", описаним вище.

Задача алгоритму знайти таку нелінійну модель $y = f(\mathbf{x};\beta_{\mathbf{o}}) + \xi$, де $f(\cdot)$ нелінійний мультиваріативний-поліном апріорі меншого ступеню ніж p_{max} та такий, що складається менше ніж з t_{max} термів. За умови наявності викидів у вхідних і залежній змінних ми можемо записати, що $\mathbf{X}' = \mathbf{X} + \xi_x$ та $\mathbf{y}' = \mathbf{y} + \xi$, де ξ_x та ξ некорельовані випадкові величини, які описують викиди у вхідних та залежних змінних відповідно.

Процедура пошуку моделей за допомогою удосконаленого методу РПНМ виглядає таким чином:

1. обчислити *стійкі оцінки* центру μ та коваріації S вхідної множини точок $\mathbf{X}'$;
2. ініціалізувати вектор ваги w_x, що базується на обчислених оцінках $\hat{\mu}$ та $\hat{S}$ для подальшого використання "ядром" РПНМ алгоритму, як вказано у (10);
3. ініціалізувати множину найкращих моделей $M_{best}^{'} = \varnothing$;
4. для кожної пари обмежень (t, p) на терми та ступені нелінійного поліному (тобто $\forall(t, p)$, $t \in \{1,...,t_{max}\}$ та $p = p_{max}$):
   - 4.1. виконати алгоритм "ядра" РПНМ поклавши множину початкових моделей M_{start} рівну $M_{best}^{'}$ з обмеженнями t і p на терми та ступінь моделі $M \in M_{best}^{'}$;
   - 4.2. отримати нову множину M_{best} з виходу алгоритму;
   - 4.3. оцінити кожну з моделей множини M_{best} у відповідності до стійкого критерію Акаіке (AICr, вираз (14));
   - 4.4. оновити множину $M_{best}^{'}$ тими моделями $M \in M_{best}$, що мають меншу оцінку критерію AICr та відрізняються за структурою від наявних у множині поточних моделей $M_{best}^{'}$;
5. вибрати модель $M \in M_{best}^{'}$ з найменшим значенням критерію AICr (або використати множину $M_{best}^{'}$ упорядковану за значенням AICr критерію);
6. для вибраної моделі (моделей) знайти *явну* форму функції залежності для подальшого аналізу за допомогою «оберненої» прогонки алгоритму.

Зауважте, що алгоритм наведений вище орієнтований на вибір оптимальної моделі, у той час, як фактичний пошук параметрів моделі та її структури здійснюється «ядром»" РПНМ (п. 4.1).

Важливою відмінністю даного алгоритму від МГУА є те, що замість класичного МНК для оцінки параметрів моделі використовується стійкий IRLS. Ця зміна «ядра» у свою чергу призвела до необхідності адаптації всіх критеріїв оцінки у РПНМ.

Ядро РПНМ алгоритму здійснює пошук структури моделей та оцінки їх параметрів виходячи з пари обмежень (t, p) на кількість термів та максимального ступеню моделі, вектору ваги вхідних змінних $\mathbf{w_x}$, а також множини «спадкових»" моделей M_{start} за наступною процедурою:

1. ініціалізувати робочі множини:
   1.1. множину найкращих моделей $M_{best} = M_{start}$;
   1.2. оцінки залежної змінної $\mathbf{Y_{best}}$ для найкращих моделей M_{best} на множині $\mathbf{X}$;
   1.3. ввести у робочу множину вхідних змінних обчислені оцінки залежних змінних $\mathbf{X_{all}} = [\mathbf{X}; \mathbf{Y_{best}}]$;
2. $\forall\{i, j, k\}$, де $i, j, k \in \{1, ..., |\mathbf{X_{all}}|\}$ виконати:
   2.1. побудувати модель $M_{ijk} = \alpha_1 x^i + \alpha_2 x^j x^k$, де функція-генератор $G(i, j, k) = \mathbf{x^i} + \mathbf{x^j x^k}$, $\mathbf{x^p}$ позначає p-й стовпець матриці $\mathbf{X}$;
   2.2. відкинути побудовану модель M_{ijk}, якщо вона не відповідає обмеженням (t, p) на кількість термів та ступінь;
   2.3. Обчислити коефіцієнти лінійної регресії α_1 та α_2 за допомогою IRLS методу та з використанням вектору $\mathbf{w_x}$ так, що $\alpha = \{\alpha_1; \alpha_2\}$ задовольняє лінійній моделі $y \approx [\mathbf{x^i}; \mathbf{x^j} \cdot \mathbf{x^k}]^T \cdot \alpha$;
   2.4. обчислити стійку функцію оцінки для моделі M_{ijk}:

$$RSS_w(r) = \sum_{t=1}^{n} \rho(r_t(M_{ijk}) \cdot w_{xt}^2 / \hat{\sigma})) = \rho(\mathbf{w_x}([\mathbf{x^i}; \mathbf{x^j} \cdot \mathbf{x^k}]^T \cdot \alpha - \mathbf{y})/\hat{\sigma});$$

   2.5. відхилити модель M_{ijk}, якщо

$$RSSw(r(M_{ijk})) \geq RSSw(r(M_i)), \forall M_i \in M_{best};$$

   2.6. Відхилити модель M_{ijk}, якщо

$$\begin{cases} \exists M_i \in M_{best} : M_{ijk} \text{ має однакову стуктуру з } \mathrm{M_i} \\ RSSw(r(M_{ijk})) > RSSw(r(M_i)); \end{cases}$$

   2.7. Включити модель M_{ijk} в множину M_{best}, а оцінку моделі $\mathbf{x_{ijk}} = \alpha_1 \mathbf{x^i} + \alpha_2 \mathbf{x^j x^k} = [\mathbf{x^i}; \mathbf{x^j} \cdot \mathbf{x^k}]^T \cdot \alpha$ в множину $\mathbf{X_{best}}$, замінивши модель з аналогічною структурою, якщо вона вже існує в множині M_{best};
3. залишити у множинах M_{best} та $\mathbf{X_{best}}$, тільки моделі з першими K найкращими значеннями критерію RSS_w, щоб запобігти надмірному росту множини M_{best} та зв'язаних з цим додаткових обчислень;
4. оновити множину вхідних змінних $\mathbf{X_{all}} = [\mathbf{X}; \mathbf{X_{best}}]$;
5. повторити кроки з 2 по 4 до сходження множини моделей.

3. Результати

Описаний алгоритм був перевірений на штучних даних. Моделі та відповідні вхідні та залежні змінні були згенеровані так, як це описано нижче.

Множина штучно згенерованих початкових даних $\mathbf{X}$ була створена у відповідності до гаусового розподілу $\eta(0,\sigma_x^2)$, у всіх тестах величина дисперсії становила $\sigma_x^2 = 10$. Для кожної точки множини $\mathbf{X}$ було отримано оцінку для згенерованої моделі M_{init}. Ця оцінка є вектором справжнього відклику моделі $\mathbf{y} = M_{init}(\mathbf{X})$. Вхідними даними для алгоритмів регресійного аналізу, що використовувались в тестах, були $\mathbf{X}$ та $\mathbf{y}$ початкових множин із вкрапленнями викидів – $\mathbf{X_{fit}}$ та $\mathbf{y_{fit}}$ відповідно. Множина вхідних змінних для регресії була створена за схемою реальний $\mathbf{X}$ плюс викиди:

$$\mathbf{X_{fit}} = \mathbf{X} + \eta(0,7\sigma_x^2) = \eta(0,\sigma_x^2) + \eta(0,7\sigma_x^2). \qquad (15)$$

У залежні змінні також додавався систематичний шум ξ:

$$\mathbf{y_{fit}} = \mathbf{y} + \eta(0,1) + \eta(0,3\sigma_y^2). \qquad (16)$$

Знайдені алгоритмами моделі M_{fit} перевірялися на даних вільних від викидів $\mathbf{X_{test}}$. Котрі, як і оригінальні дані, мали гаусівский розподіл, але з тричі більшою дисперсією $\mathbf{X_{test}} \sim \eta(0,3\sigma_x^2)$. Це було зроблено для виявлення “перенавчання”. Мірою якості моделей було середнє різниці квадратів відхилення моделей (RS) на точках тестової вибірки

$$RS = \frac{1}{|\mathbf{X_{test}}|} \cdot \sum_{\forall \mathbf{x_i} \in \mathbf{X_{test}}} (M_{fit}(\mathbf{x_i}) - M_{init}(\mathbf{x_i}))^2. \qquad (17)$$

На рис.1. наведено розподіл залишків отриманих за допомогою IRLS, РПНМ та УРПНМ для експериментів з квадратичними моделями.

Ширина стовпця - 300. Стовбці, що відповідають інтервалу $[-300;300)$ “обрізані” на відмітці 1500, справжні значення стовпців: 7749, 4831, та 18637 точок для IRLS, РПНМ, та УРПНМ відповідно. Всі експерименти мали множину вхідних даних $\mathbf{X_{init}}$ та $\mathbf{X_{test}}$ - множину тестових даних потужністю 100 точок кожна.

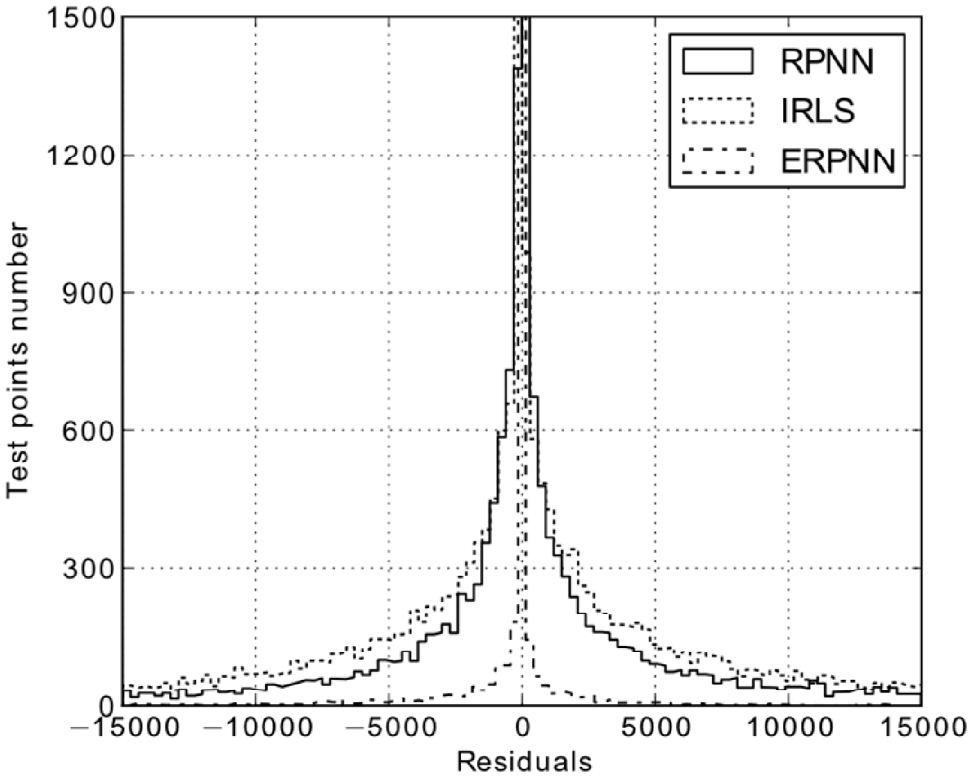

Рис. 1. Розподіл залишків отриманих за допомогою IRLS (суцільна лінія), РПНМ (штрихована лінія) та УРПНМ (штрих-крапка) у експерименті з квадратичними моделями

Оцінка алгоритму здійснювалась трьома типами тестів. Загальна якість прогнозування моделі порівнювалась з базовим алгоритмом РПНМ та алгоритмом IRLS (в останньому використовувався стійкий критерій якості, як описано в п.2). Якість структурного моделювання алгоритму протестовано за допомогою двох критеріїв вибору в еволюційній частині алгоритму. Досліджено стійкість алгоритму в залежності від кількості точок-викидів.

Якість регресії удосконаленого алгоритму порівняно з оригінальною РПНМ [2] та з стійкою до викидів модифікацію IRLS і наведено в табл.1.

Таблиця 1

Якість прогнозування при наявності 10 викидів у залежній змінній та 15 викидів у вхідних змінних

Метод	*Критерій RS*	*Моделі*
Лінійні	$mean \pm std$	$RS > 10^3$
RPNN	9336.239 ± 44850.23	11.0%
IRLS	424.638 ± 5278.51	1.0%
ERPNN	0.471 ± 0.43	0.0%

Метод	*Критерій RS*	*Моделі*
Квадратичні моделі	$mean/10^3 \pm std/10^3$	w. $RS > 10^3$
RPNN	$346\,164 \pm 451\,348$	99.5%
IRLS	$198\,490 \pm 453\,752$	78.5%
ERPNN	$2\,106 \pm 13\,404$	8.5%
Кубічні моделі	$mean/10^3 \pm std/10^3$	w. $RS > 10^3$
RPNN	$266\,120\,481 \pm 878\,222\,086$	99.5%
IRLS	$273\,272\,658 \pm 945\,446\,327$	85.5%
ERPNN	$231\,003\,849 \pm 876\,882\,759$	38.5%

Алгоритми тестувалися на лінійних, квадратичних та кубічних моделях (останні містили один терм 3-ого ступеня та решту термів 2-ого ступеню).

Для IRLS алгоритму, лінійного по суті, у випадку квадратичних та кубічних моделей масив вхідних даних $\mathbf{X}_{\mathbf{init}}$ був трансформований у $\mathbf{X}'_{\mathbf{init}}$ додаванням всіх можливих нелінійних комбінацій початкових змінних, що дають терми 2-ого та 3-ого ступеню відповідно. Всі згенеровані моделі складалися з 5 термів (4 терми створені довільним чином з 4 наявних змінних плюс терм-константа). Необхідно зауважити, що в деяких випадках моделі могли містити не всі вхідні змінні. На приклад: модель $f(\mathbf{x}) = x_3^2 + x_4^2 + x_3 x_4 + 1$ використовує тільки 2 змінні з усіх можливих, хоча складається з 4-х термів. УРПНМ та РПНМ алгоритми здійснювали пошук тільки серед моделей 2-ю ступеню та таких, що складалися з не більше, ніж з 6 термів. Дані для всіх експериментів містили 25% викидів (15 викидів серед вхідних змінних $\mathbf{x}$ та 10 викидів у залежній змінній $\mathbf{y}$). Кожен окремий експеримент повторювався 200 разів на нових моделях та даних.

Отримані результати узагальнені у Таблиці 1. Для експерименту з квадратичними моделями розподіл залишків між знайденими та справжніми моделями показано на рис. 1. “Хвости” розподілів, що лежать далі ніж 15000 від 0 – не показано. Кількість таких точок була 2111 (10.56%), 3821 (19.11%) та 34 (0.17%) для IRLS, РПНМ та УРПНМ відповідно. Значення “нульових” стовпців $[-300;300)$ на гістограмі обмежені 1500. Цей інтервал містить: 7749 (38.75%), 4831 (24.16%) та 18637 (93.18%) точок для IRLS, РПНМ та УРПНМ відповідно.

Значення RS критерію для кожного алгоритму та кожного з 200 експериментів з квадратичними моделями показано на Рис. 2.

Треба зауважити, що для IRLS алгоритму через трансформацію масив вхідних даних $\mathbf{X}_{\mathbf{init}}$ (розміром 100×5) перетворився на масив $\mathbf{X}_{\mathbf{fit}}$ розміром 100×15 та 100×35 для квадратичного та кубічного випадків відповідно.

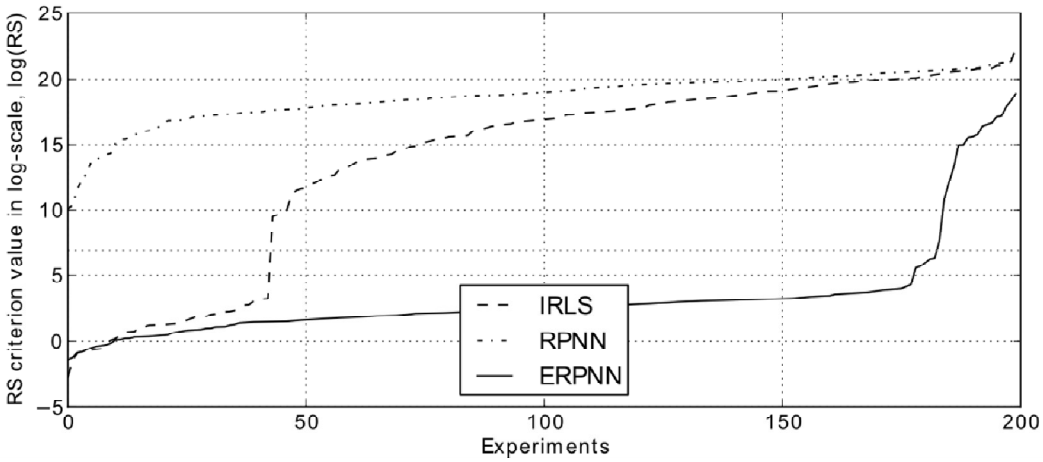

Рис. 2. Порівняння IRLS (штрихована лінія), РПНМ (штрих-крапка) та УРПНМ (суцільна лінія). Експерименти, відсортовані за значенням RS критерію, розташовані на осі абсцис, значення RS критерію на осі ординат

Вплив критеріїв AICr та стійкої суми квадратів залишків (RSS_w) на якість структурного моделювання був перевірений наступним тестом. Область пошуку моделей була розширена на такі, що містять до 12 термів (замість стандартних 6 і таким чином на 7 більше ніж реально необхідно). Результати експерименту узагальнено в таблицях 2 та 3.

Таблиця 2

Якість вибору моделей УРПНМ з критеріями AICr та RSS_w при максимальній допустимій довжині моделей 6 (Т6) або 12 (Т12) термів

Критерій	*Кіл-сть моделей*	*Середня довжина моделі*
(кіл-сть термів)	з $RS > 10^3$	(термів в моделі)
AIC_r (Т6)	8.5%	5.11
AIC_r (Т12)	7.5%	5.74
RSS_w (Т12)	7.5%	10.56

Таблиця 3

Якість прогнозування моделей УРПНМ з критеріями AICr та RSS_w при максимальній допустимій довжині моделей 6 (6Т) або 12 (12Т) термів

Критерій	*RS, найкращі 80%*	*RS, гірші 20%*
(кіл-сть термів)	$mean \pm std$	$mean/10^3 \pm std/10^3$
AIC_r (Т6)	10.80 ± 8.22	$10\,532 \pm 28\,453\,546$
AIC_r (Т12)	12.31 ± 8.89	$6\,536 \pm 16\,842\,335$
RSS_w (Т12)	15.75 ± 9.95	$6\,536 \pm 16\,842\,333$

Зверніть увагу на покращення значення RS критерію при пошуку більш

довгих моделей, що в решті-решт дозволило отримати на 2 моделі більше (на 1% більше успішних експериментів) з низьким (менше 1000) значенням цього критерію.

Чутливість алгоритму тестувалася на даних, що містили $N_O = \{5,10,20\}$ викидів у залежній змінній та $N_L = \{0,5,10,15,20,25\}$ викидів у вхідних змінних. Найбільш показові результати наведено у таблиці 4.

Таблиця 4

Чутливість алгоритму УРПНМ до кількості викидів серед вхідних даних

Кількість викидів		*Значення критерію RS*	*К-сть моделей*
в X	в Y	$mean/10^3 \pm std/10^3$	w. $RS > 10^3$
0	10	1.92 ± 19.26	1.0%
5	10	193.1 ± 931.9	6.5%
10	10	$1128 \pm 7\,503$	8.5%
15	5	$1\,227 \pm 8\,922$	5.5%
15	10	$2\,106 \pm 13\,404$	8.5%
15	20	$7\,795 \pm 33\,049$	15.5%
20	10	$15\,051 \pm 75\,973$	17.5%
25	10	$67\,459 \pm 177\,511$	38.5%

Залежність кількості моделей з високим середнім значення квадратів залишків (RS) від кількості викидів N_O та N_L серед вхідних даних наведено на рис. 3.

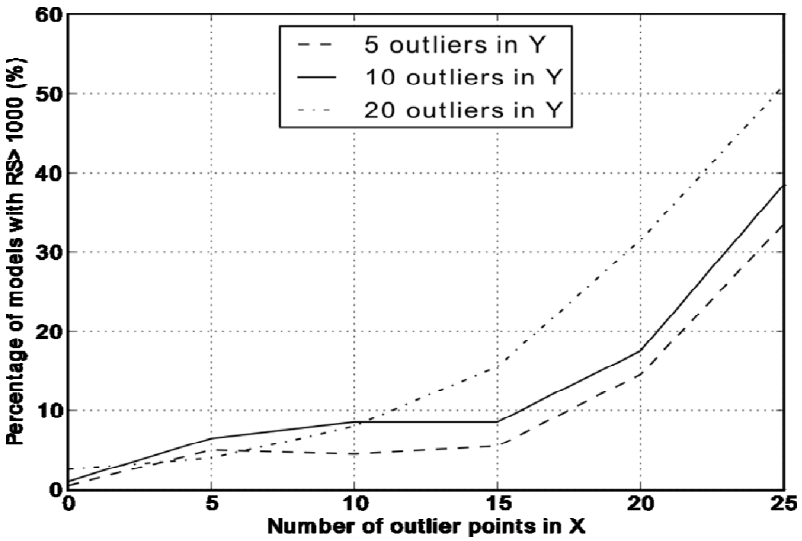

Рис. 3. Відсоток моделей з високим значенням RS критерію ($RS > 10^3$), знайдених за допомогою УРПНМ в залежності від кількості точок-викидів (від 0 до 25) у вхідних змінних (ось x) та у залежній змінній: 5 (штрихована лінія), 10 (суцільна лінія) та 20 (штрих-крапка)

Зміна RS критерію у експериментах для найбільш цікавих початкових умов відображено на рис. 4.

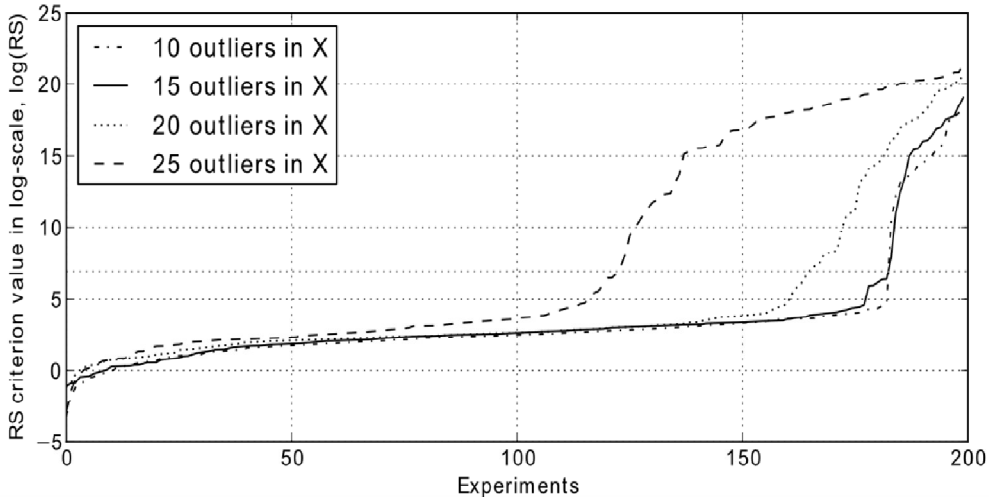

Рис. 4. Стійкість УРПНМ до точок-викидів у множині вхідних даних (експерименти, відсортовані за RS критерієм, розташовані на осі абсцис)

Треба зауважити, що у всіх випадках алгоритм залишався налаштований на 25% викидів (тобто значення константи налаштування C_τ завжди було однаковим).

На рис. 4. значення критерію RS на осі ординат наведено у логарифмічній шкалі. 10 викидів були присутні у залежній змінній і 10 (штрих-крапка), 15 (суцільна лінія), 20 (крапки) або 25 (штрихована лінія) викидів були присутні у вхідних змінних.

4. Висновки

З таблиці 1 видно, що оригінальний РПНМ не може ефективно працювати при наявності цього типу викидів. У квадратичному випадку РПНМ правильно віднайшов лише 1 модель, у той час як удосконалена версія не знайшла 17 моделей (з двохсот). Порівняння з стійким IRLS показало, що УРПНМ трохи краща на лінійних моделях і значно краща на моделях більш високого ступеню. Проте для IRLS було застосовано метод “розширення” базових змінних, що призведе до набагато гірше обумовлених вхідних матриць, особливо при наявності великої кількості змінних, значного збільшення обчислень та збільшення впливу викидів, які будуть піднесені у ступінь моделі. УРПНМ у свою чергу працює з обмеженою кількістю змінних на кожному окремому кроці і може більш успішно боротися з викидами.

Як можна бачити з рисунків 2 та 4, моделі, побудовані алгоритмом, добре апроксимують дані до певної межі, але в певний момент спостерігається різкий “скачок” RS-критерію до значень 10^6 та вище. Це свідчить про те, що алгоритм не знаходить один або декілька квадратичних термів справжньої моделі і, натомість, заміняє їх іншими *невірними* квадратичними термами. Враховуючи “квадрати” моделі та “квадрат” критерію оцінки якості, це призводить до появи надвисоких значень RS-критерію. В інших випадках, модель може мати

неточно оцінені параметри або надлишкові терми з коефіцієнтами близькими до нуля, проте в цих випадках значення критерію будуть відносно низькими – до 10^3.

Необхідно зауважити, що результати для лінійних та кубічних моделей (наведених у таблиці 1) не можуть бути напряму співвіднесені з відповідними даними квадратичних моделей через значно іншу динаміку зміни RS оцінки: у першому випадку критерій зростає значно повільніше, а у другому – значно швидше.

Різницю між трьома методами найкраще видно на рис. 1. З рисунку видно, що для всіх алгоритмів більшість точок згрупована навколо нуля. Проте удосконалений алгоритм має 93% точок у стовпцях сусідніх до нуля, проти 39% для IRLS та 24% для РПНМ. Що свідчить про набагато менший рівень навіть дрібних похибок знайденої моделі. Як наслідок, розподіли залишків інших алгоритмів мають набагато масивніші “хвости”, тобто, більшу кількість набагато зовсім не вірних прогнозів.

З таблиці 2 видно, що дозвіл на пошук двічі довших моделей дозволяє зменшити кількість моделей з великими значеннями RS критерію на 11%, у порівнянні із варіантом пошуку серед моделей в 6 термів. Окрім того критерій Акаікі дозволяє знаходити майже у двічі коротші моделі, ніж RSS_w – у середньому 5.74 термів, проти 10.56 – що дозволяє більш ефективно запобігти з перенавчанням. IRLS алгоритм, у свою чергу, взагалі не має засобів вибору оптимальної структури моделі: він завжди знаходить моделі, що складаються з усіх можливих термів.

Не зважаючи на успішні отримані результати тестування алгоритму, існує ще достатньо засобів для його поліпшення. Якість виявлення викидів, може бути покращена за допомогою заміни попарного-алгоритму стійкої оцінки коваріацій на такий, що аналізує весь масив даних загалом, а не тільки попарні комбінації змінних, як це робить ОГК-алгоритм. З іншого боку такі алгоритми набагато більш ресурсоємкі, тому можливо, що в залежності від поставлених задач, їх використання буде небажаним. Та окрім того, ефективність протидії викидам у термах високих ступенів можна покращити більш ретельно дослідивши вибір границі чутливості алгоритму.

Попри деякі наявні недоліки запропонований алгоритм удосконаленої робастної поліноміальної нейтронної мережі зберігає високу точність вибору синтезу моделей характерну для його попередника та демонструє стійкість до викидів, як у вхідних, так і у залежних змінних.

Подяка

Автори приносять подяку за підтримку проекту EC FP6 грант #034632 (PERPLEXUS), а також фонду ICOBI of Foundation “Nanoscience at the limits of Nanoelectronics”.

Література

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] *Aksenova, T., Volkovich, V., Villa A.E.P.* Robust Structural Modeling and Outlier Detection with GMDH-Type Polynomial Neural Networks, LNCS, 3697, 881-886, 2005.

[3] K.P. Burnham and D.R. Anderson. Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods Research*, 33(2):261–304, 2004.

[4] D. L. Donoho and P. J. Huber. The notion of breakdown point. *Festschr. for Erich L. Lehmann*, 157–184, 1983.

[5] R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124, 1972.

[6] Ali S. Hadi, A. H. M. Rahmatullah Imon, and M. Werner. Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70, 2009.

[7] F. R. Hampel. A general qualitative definition of robustness. *Ann. Math. Stat.*, 42:1887–1896, 1971.

[8] *Хьюбер П.* Робастность в статистике. - М.: Мир, 1984. - 304 с.

[9] Hendrik P. Lopuhaa. Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics*, 27(5):1638–1665, 1999.

[10] *Ивахненко А.Г., Степашко В.С.* Помехоустойчивость моделирования. — Киев: «Наук. думка», 1985.

[11] P.C. Mahalanobis. On the generalized distance in statistics. *Natl. Inst. Science*, 49–55, 1936.

[12] Ricardo A Maronna and Ruben H Zamar. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317, 2002.

[13] Dianne P. O'Leary. Robust regression computation computation using iteratively reweighted least squares. *SIAM J. Matrix Anal. Appl.*, 11(3):466–480, 1990.

[14] E. Ronchetti. Robustness aspects of model choice. *Statistica Sinica*, 7:327–338, 1997.

[15] P. Rousseeuw and V. Yohai. Robust regression by means of s-estimators. *Robust and nonlinear time series analysis*, SMC-1(26):256–272, 1983.

[16] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[17] V. Yohai. High breakdown-point and high efficiency robust estimates for regression. *Ann. Stat.*, 15:642–656, 1987.

[18] *Ивахненко А.Г., Юрачковский Ю.П.* Моделирование сложных систем по экспериментальным данным. — М.: «Радио и связь», 1987.