УДК 681.513

ПОРІВНЯЛЬНИЙ АНАЛІЗ ТИПОВИХ СТРУКТУРНИХ ЕЛЕМЕНТІВ МЕТОДІВ ПОБУДОВИ МОДЕЛЕЙ

С.М.Єфіменко
Міжнародний науково-навчальний центр інформаційних технологій та систем (МННЦ ІТС) НАН та МОН України,
syefim@ukr.net

В роботі виконано порівняльний аналіз поширених методів побудови моделей за даними спостережень. Виділено чотири спільні для кожного методу моделювання компоненти, що дає можливість порівнювати функціональні особливості цих методів.
Ключові слова: структурно-параметрична ідентифікація, регресійний аналіз, індуктивне моделювання, МГУА.

The comparative analysis of the widespread modeling methods from data observed is executed in the paper. Four common components for every modeling method are extracted that enables to compare the functional features of these methods.
Keywords: structural and parametric identification, regression analysis, inductive modeling, GMDH.

В работе выполнен сравнительный анализ распространенных методов построения моделей по данным наблюдений. Выделены четыре общие для каждого метода моделирования компоненты, что дает возможность сравнивать функциональные особенности этих методов.
Ключевые слова: структурно-параметрическая идентификация, регрессионный анализ, индуктивное моделирование.

Важливим етапом розробки систем управління є ідентифікація об'єкта управління, природного або технологічного процесу. Ця задача передбачає побудову моделі за наслідками експериментальних спостережень за їх функціонуванням.

Підходи до вибору кращої структури моделі. Серед існуючих методів ідентифікації та регресійного аналізу можна умовно виділити три групи методів, кожна з яких характеризує один з трьох основних підходів до вибору кращої структури, які в [1] названі так: ***класичний***, ***сучасний*** і ***нетрадиційний***.

Класичний підхід ґрунтується на апараті перевірки статистичних гіпотез і представлений в роботах С.А.Айвазяна [2], М.Г.Загоруйка [3], В.Камінскаса [4], Л.Льюнга [5], Н.С.Райбмана [6], Я.З.Ципкіна [7], П.Ейкхофа [8] та інших.

Сучасний підхід передбачає дотримання принципу компромісу між точністю і складністю (числом оцінюваних параметрів) моделі; найбільш відомими тут є напрями, розроблені Х.Акаіке [9] (мінімізація ентропії моделі), В.Н.Вапніком [10] (гарантоване оцінювання величини середнього ризику) і К.Л.Маллоузом [11] (незміщене оцінювання теоретичної помилки моделі), істотні результати містяться в роботах Т.Амемія [12], П.Янга [13] та інших.

Нетрадиційний підхід „перехресного обґрунтування" заснований на розділенні вибірки даних для отримання додаткової інформації; найбільший внесок належить тут О.Г.Івахненку [14] (метод групового урахування

аргументів (МГУА)) і Дж.Тьюки [15] (метод "джекнайф"). Останнім часом інтерес до методів, які використовують розділення вибірки, особливо до МГУА, значно зріс, передусім через їх ефективність при розв'язанні широкого спектру задач ідентифікації, прогнозування, розпізнавання.

Аналіз методів структурно-параметричної ідентифікації

Проаналізуємо деякі найчастіше використовувані методи структурно-параметричної ідентифікації з метою пошуку їхніх спільних етапів.

Айвазян виділяє таку послідовність дій при моделюванні за даними спостережень [2]:

- генерація підмножини змінних;
- обчислення критерію якості рівняння регресії для цієї підмножини;
- перевірка умови закінчення моделювання.

Відновлення невідомої залежності вихідної змінної від вхідних розпочинається з вибору класу функцій, в якому буде виконуватися пошук найкращої апроксимації. Цей етап автор вважає ключовим та рекомендує максимально використовувати апріорну інформацію про залежність, що аналізується. У якості допустимих він розглядає:

- лінійні функції

$$f(X;\Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)},$$

де X – матриця незалежних змінних, Θ — вектор параметрів розмірності $(p+1)\times 1$;

- степеневі

$$f(X;\Theta) = \theta_0 (x^{(1)})^{\theta_1} (x^{(2)})^{\theta_2} ... (x^{(p)})^{\theta_p};$$

- поліноміальні

$$f(X;\Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)} +$$

$$+ \sum_{k_1=1}^{p} \sum_{k_2=1}^{p} \theta_{k_1 k_2} x^{(k_1)} x^{(k_2)} + ... + \sum_{k_1=1}^{p} ... \sum_{k_m=1}^{p} \theta_{k_1 k_2 ... k_m} x^{(k_1)} x^{(k_2)} .. x^{(k_m)}$$

та інші. Останній вираз часто називають поліномом Колмогорова-Габора.

Наступним кроком є пошук невідомого значення параметра, що буде найкращим у розумінні заданого критерію адекватності. Тут відзначається, що на практиці найчастіше використовуються такі методи оцінювання:

- метод найменших квадратів;
- метод найменших модулів;
- байєсівське оцінювання;
- робастні методи.

При відборі кращих моделей Айвазян вказує на необхідність досягнення компромісу між складністю регресійної моделі та її точністю, та використовує для мінімізації критерію адекватності ємнісні характеристики класу базисних функцій.

Також для відбору кращих моделей автор пропонує використовувати статистичні властивості оцінок складності моделей.

Льюнг виділяє три основні складові методів моделювання за даними спостережень [5]:

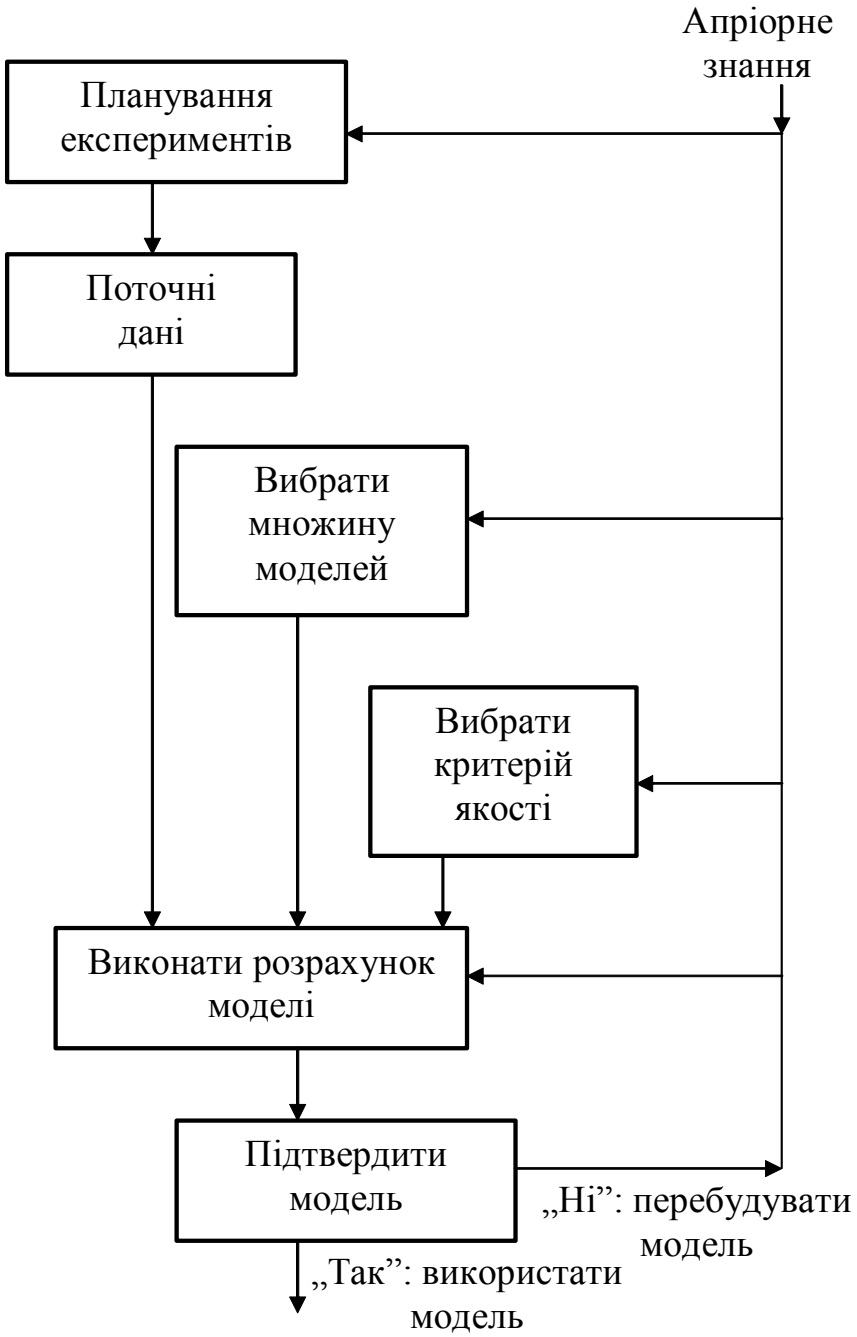

Рисунок 1. Етапи моделювання за Льюнгом

– дані. У випадку, коли користувач має можливість виділяти набір вхідних і вихідних змінних та спосіб їх отримання, він повинен вибирати лише найбільш інформативні дані;

– множина моделей. Льюнг вважає вибір множини (або класу) моделей, серед яких користувач буде виконувати пошук найкращої, найскладнішим та найважливішим етапом процесу моделювання. На цьому етапі формуються структури моделей (тобто моделі з невизначеними параметрами);

– визначення за даними спостережень найкращої моделі. Цей етап і становить власне процес ідентифікації.

Після отримання найкращої моделі виконується перевірка, чи задовільно модель описує поведінку системи.

Відповідно автор виділяє такі етапи ідентифікації:

1. Вибір даних.
2. Вибір множини моделей.
3. Вибір найкращої з цієї множини моделей.

На рисунку 1 представлено послідовність розв'язання задачі моделювання відповідно до описаних етапів.

Суть методу Вапніка **мінімізації емпіричного ризику** [10] полягає у тому, щоб за заданою навчальною вибіркою $X^m = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, побудувати функціональну залежність $F(x, \theta)$, яка мінімізує функціонал емпіричного ризику

$$I_e(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - F(x_i, \theta))^2 ,$$

де n – кількість точок.

При побудові регресійних моделей автор використовує передусім клас лінійних функцій

$$f(x, \theta) = \sum_{i=1}^{m} \theta_i x_i + \theta_0 ,$$

де m – кількість регресорів.

Для генерації моделей використовуються структури

$$S_1 \subset S_2 \subset ... \subset S_m .$$

Кожен елемент S_i, $i = \overline{1, m}$ є лінійною комбінацією аргументів множини $x_1, x_2, ..., x_m$, структура якої задається апріорно.

Автор визначає такі етапи структурної мінімізації ризику для класу лінійних функцій:

1. Визначити аргументи, які входять до елементів S.
2. Для кожного елемента структури S знайти функцію, що мінімізує емпіричний ризик.
3. Визначити такий елемент структури S, для якого величина сумарного ризику

$$I_{\sum}(\theta) = \frac{1}{k}\sum_{i=1}^{k}(y_i - f(x_i, \theta))^2 \qquad (1)$$

мінімальна. В (1) використовується окрема частина (автор називає її робочою) вибірки з k векторів $x_1, x_2, ..., x_k$.

Пошук найкращого рівняння регресії [16]. В **методі виключення** розгляд починається з моделі максимальної складності, коли будується залежність вихідної величини від усіх можливих вхідних. Наступним кроком обчислюється значення частинного F-критерію для кожної змінної:

$$F_{част\_x_i} = \frac{R^2_{yx_1x_2...x_i...x_m} - R^2_{yx_1x_2...x_{i-1}x_{i+1}...x_m}}{1 - R^2_{yx_1...x_i...x_m}}(n-m-1), \qquad (2)$$

де

$$R^2_{yx_1...x_i...x_m} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

– коефіцієнт множинної детермінації, $\hat{y}$ – вихід моделі, $\bar{y}$ – середнє значення вектора y. Якщо найменше значення критерію менше за обране попередньо деяке критичне значення, то відповідна змінна виключається з моделі. Модель перераховується і знову обчислюються значення частинного F-критерію з порівнянням з критичним значенням і т.д. Процедура моделювання закінчується тоді, коли отримане найменше значення критерію буде більшим за критичне.

Суть **методу включення** полягає у тому, що найкраща модель шукається шляхом послідовного додавання аргументів, як з множини початкових $x_i, \quad i = \overline{1, m}$, так і деяких функцій від них $f_j(x_i), \quad j = 1, 2, ...$ При додаванні аргументів у модель також використовується значення критерію (2).

За допомогою методу можлива побудова як лінійних, так і нелінійних за аргументами (але в будь-якому разі лінійних за параметрами) моделей $y = \sum_i \theta_i f_i(x)$.

Метод включення-виключення є поєднанням двох описаних методів та після етапу включення містить перевірку аргументів на значущість з виключенням незначущих аргументів. Цей метод також називається методом *покрокової регресії*.

Метод еволюційної ідентифікації [17]. Суть цього методу, що є модифікацією алгоритму включення полягає у використанні генератора випадкових чисел для вибору аргумента, який додається в модель. Перед початком роботи алгоритму задаються ваги w_j як для кожного аргумента x_j,

так і для заданих функціональних перетворень (використовуються різні класи базисних функцій $f(x_j)$). Аргументи та функції вибираються за допомогою генератора випадкових чисел з імовірністю $w_j / \sum_{j=1}^{m} w_j$. Потім у залежності від значущості випадково вибраних аргумента та функціонального перетворення ці ваги коригуються. Процес включення-виключення аргументів також триває до моменту, коли критерій якості перестає зменшуватись.

Комбінаторний алгоритм МГУА. З часу виходу в 1968 році першої статті, що започаткувала метод групового урахування аргументів (МГУА) [18], розроблено значну кількість різноманітних алгоритмів.

Серед них є такі, що орієнтовані на різні класи задач моделювання – поліноміальні алгоритми для статичних об'єктів, гармонічні та авторегресійні для часових рядів (коливальних процесів), алгоритми побудови різницевих рівнянь для динамічних об'єктів і процесів.

Існують три основних класи алгоритмів МГУА, що відрізняються способом організації перебору структур [19]:

– комбінаторні (однорядні);

– комбінаторно-селекційні (багатоетапні);

– селекційні (багаторядні).

Далі описується комбінаторний алгоритм МГУА.

Алгоритм містить такі основні блоки: перетворення даних згідно з вибраним класом структур моделей, лінійних за параметрами; формування моделей різної складності та оцінювання параметрів за МНК; обчислення значень зовнішніх критеріїв якості та відбір кращих моделей; оцінка якості отриманих моделей на екзаменаційній частині вибірки.

Розглянемо принципи організації обчислень в комбінаторному генераторі моделей, де концентруються основні обчислювальні витрати.

Основними операціями в блоці генерації моделей є: формування структури моделі (системи умовних рівнянь) та розв'язання відповідної нормальної системи рівнянь.

У випадку лінійного об'єкту з m входами в процесі повного перебору порівнюються моделі виду $\hat{y}_\nu = X_\nu \hat{\theta}_\nu, \nu = 1,...,2^m - 1,$, де десятковому числу ν ставиться у відповідність двійкове число d_ν – його ще називають двійковим структурним вектором $d_\nu = (d_1 ... d_m)^T$, одиничні елементи якого вказують на включення в модель регресорів з відповідними номерами.

Зміну станів вектора d можна організувати різними способами: усіма можливими варіантами розміщення у векторі однієї, двох, і т.д. до m одиниць;

за принципом двійкового лічильника, в останній розряд котрого додається одиниця, тощо.

Суть **методу оцінювання за мінімумом інформаційного критерію Акаіке (*MAICE*)** [8] полягає в тому, що для вибору кращої моделі використовується інформаційний критерій Акаіке

$$AIC(s) = -2\ln f_m(x, \hat{\theta}(s)) + 2s,$$

де $f_m(x, \hat{\theta})$ – густина розподілу випадкового вектора x, s – складність моделі (число оцінюваних параметрів), $s = \overline{1, m}$.

Перебір серед моделей-кандидатів при використанні методу Акаіке виконується у класі вкладених структур, коли спочатку розглядається модель, що містить один перший аргумент, потім перших два і т.д. до повної моделі складності m. Як правило, при цьому мова йде про різницеві моделі динаміки.

Оцінки параметрів при використанні методу Акаіке отримуються як розв'язок рівняння Юла-Уолкера, яке є аналогом нормального рівняння для оцінювання параметрів регресії.

Висновки

З представленого короткого огляду методів моделювання за статистичними даними, який не претендує на вичерпну повноту, можна зробити висновок про наявність основних спільних для них етапів (які формують основні компоненти методів побудови моделей):

– вибір класу базисних функцій та відповідне перетворення початкових даних;

– генерація різних структур моделей у вибраному класі;

– оцінювання параметрів кожної згенерованої структури;

– мінімізація значення заданого критерію селекції та вибір кращої моделі.

Очевидно, що деякі етапи в конкретному методі моделювання можуть бути відсутніми, залежно від наявної апріорної інформації. Так, скажімо, якщо істинна структура моделі (набір аргументів, що реально формують вектор виходу) відома, то відпадає необхідність в етапах генерації різних структур та мінімізації значення критерію селекції.

Література

1. Степашко В.С. Автоматизована структурна ідентифікація прогнозуючих моделей складних об’єктів / Автореферат дисертації на здобуття наукового ступеня доктора технічних наук. – Київ: Ін-т кібернетики НАН України, 1994. – 46 с.

2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. - М.: Финансы и статистика, 1985. - 487с.

3. Загоруйко Н.Г. Эмпирические предсказание. - Новосибирск: Наука, 1979. - 124с.

4. Каминскас В.А. Идентификация динамических систем по дискретным наблюдениям. Часть 1. Основы стохастических методов оценивания параметров линейных систем. - Вильнюс: Мокслас, 1982. - 245с.

5. Льюнг Л. Идентификация систем. Теория для пользователя. - М.: Наука, 1991.- 432с.

6. Райбман Н.С. Что такое идентификация? – М.: Наука, 1970. – 345 с.

7. Цыпкин Я.З. Информационная теория идентификации.- М.: Наука, 1995.- 336с.

8. Современные методы идентификации систем / Под ред.П.Эйкхоффа.- М.: Мир, 1983.- 327с.

9. Акаике Х. Развитие статистических методов / В кн. Современные методы идентификации систем.- М.: Мир, 1983.- С.148-176.

10. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. - М.: Наука, 1979.- 447с.

11. Mallows C.L. Some comments on Cp // Technometrics.- 1973.- v.15.- P.661-667.

12. Amemiya, T. Advanced Econometrics, Cambridge, MA: Harvard University Press, 1985. – 507 p.

13. Young P.S. Data-Based Mechanistic Modeling of Engineering Systems // Journal of Vibration and Control. – 1998. - v. 4, No. 1. - P.5-28.

14. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. - Киев: Наук.думка, 1982.- 295с.

15. Тьюки Дж. Анализ результатов наблюдений: Разведочный анализ.-М.: Мир, 1981.- 693с.

16. Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 456 с.

17. Качала В.В., Чагоровская О.А. Алгоритм эволюционной идентификации сложных объектов // Распределенные информационно-управляющие системы. Саратов, 1988. С. 159.

18. Ивахненко А.Г. Метод групового урахування аргументів – конкурент методу стохастичої апроксимації // Автоматика. 1968. № 3. С. 57-72.

19. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – Киев: Наукова думка, 1985. – 216 с.