Принципи зовнішнього доповнення у методі групового урахування аргументів (МГУА) та в методі граничних спрощень (МГС)

Васильєв В.І., Ланге Т.І., Кобець Н.М.

В. І. Васильєв: Міжнародний науково-навчальний центр ЮНЕСКО/МПІ інформаційних технологій і систем НАН і МО України (МННЦ ІТІС), Україна, 252022, Київ, пр. Академіка Глушкова 40, т. (044)2664187, факс (044)2661570
Н. М. Кобець: Національний технічний університет України "Київський політехнічний інститут", пр. Перемоги 37,т. (044)4411892
Т. І. Ланге: Fachhochschule Merseburg, Fachbereich Elektrotechnik, FH Merseburg,Geusaer Straβe, D-06217 Merseburg, Deutschland, т.(03461)462528, факс (03461)462919

Анотація - Розглядаються та аналізуються два індуктивних методи відновлення функції - метод групового урахування аргументів (МГУА) і метод граничних спрощень (МГС). Вказано їхні переваги перед іншими методами та можливість їхнього поєднання з метою поліпшення їхніх екстраполяційних властивостей.

ВСТУП

Всі індуктивні методи, які засновані на неповній індукції, а до них безумовно відносять МГУА [1] та МГС[2], відрізняються тим, що в них загальні висновки робляться на підставі часткових фактів, а це може призвести як до вірних, так і до помилкових рішень. Причина такої невизначеності полягає в тому, що окремі факти, на яких засновуються загальні висновки, не завжди достатньо добре характеризують досліджуване явище. Разом з цим отримані загальні висновки повинні пояснювати не тільки окремі факти, але і все досліджуване явище в цілому, тобто ці загальні висновки не повинні змінюватися при практично нескінченному збільшенні кількості експериментів. Тому якість індуктивного висновку визначається не тільки і не стільки поясненням окремих, отриманих у процесі експериментів фактів, скільки екстраполяційними спроможностями цих висновків, їхньою здатністю до експансії в область явища, що не охоплена експериментами.

Кожний раз, коли модель вибирається із надто складного класу, здебільшого не вистачає емпіричних даних для її однозначного пояснення, тобто модель виявляється складніше того, що несуть у собі накопичені факти, і ці факти просто не в змозі відтворити таку модель.

У задачах відновлення багатовимірних залежностей будь-яке спрощення моделі приводить до згладжування тих або інших деталей. Навпаки, надмірне ускладнення моделі без урахування обсягу експериментальних даних призводить до непомірної свободи поводження апроксимуючої функції в області, що не була охоплена експериментом, у той час як більш прості моделі в цій області поводяться більш "обережно".

Далі будуть розглядатися індуктивні методи стосовно до вирішення задачі відновлення функцій, суть якої у виявленні та моделюванні деякої закономірності, що зв'язує основні характеристики досліджуваного явища у функціональну залежність наступного виду

$$y = F(X) \quad . \qquad (1)$$

Зробити це потрібно на підставі обмеженої кількості експериментів, тому тут постає центральна проблема усіх індуктивних методів, яка полягає в правильному співвідношенні складності моделі, що синтезується, з кількістю емпіричних даних.

Серед деяких методів, у яких особлива увага приділяється пошуку такого співвідношення, виділяються метод групового урахування аргументів (МГУА) [1] та метод граничних спрощень (МГС) [2], [3]. Тому тут будуть розглянуті ці два методи з погляду на можливість їх взаємодоповнення.

Особливості методу групового урахування аргументів (МГУА) [1]

У основу методу покладено декілька основних принципів: неостаточність проміжних рішень, зовнішнього доповнення, самовідбору проміжних рішень, єдиності остаточного рішення. Одночасно з оптимізацією коефіцієнтів апроксимуючого полінома відбувається й оптимізація складності, що дозволяє не піклуватися про вибір структури цього полінома. Спеціальний критерій (зовнішнє доповнення) є перешкодою переускладненню моделі. Цей критерій відкидає усе, що не в змозі відтворити коротка навчальна вибірка (відкидання

«сміття»). В результаті залишається тільки те, що може бути надійно «підтверджено» конкретною вибіркою, а відкидаються «фантазії» моделі.

Задача розв'язується у декілька етапів. Спочатку для всіх незалежних змінних (аргументів) створюються усі можливі групи комбінацій. У практичних алгоритмах до кожної комбінації належить лише два аргументи (попарне урахування аргументів). Для кожної пари (групи) створюється частковий опис, тобто деяке просте рівняння не вище другого порядку, аргументами якого є обрана пара. Вид часткового опису однаковий для всіх груп протягом усього процесу навчання. А всю вибірку поділяють на дві частини: навчаючу та перевірочну. Таким чином утворюється зовнішнє доповнення (перевірочна вибірка), що відіграє роль сита, яке відсіває усе надмірно складне, що не має права на існування при таких обмеженнях інформації. Коефіцієнти часткових описів визначаються за даними навчаючої вибірки. В результаті утворюються множини вирішень, оскільки часткове рівняння кожної пари розглядається як деяка спрощена модель функції, яку відновлюють.

У якості апроксимуючого полінома дуже часто використовується поліном Колмогорова-Габора [1]:

$$y = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \alpha_i x_i x_j x_k + \ldots \quad (2)$$

тому що за допомогою такого полінома можна досягнути достатньо точної апроксимації будь-якої диференційованої функції F.

Цю складну залежність заміняємо множиною простих функцій:

$$y_1=f(x_1,x_2);\ y_2=f(x_1,x_3);\ y_s=f(x_{m-1},x_m), \quad (3)$$

де $s = C_m^2$, причому функція f усюди однакова.

Дуже часто в якості функції f вибираються прості залежності.

$$y(x_i,x_j) = a_0+a_1x_i+a_2x_j+a_3x_ix_j \quad (4)$$

У другому ряду розглядаються моделі виду

$$z_1= f(y_1,y_2);\ z_2= f(y_1,y_3);\ldots;\ z_s= f(y_{s-1},y_s) \quad (5)$$

Складність полінома зростає від ряду до ряду. Так, наприклад, у другому ряду будуть отримані поліноми, що містять нелінійні члени виду x_1^2, $(x_1\ x_3)$, $(x_2\ x_3)$, $(x_1^2\ x_3)$, $(x_1^2 x_2\ x_3)$ і т.д., коефіцієнти цих поліномів можуть бути визначені по тих же точках навчальної послідовності і не потребують додаткової інформації, хоча їхня складність увесь час зростає. При цьому кількість коефіцієнтів, що визначаються, значно перевищує кількість точок навчальної послідовності. Якби не було зовнішнього доповнення, тобто перевірочної вибірки, то алгоритми МГУА, завдяки складності синтезованих поліномів, могли б абсолютно апроксимувати функцію (1) у всіх точках навчальної вибірки, але при цьому не залишилося б ніяких гарантій задовільного поводження функції, яку відновлюють, на нових точках.

Зупинка алгоритму відбувається відразу ж по досягненні єдиного мінімуму відхилень, отриманих на перевірочній вибірці. Тим самим вибирається модель оптимальної складності, що встановлює компроміс між складністю й обсягом інформації, що використовується при синтезі моделі. Як тільки складність перевищить можливості навчальної інформації, процес синтезу моделі припиняється.

Особливості методу граничних спрощень (МГС)

Будь-яка індукція містить у собі елементи дедукції і з нею пов'язана, та, навпаки, дедукція пов'язана з індукцією. Такий діалектичний взаємозв'язок індукції і дедукції наводить на думку про те, що будь-які індуктивні методи повинні містити в собі елемент дедукції, тобто спиратися на деякі твердження, перевіряти на них правильність індуктивного висновку на всіх етапах його формування. Саме ці міркування були прийняті до уваги при створенні МГС. В цьому методі зовнішнім доповненням є дедуктивне твердження (теорема), що не дозволяє надмірно ускладнювати результуючу модель, побудовану методами індукції. У цьому методі, як і у МГУА, зовнішнє доповнення є фільтром, що відкидає "сміття" і залишає тільки те, що формує аналізовану закономірність (1).

Задача відновлення багатовимірної функції (1) у МГС спочатку зводиться до стандартної задачі навчання розпізнаванню образів (НРО), а потім рішення задачі НРО забезпечує рішення задачі відновлення із заданою якістю і надійністю. Нехай задана навчальна вибірка пар $x_1y_1, \ldots, x_ly_l$, де x_v - вектор, а y_v - скаляр, який визначає значення функції F (1) у точці, що відповідає вектору x_v. Потрібно відновити багатовимірну неперервну функцію F так, щоб для будь-якого x_v, виконувалася нерівність

$$|y_v-F(x_v\alpha)|\le\xi, \quad (6)$$

де α - параметри функції, що відновлюється.

Хай, як і в [3], кожному x_v відповідає два значення y:

$$y_{v1}= y_v+\xi; \qquad y_{v2}= y_v - \xi; \quad (7)$$

При цьому навчальна вибірка збільшиться вдвічі і розділиться на дві підмножини V_1 і V_2, одна з яких містить елементи (x_{v1}, y_{v1}), а інша (x_{v2}, y_{v2}). Підмножини V_1 і V_2 можна розглядати як образи, які є явно відокремленими, і якщо

вдасться розділити ці образи, то тим самим буде відновлено функцію $F(x, \alpha)$, а це гарантує виконання (6) для всієї навчальної вибірки. Порушення співвідношення (6) буде відбуватися з частотою помилкового розпізнавання образів V_1 і V_2. Для поділу на образи можна використовувати будь-який алгоритм навчання розпізнаванню образів, у тому числі й альфа-процедуру [3], що, у свою чергу, є однією з реалізацій методу граничних спрощень. Як вирішальна, тобто, і апроксимуюча функція $F(x, \alpha)$ використовується поліном (2), у якому m-розмірність вектора $X=x(x_1,\ldots,x_m)$, а α - параметри, що настроюються.

З геометричної точки зору поліном (2) є гіперплощиною у спрямляючому просторі узагальнених координат x_i; x_ix_j; $x_ix_jx_k$, і т.д. Процедура відбере тільки ті доданки, які не будуть суперечити нерівності (6), тобто процедура як би пропускає поліном (6) через сито, залишивши тільки те, що визначає залежність (1), а все «сміття» буде відкинуте.

У процесі навчання частина членів полінома відкидається, а частина використовується як аргументи функції $F(x,\alpha)$. Таке сортування членів полінома здійснюються за допомогою зовнішнього доповнення, що є деяким дедуктивним твердження, яке засноване на передумовах теорії емпіричного ризику. Основний зміст цієї теорії - це вказування тих умов, при виконанні яких емпіричний ризик (тобто ризик, що обчислюється по емпіричним даним) збігається або майже збігається із середнім ризиком, обчисленим за нескінченними вибірками.

Одне з центральних тверджень теорії емпіричного ризику, що використовується в рамках проблеми НРО, міститься в теоремі [4], в якій доводиться те, що якщо з N вирішальних правил вибирається одне, що безпомилково розділяє випадкову і незалежну вибірку довжини l, то з ймовірністю $(1-\eta)$ можна утверджувати, що ймовірність помилкової класифікації за допомогою цього правила не перевищує наперед задане число

$$\varepsilon = \frac{\ln N - \ln \eta}{l}. \qquad (8)$$

Ця теорема лягла в основу теорії редукції [2], на якій базується метод граничних спрощень. Відповідно до цієї теорії задача синтезу складних поверхонь, що розділяють, у вихідному "засміченому" просторі величезної розмірності заміняється задачею синтезу такого простору малої розмірності ($n_0 << m$), у якому образи, які подані навчальною вибіркою, легко розділяються простим (частіше усього лінійним) вирішальним правилом. Іншими словами, насамперед приводиться у відповідність складність конструкцій, які індукують, із наявним обсягом інформації. Вихідна задача зводиться до більш простої таким чином, щоб її складність не перевищувала можливостей емпіричних даних, тобто складність повинна бути не вище тієї, що може упевнено проявитися в рамках наявної інформації.

Основою редукції при індуктивному синтезі простору є дедуктивне твердження приведеної вище теореми. Розширення простору обмежено дедуктивним твердженням, що не припускає ніяких ускладнень без істотного поліпшення результату. Дедуктивне твердження теореми як би зважує кожне ускладнення і визначає його вартість. Здійснюється послідовний синтез простору, у якому можливий лінійний поділ. При цьому дедуктивне твердження дуже просто перевіряється на навчальній послідовності.

Різновидом МГС є альфа-процедура [3], основна особливість якої полягає в тому, що кожна з m властивостей піддається особливій перевірці за допомогою критерію зовнішнього доповнення, який визначається дедуктивним твердженням теореми. При цьому, перед тим, як будувати індуктивну конструкцію, кожна властивість перевіряється на відповідність теоремі, і якщо практика не суперечить вимогам теорії, то продовжується ускладнення індуктивного висновку.

Якщо проводити аналогію між МГУА і МГС, то в МГУА зовнішнє доповнення формується на основі індуктивного висновку і на кожному кроку порівнюються дві індуктивних конструкції, побудовані на двох різноманітних вибірках: на навчальній і перевірочній. І якщо ці конструкції близькі, то ускладнення продовжується. У МГС порівнюються дві моделі: індуктивна і дедуктивна. Ускладнення є припустимим, якщо відокремлююча силу цього ускладнення, яка легко обчислюється за навчальною вибіркою, перевищує мінімально припустиму відокремлюючу силу, що обчислюється відповідно до теореми і виконує роль зовнішнього доповнення. Як у тому, так і в іншому методі ускладнення продовжується доти, поки воно дає позитивний результат, тобто поки моделі, які порівнюються, не суперечать одна одній.

Можливості поєднання МГУА та МГС

Обидва методи базуються на застосуванні принципу зовнішнього доповнення, щоб усунути некоректності у постановці задачі. У МГУА - це додаткова перевірка проміжних моделей, які отримані на перевірочній вибірці, а в МГС індуктивна модель будується з урахуванням її відповідності з теоретичними рекомендаціями, отриманими в результаті дедуктивних міркувань.

В обох випадках застосовуються засоби проти зайвого переускладнення моделі, але в МГУА це здійснюється шляхом додаткового контролю екстраполяційних властивостей, а в МГС - шляхом забезпечення умов, при яких ці екстраполяційні властивості гарантовані за теорією.

Можна сподіватися, що поєднання цих методів дозволить помітно поліпшити розв'язання задач відновлення залежностей в умовах коротких вибірок шляхом послідовного застосування відразу двох критеріїв зовнішнього доповнення.

Можна використовувати проміжні поліноми МГУА як вихідні дані альфа-процедури, які уже пройшли додаткову перевірку на перевірочній вибірці. Це дає явну перевагу при ускладненні результуючого полінома. У звичайній альфа-процедурі (МГС) після перевірки всіх лінійних членів полінома (2) використовуються всі нелінійні члени цього полінома як додаткові змінні, що складаються з двох аргументів (x_i, x_j), а в комбінованому методі для цієї мети можна використовувати моделі, отримані в першому ряду МГУА, які пройшли самовідбір за будь-яким критерієм зовнішнього доповнення (наприклад, за критерієм середньоквадратичного відхилення на перевірочній вибірці). У цьому випадку альфа-процедура здійснить перевірку тільки тих моделей $y(x_i,x_j)$, які були обчислені за (5), що пройшли поріг самовідбору за критерієм зовнішнього доповнення МГУА. У результаті кожна проміжна модель, тобто кожне значення y пройде подвійну перевірку: на спроможність працювати на нових даних (перевірочна вибірка) і на відповідність теорії, тобто на спроможність взяти на себе навантаження, обумовлене мінімально припустимою відокремлюючою силою. Таким чином, у цьому випадку МГУА відіграє роль відбору нелінійних змінних для альфа-процедури. Лінійні змінні відбираються звичайним алгоритмом альфа-процедури, а якщо серед них не знайдеться достатньої кількості змінних, що мають відокремлюючу силу, більшу за мінімально припустиму, то починає працювати МГУА, який на першому ряду генерує, а потім перевіряє по зовнішньому критерію проміжні поліноми другого ступеня, частина з яких використовуються як вхідні в альфа-процедуру. Якщо ж і це не призводить до повного лінійного поділу навчальної вибірки, то використовуються поліноми більш високих ступенів, що пройшли перевірку на перевірочній вибірці.

Можна поміняти місцями МГС і МГУА, використовуючи МГС для попередньої перевірки змінних, що використовуються у подальшому в МГУА. У цьому випадку відповідно до МГС будується лінійна модель по загальній схемі. Якщо ця модель виявиться не повною, тобто не дасть безпомилкового поділу вибірки, то серед усіх $\mathbf{x_i}$ **(i=1,2,...,m)** вибираються тільки m*, відокремлююча сила яких максимальна, і ці змінні використовуються в першому ряду МГУА для побудови квадратичних часткових поліномів. Коефіцієнти проміжних поліномів визначаються по МНК, а самі поліноми перевіряються на перевірочній вибірці. З побудованих поліномів вибирається m* найкращих і з них формується $C^2_{m^*}$ пар, кожна з яких перевіряється і ранжирується за алгоритмом альфа-процедури по дедуктивному критерію мінімально припустимої відокремлюючої сили. Усі пари, побудовані на змінних, які обчислюються алгоритмом МГУА, ранжируються за дедуктивним критерієм відокремлюючої сили. Кожній парі буде відповідати узагальнена змінна $\tilde{x}_{ij}$ [3]. Кращі пари, тобто відповідні їм змінні $\tilde{x}_{ij}$, використовуються для формування пар нового ряду МГУА. Далі працює звичайна схема МГУА, тобто всім парам приписується свій поліном, і знаходяться по навчальній послідовності його коефіцієнти. Всі поліноми перевіряються на перевірочній вибірці і m* кращих із них формуються в пари, кожна з яких перевіряється за критерієм відокремлюючої сили. Серед усіх пар вибирається m* найкращих і відповідні їм узагальнені змінні $\tilde{x}$ використовуються для формування нових пар для алгоритму МГУА. Така схема спільної роботи двох методів припускає паралельну роботу двох алгоритмів - алгоритму МГУА й алгоритму МГС. Алгоритм МГС припиняє свою роботу тоді, коли на якомусь етапі або відбудеться безпомилковий поділ вибірки, або алгоритм МГУА призведе до погіршення зовнішнього критерію. Як і в попередній комбінованій схемі правило, згідно з яким зупиняється робота алгоритму, можна лишити за конструктором, що вибере момент зупинки роботи або в МГС, або - МГУА.

ЛІТЕРАТУРА

1. Ивахненко А. Г., Степашко В.С. *Помехоустойчивость моделирования*.- Киев: Наукова думка, 1985 - 215 с.

2. Васильев В.И. *Теория редукции в проблемах экстраполяции.* //Проблемы управления и информатики. 1996 - №1-2, с.

3. Васильев В.И., Суровцев И.В. *Индуктивные методы обнаружения закономерностей, основанные на теории редукции* //Управляющие системы и машины - 1998 - №5 с.3-14

4. Вапник В.Н. *Алгоритмы и програмы восстановления зависимостей.* - М.: Наука, 1984 - 815 с.

5. Вапник В.Н., Червоненкис А.Я. *Теория распознавания образов*.- М.: Наука, 1974. -416 с.