Лапач С.М. Забезпечення необхідних властивостей вибірки для побудови регресійної моделі // Физические и компьютерные технологии. Труды 15-й Международной научно-технической конференции, 2-3 декабря 2009г. – Харьков: ХНПК «ФЭД», 2009. – С.179–182.

Abstract

The Paper investigates the problem of applying the idea theory design of experiments to matrix of passive experiments. The scientific idea and algorithms of forming subsamples from matrix of passive experiments and completing the construction of samples are described. The work results have been reduced to practice in R&D institute of aviation.

Кл. сл.
Планирование эксперимента, пассивный эксперимент, формирование выборки

design of experiments, passive experiments, construction of samples

Лапач С.М. Забезпечення необхідних властивостей вибірки для побудови регресійної моделі

НТУУ «Київський політехнічний інститут»

Теорія планування експерименту конструює матриці експериментів, які забезпечують побудову регресійних моделей з необхідними дослідику властивостями [1]. Разом з тим існує велика кількість задач, в яких застосування планування експерименту неможливе. Дослідник отримує вибірку пасивного експерименту з характеристиками, які не дають змогу побудувати регресійних моделей з задовільними характеристиками [2].

Проблема.

Відсутність апарату формування вибірки з заданими властивостями для пасивного експерименту.

Мета.

Застосування методології планування експерименту для випадку пасивного експерименту.

Наукова ідея.

І D-оптимальність [1], і робастність [3,4], які покладені в основу найбільш ефективних планів експерименту, вимагають певної організації розміщення точок в факторному просторі. Для обох систем точки розміщуються рівномірно в факторному просторі, Тільки першому випадку вони утворюють регулярну, а в другому – випадкову просторову решітку. Пропонується з існуючої вибірки сформувати підвибірку, точки якої будуть якомога більш рівномірно розміщені у факторному просторі.

Точки мають бути рівномірно розподілені в факторному просторі. Простір має мати форму гіперкубу, або гіперкулі. Для цього необхідного видалити частину точок з вибірки, зберігаючи по можливості загальні інтервали існування факторів. Видалені точки перейдуть до контрольної підвибірки.

Якщо форма простору занадто відрізняється від ідеальних, то вибір неможливий і необхідні інші методи роботи: нова формалізація задачі або перетворення факторного простору, або додавання додаткових точок, які змінять форму факторного простору.

Необхідна передумова.

Деформація вважається викликаною особливостями збору даних, а не функціональною залежністю між змінними, чи обмеженнями, які між ними існують. В такій ситуації намагання «виправити» наявну ситуацію стає по суті фальсифікацією експериментальних даних. Це видалення з даних інформативної частини.

Визначення степені деформованості факторного простору

Ідеальний простір розглядається як одиничний гіперкуб, або гіперкуля, вписана в одиничний гіперкуб.

Для нормованого в одиничний гіперкуб простору координати вершин можна записати як двійкові числа, які відповідають номеру вершини мінус 1.

Наприклад, для трьохвимірного простору координатами вершин приведені в табл.. 1.

Табл. 1. Координати вершин для трьохвимірного простору

Номер вершини	Координати
1	0,0,0
2	0,0,1
3	0,1,0
4	0,1,1
5	1,0,0
6	1,0,1
7	1,1,0
8	1,1,1

Для визначення ступеню деформації гіперкубу необхідно визначити відстані між протилежними вершинами.

Кількість вершин гіперкубу визначається як $N_{вершин} = 2^m$. З табл.. 1легко побачити, що протилежні вершини це такі, для яких сума номерів дорівнює $N_{вершин} - 1$. Наприклад, перша і восьма, четверта і п’ята, тощо.

Ступінь реформованості визначається за наступним показником $Q_{гіперкуб} = \prod_{i=1}^{K} \frac{D_i}{D_{діаг}}$, де K –кількість діагоналей, D_i – значення діагоналі в реальному факторному просторі, $D_{діаг}$ – теоретичне значення діагоналі. Кількість діагоналей $K = 2^{m-1}$, а відстань між протилежними вершинами для ідеального гіперкубу $D_{діаг} = \sqrt{m}$.Тут m – розмірність факторного простору. Для гіперкулі радіус, який використовується замість $D_{діаг}$=1. Тоді значення показнику розраховується $Q_{гіперкуля} = \prod_{i=1}^{K} D_i$. Може бути перераховане $Q_{гіперкуля} = Q_{гіперкуб} \times \frac{1}{(\sqrt{m})^m}$. D_i – довжини осей по різним векторам. При цьому критерій якості змінюється в інтервалі 0≤Q≤1. Значення 1 відповідає ідеальному найкращому варіанту, коли множина точок в факторному просторі утворює гіперкуб чи гіперсферу. Значення близько 0 відповідає виродженому факторному простору, коли дві чи більше координатних осей майже паралельні. Чим ближче значення Q до 1, тим ближче форма простору до ідеального. Початкові дані, які далекі від ідеальної форми, не можуть бути задовільним чином розбиті на навчальну та контрольну вибірку. Таким чином, можливо до початку обробки визначити можливість ефективного розбиття. В табл.. 2 представлено зв'язок між коефіцієнтом реформованості і закорельваністю.

Табл. 2 Співвідношення між показником реформованості і закорельованістю

Коефіцієнт деформації для однієї осі (двовимірний випадок)	Коефіцієнт кореляції
0,25	>0,8

0,5	>0,45
0,75	>0,25

Для визначення граничного коефіцієнту для багатовимірного випадку необхідно вказане значення піднести до степеню, який дорівнює числу осей.

Алгоритм визначення деформованості простору

1. Визначення координат теоретичної вершини за двійковим кодом.
2. Визначення координат реальної вершини. Для цього знаходиться точка з реальної вибірки, найближча до теоретичної.
3. Для протилежних вершин знаходиться відстань між реальними точками.
4. Обчислюється коефіцієнт деформованості як відношення реальної відстані до теоретичної для даної осі.
5. Якщо коефіцієнт менше критичного, то діагональ і реальні координати вершин для неї запам'ятовуються.
6. п.п. 1-5 виконуються для всіх теоретичних діагоналей гіперпростору.

В зв'язку з тим, що в багатофакторній вибірці число теоретичних діагоналей дуже велике, кожна з реальних точок біде використовуватись в кількох діагоналях.

Опис алгоритму формування підвибірки

1. Найти граничні точки $x_{ep} \in \{\min X_i; \max X_i\} \forall i \in (1, m)$.
2. Нормувати факторний простір в одиничний гіперкуб. $x'_{ij} = \dfrac{x_{ji} - \min x_i}{\max x_i - \min x_i}$
3. Задати гіпотетичну складність залежності відгуку від конкретної змінної за допомогою степені апроксимуючого поліному по і-му фактору f_i.
4. Задати критичне значення коефіцієнту спотворення простору $Q_{кр}$.
5. Визначити необхідну кількість точок, виходячи з гіпотетичної складності залежності відгуку від конкретної змінної N_{min} і N_{max}(за [3]).
6. Розрахувати можливість ефективного розбиття на підвибірки.
7. Якщо розбиття неможливе – роботу припинити. Або перейти в п.3 для завдання нових степенів і перерахунку кількості дослідів.
8. Знайти критичну відстань між точками. $d_{кр} = \sqrt{\sum_{i=1}^{m} \Delta_{x_i}^2}$, де $\Delta_{x_i} = (x_{i\max} - x_{i\min}) / (f_i + 1)$.
9. Найти матрицю відстаней між точками $d_{ij} = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2}$.
10. Взявши за основу точки, які утворюють вісь з найбільшою довжиною, залишити в навчальній вибірці тільки точки, відстань між якими більша граничної.
11. Якщо кількість дослідів менше необхідного, то додати до вибірки нові точки, які знаходяться якнайдалі, до вже включених. Визначаються за умовою $\sum_{i=1}^{N} \left| D_{\max} - d_{ij} \right| \Rightarrow \min$. Тут d_{ij} – відстань між точками, D_{max} –

максимальна довжина теоретичної діагоналі, і – номер уже вибраної точки, j – номер точки-кандидату, N – кількість вже вибраних точок.

По. п.11 можна застосувати покроковий відбір. При додаванні кожної нової точки розраховуються статистичні і обчислювальні характеристики матриці, і приймається рішення про зупинку, продовження процесу включення нових точок або повернення на попередній крок.

В тих ситуаціях, коли можливо проведення додаткових експериментів пропонується алгоритм добудови точок для деформованого простору.

1. Вибирається діагональ з деформацією більше критичної.
2. Для даної діагоналі знаходяться координати теоретичних вершин.
3. Визначені в п. 2 вершини додаються до списку додаткових точок.
4. п.п.1-3 виконуються для всіх деформованих діагоналей.

Після додавання до вихідної матриці експериментів, визначених за вказаним алгоритмом, вибірка буде мати необхідні статистичні властивості.

Висновки

Представлено ефективні алгоритми формування підвибірки з пасивного експерименту і добудови пасивного експерименту для отримання заданих властивостей, відповідних вимогам робастності. Розроблено відповідні програмні модулі, які розширюють можливості програмних засобів [3, 5]. Результати роботи впроваджені в НДІ Авіації. Застосування запропонованих алгоритмів дозволило формувати з матриць пасивного експерименту навчальні матриці з задовільними властивостями (зменшення закорельованості від 0,7-0,9 до 0,3-0,4) для отримання регресійних моделей прогнозування показників апаратури.

Список літератури

1. Налимов В.В., Голикова Т.И. Логические основания планирования эксперимента. 2-е изд.перераб. и доп. –М.: 1981. «Металлургия». – 152с.
2. Лапач С.Н. Проблемы построения математических моделей экспериментально-статистическими методами // Прогресивна техніка і технологія машинобудування, приладобудування і зварювального виробництва. Праці НТУУ “КПІ”, –Т. 2, –К.: НТУУ “КПІ”, –1998. - С.25-29.
3. С.Н. Лапач, А.В. Чубенко, П.Н. Бабич Статистические методы в медико-биологических исследованиях с использованием Excel –2 изд. перераб. и доп. –К.: 2001, Морион. – 408с.
4. Радченко С.Г. Устойчивые методы оценивания статистических моделей. – К.: ПП «Саспарель». –504с.
5. Лапач С.Н., Радченко С.Г., Бабич П.Н. Планирование, регрессия и анализ моделей PRIAM (ПРИАМ) / Каталог программные продукты Украины. К.: 1993. С. 24-27.