ЛЕКЦІЯ 6. ЕВРИСТИЧНІ МЕТОДИ ЗМЕНШЕННЯ КІЛЬКОСТІ ВИХІДНИХ ОЗНАК

НАВЧАЛЬНІ ПИТАННЯ

Природа евристичних методів

Метод екстремального групування ознак

Метод кореляційних плеяд

1. Природа евристичних методів.

Наведені раніше методи скорочення розмірності досліджуваного ознакового простору, і в першу чергу, методи факторного аналізу, припускали інтерпретацію в термінах тієї чи іншої строгої імовірнісної моделі і, отже, мали на увазі можливість дослідження властивостей даних процедур в рамках теорії математичної статистики. У даному випадку мова піде про методи, підпорядковані деяким частковим цільовим установкам (найменше спотворення геометричної структури початкових «вибіркових точок», найменше спотворення їх еталонного розбиття на класи і т.д.), але не формульованих в термінах ймовірносно-статистичної теорії. Процедура вибору цільової установки, відповідної саме для даного конкретного завдання, практично не формалізована, носить евристичний характер, т. е., як правило, обумовлюється лише досвідом й інтуїцією дослідника. Тому такі методи називають *евристичними*.

За відсутності апріорної або вибіркової попередньої інформації про природу досліджуваного вектора спостережень і про генеральні сукупності, з яких ці спостереження витягуються, точно в такому ж невигідному положенні знаходяться і методи факторного аналізу і головних компонент. Проте для них все-таки існує принципова можливість теоретичного обґрунтування (за наявності відповідної додаткової інформації), тоді як лише деякі з евристичних методів вдається згодом теоретично обґрунтувати в рамках строгої математичної моделі.

Підкреслимо, що факт опису тут методів зниження розмірності, що не використовують попередньої інформації, наприклад, навчальних і квазінавчальних вибірок, доцільно розцінювати лише як наслідок визнання неминучості ситуацій, в яких такій інформації не є, але не як прагнення рекламувати ці методи як найбільшефективних. Насправді ж обґрунтування і ефективне рішення задач зниження розмірності без сліпої надії на успіх можна, на нашу думку, одержати лише на шляху глибокого професійного аналізу, доповненого статистичними методами, що використовують попередню вибіркову (навчальну) інформацію

2. Метод екстремального групування ознак

При вивченні складних об’єктів, які задані багатьма параметрами, виникає завдання розбиття параметрів на групи, кожна з яких характеризує об’єкт з

певного боку. Однак одержання результатів, що підлягають легкій інтерпретації, ускладнюється тим, що в багатьох застосуваннях вимірювані параметри (ознаки) лише побічно відображають істотні властивості, що характеризують даний об'єкт.

Отже, в багатьох випадках зміна вимірювання деякого загального фактора позначається не однаково на вимірюваних ознаках. Наприклад, вихідна сукупність з n ознак може мати природне розбиття на відносно невелику кількість груп. При цьому вимірювання ознак, що належать певній групі, обумовлюється одним загальним фактором, своїм для кожної групи. Після прийняття такої гіпотези розбиття на групи природно будувати таким чином, щоб параметри, що належать одній групі, були корельовані досить сильно, а параметри, що належать різним групам, навпаки, слабко. Після такого розбиття для кожної групи ознак будується випадкова величина, яка найбільш сильно корельована з представниками даної групи. Ця величина інтерпретується як шуканий фактор, від якого залежать всі параметри даної групи.

Очевидно, що подібна схема побудови діагностичних ознак є частковим випадком логічної схеми факторного аналізу. Однак на відміну від розглянутих класичним моделей факторного аналізу при евристично-оптимізаційному підході групування ознак і виокремлення загальних факторів здійснюються на основіекстремізації деяких евристично введених функціоналів. Розбиття, що оптимізують функціонал, називається*екстремальним групуванням параметрів*. Взагалі під завданням екстремального групування набору випадкових величин $X_1, X_2, \dots, X_n$ на наперед визначену кількість класів p розуміють відшукання такого набору підмножин$S_1, S_2, \dots, S_p$ натурального ряду чисел 1, 2, … , p і такої ж кількості p нормованих факторів $F_1, F_2, \dots, F_p$, які оптимізують певний критерій оптимальності. При цьому підмножини $S_1, S_2, \dots, S_p$ не перетинаються і разом з тим в об'єднанні охоплюють весь досліджуваний ряд чисел.

3. Перший алгоритм екстремального групування.

Перший алгоритм екстремального групування ознак в ролі критерію оптимальності використовує функціонал

$$J_1 = \sum_{i \in S_1} [corr(X_i, F_1)]^2 + \sum_{i \in S_2} [corr(X_i, F_2)]^2 + \dots + \sum_{i \in S_p} [corr(X_i, F_p)]^2 , \quad (7.1)$$

де $corr(X, F)$ – парний коефіцієнт кореляції між ознакою X та фактором F.

Позначимо через $A_k = \{X_i , i \hat{I} S_k\}$, $k = 1, 2, \dots, p$ – групу відповідних ознак при заданому розбитті. Максимізація функціонала (7.1) (як по розбиттю вихідних ознак, так і по вибору факторів) відповідає вимогам такого розбиття параметрів, коли в одній групі виявляються найбільш близькі між собою ознаки в сенсі їх кореляційного зв'язку. Дійсно, при максимізації функціонала (7.1) для заданого набору факторів $F_1, F_2, \dots, F_p$ в одну групу A_k потраплять такі

ознаки, що найбільш корельовані з фактором F_k . В той же час серед можливих наборів випадкових величин F_1, F_2, … , F_p буде обиратись такий, щоб кожна з F_k була найбільш близькою до представників своєї групи.

Очевидно, що при заданих класах S_1, S_2, … , S_p оптимальний набір факторів F_1, F_2, … , F_p утворюється в результаті незалежної максимізації кожного з доданків $\sum_{i \in S_k} [corr(X_i, F_k)]^2$ виразу (7.1) . Звідси маємо, що

$$\max_{F_1, F_2, \ldots, F_k} J_1 = \sum_{i=1}^{p} \lambda_i^2 , \qquad (7.2)$$

де λ_i – власне значення матриці R_i , що складається з коефіцієнтів кореляції змінних, які входять складу групи A_i .

При цьому оптимальний набір факторів F_1, F_2, … , F_p задається формулами

$$F_k = \frac{\sum_{i \in S_k} a_i^{(k)} X_k}{\sqrt{\sum_{i,j \in S_k} a_i^{(k)} a_j^{(k)} r_{ij}}} , \qquad (7.3)$$

де $r_{ij} = corr(X_i, X_j)$,

$a^{(k)} = (a_1^{(k)}, a_2^{(k)}, \ldots , a_{mk}^{(k)})$ – власний вектор матриці R_k , який відповідає власному значенню λ_k .

З іншого боку, вважаючи відомими фактори F_1, F_2, … , F_p неважко побудувати розбиття S_1, S_2, … , S_p, що максимізує J_1 при заданому наборі факторів, а саме:

$$S_k = \{ i : corr^2(X_i, F_k) \geq corr^2(X_i, F_q), \quad q = 1,2,\ldots, p \} . \qquad (7.4)$$

Співвідношення (7.3) та (7.4) є необхідними умовами максимізації J_1 .

Для одночасного знаходження оптимального розбиття S_1, S_2, … , S_p і оптимального набору факторів S_1, S_2, … , S_p пропонується ітераційних алгоритм, який послідовно здійснює оптимальних (по відношенню до розбиття, одержаного на попередньому кроці) факторів, а потім вибір розбиття, оптимального до факторів, одержаних на попередньому кроці. Тобто, спочатку обирається деяке розбиття вихідних ознак на групи. Далі для кожної групи будується фактор за формулою (7.3). Після цього здійснюється нове розбиття вихідної сукупності ознак на групи за правилом (7.4), тобто, ознака X_i відноситься до групи A_k , якщо квадрат коефіцієнта кореляції цієї ознаки з фактором F_k не менший за квадрат коефіцієнта кореляції цієї ознаки з іншими факторами. При цьому якщо буде мати місце випадок, коли декілька таких значень будуть однаковими, то віднесення ознаки X_i до однієї з груп здійснюється на основі змістовного аналізу. Очевидно, що на кожному кроці значення функціоналу (7.1) не зменшується. А отже, процес буде вести до його максимізації. Недоліком наведеного алгоритму є те, що одержаний максимум функціоналу може бути локальним.

4. Другий алгоритм екстремального групування.

Інший підхід до визначення екстремального групування параметрі полягає в максимізації функціоналу

$$J_2 = \sum_{i \in S_1} |corr(X_i, F_1)| + \sum_{i \in S_2} |corr(X_i, F_2)| + \ldots + \sum_{i \in S_p} |corr(X_i, F_p)| \quad . \quad (7.5)$$

В змістовному сенсі функціонал J_2 схожий на J_1 , і його Максимізація також відповідає основній вимозі до характеру розбиття вихідних ознак на групи. Необхідними і достатніми умовами максимуму функціоналу (7.5) є наступні:

- розбиття параметрів на групи $A_1, A_2, \ldots, A_p$ є таким, що функціонал $J_3 = \sum_{i=1}^{p} \sqrt{D(\sum_{i \in S_k} g_i X_i)}$ сягає максимуму як по розбиттю на групи, так і за значеннями коефіцієнтів g_i , які приймають значення або +1, або –1; символ D означає дисперсію випадкової величини;

- фактори F_k визначаються співвідношеннями

$$F_k = \frac{\sum_{i \in S_k} g_i X_i}{\sqrt{\sum_{i,j \in S_k} g_i g_j r_{ij}}} \quad . \quad (7.6)$$

Порядок максимізації функціоналу J_2 базується на попередній максимізації J_3 . Для цього циклічно перебираються значення ознак $X_1, X_2, \ldots, X_n$ і на кожному кроці приймається рішення про віднесення ознаки X_iдо однієї з груп $A_1, A_2, \ldots, A_p$ і визначається знак g_i . Таким чином, на черговому кроці одержується розбиття вихідних ознак $X_1, X_2, \ldots, X_n$ на групи $A_1^{(t)}, A_2^{(t)}, \ldots, A_p^{(t)}$ і обчислюються відповідні коефіцієнти $g_1^{(t)}, g_2^{(t)}, \ldots, g_p^{(t)}$ з урахуванням максимізації виразу $D(\sum_{i \in S_k} g_i X_i)$. Далі будується p допоміжних коефіцієнтів $g_{i,k}^{(t+1)}$ за формулою

$$g_{i,k}^{(t+1)} = sign \sum_{X_j \in A_k^{(t)}} g_j^{(t)} r_{ij}, \quad i \neq j \quad , \quad (7.7)$$

і для всіх k=1, 2, … , p обчислюються величини

$$\Delta_k^{(t+1)} = \sqrt{D(\sum_{\substack{X_j \in S_k^{(t)} \\ i=j}} g_j^{(t)} X_j + g_{i,k}^{(t+1)} X_i)} - \sqrt{D(\sum_{\substack{X_j \in S_k^{(t)} \\ i=j}} g_j^{(t)} X_j)^2} \quad .(7.8)$$

Далі обирається такий номер $k=k^*$, для якого

$$\Delta_{k^*}^{(t+1)} = \max_{1 \le k \le p} \Delta_k^{(t+1)} \quad , \quad (7.9)$$

і ознака X_i виключається з з групи A_k і приєднується до групи A_{k*} . Тобто, на цьому кроці лише одна ознака змінює свою належність одній з груп. Інші групи ознак на цьому кроці залишаються без змін. В результаті одержується нове розбиття ознак на групи. Коефіцієнти $g_{i,k*}^{(t+1)}$ обчислюються за правилами

$$g_j^{(t+1)} = g_j^{(t)}, \quad j \neq i,$$
$$g_i^{(t+1)} = g_{i,k*}^{(t+1)}. \qquad (7.10)$$

На наступному кроці аналогічно розглядається параметр X_{i+1} . Якщо всі параметри розглянуті, починається їх повторний перегляд з першого.

Процес завершується, коли при розгляді всіх параметрів чергового циклу збереглись як розбиття ознак на групи, так і значення всіх коефіцієнтів g_i .

Даний метод має свою "ідейну" близькість з центроїдним методом факторного аналізу.

Аналізуючи обидва способи, можна відмітити, обидва є досить трудомісткими з обчислювальної точки зору.

МЕТОД КОРЕЛЯЦІЙНИХ ПЛЕЯД

Метод кореляційних плеяд, так само як і метод екстремального групування призначений для знаходження таких груп ознак – «плеяд», коли кореляційний зв'язок – тобто сума модулів коефіцієнтів кореляції між параметрами однієї групи достатньо велика, а зв'язок між параметрами з різних груп (міжплеядна) — мала. За певним правилом по кореляційній матриці ознак утворюють креслення — граф, який потім за допомогою різних прийомів розбивають на підграфи. Елементи, що відповідають кожному з підграфів, і утворюють плеяду.

Розглянемо кореляційну матрицю R початкових ознак. Намалюємо n кружків; усередині кожного кружка напишемо номер однієї з ознак. Кожен кружок з'єднується лініями зі всією рештою кружків; над лінією, що сполучає i-й і j-й елементи (ребром графа), ставиться значення модуля коефіцієнта кореляції . Одержане таким чином креслення розглядаємо як початковий граф.

Задавшись (довільним чином або на підставі попереднього вивчення кореляційної матриці) деякимипороговими значеннями коефіцієнта кореляції r_0, виключаємо з графа всі ребра, які відповідають коефіцієнтам кореляції, меншим за модулем від порогового значення. Потім задаємо деяке $r_1 > r_0$ і відносно його повторюємо описану процедуру. При деякому достатньо великому r_s граф розпадається на декілька підграфів, тобто таких груп кружків, що зв'язки (ребра графа) між кружками різних груп відсутні. Очевидно, що для одержаних таким чином плеяд внутрішньоплеядні коефіцієнти кореляції будуть більше r_s, а міжплеядні — менші його.

У іншому варіанті кореляційних плеяд пропонується упорядковувати ознаки і розглядати тільки ті коефіцієнти кореляції, які відповідають зв'язкам між

елементами у впорядкованій системі.

Впорядкування проводиться на підставі принципу максимального кореляційного шляху: всі n ознак зв'язуються за допомогою $(n - 1)$ ліній (ребер) так, щоб сума модулів коефіцієнтів кореляції була максимальною. Це досягається таким чином: у кореляційній матриці знаходять найбільший по абсолютній величині коефіцієнт кореляції, наприклад $½r_{ij}½=r^{(1)}$. При цьому коефіцієнти кореляції, що стоять на головній діагоналі матриці і рівні одиниці, не розглядаються.

Малюємо кола, відповідні параметрам X_i та X_j, і над "зв'язком" між ними пишемо значення $r^{(1)}$. Потім, виключивши значення $r^{(1)}$ з вихідної кореляційної матриці, знаходимо найбільший коефіцієнт в j-му стовпці матриці (це відповідає знаходженню ознаки, яка найтісніше після X_i пов'язана з X_j) і найбільший коефіцієнт в i-тому рядку матриці (це відповідає знаходженню ознаки, що найсильніше після X_j пов'язана з X_i). Із знайдених таким чином двох коефіцієнтів кореляції обирається більший – позначимо його через $r^{(2)}$. Малюємо нове коло, позначаємо його відповідним індексом і сполучаємо з тим, який відповідає обраному значенню коефіцієнта кореляції. Далі для кожної з цих трьох ознак знаходимо найбільш близькі до них і обираємо ту, для якої зв'язок найщільніший, тобто, коефіцієнт кореляції за абсолютною величиною найбільший. Малюємо чергове коло і з'єднуємо його з одним із тих, що вже є за аналогічним наведеному вище правилом. Для того, щоб опрацьована ознака повторно не увійшла у граф, на кожному кроці вона виключається з розгляду. В результаті такого процесу буде зображено n кіл, з'єднаних $(n–1)$ ребром. Потім задаємо порогове значення r_0 вилучаємо з графа всі ребра, які відповідають меншим за модулем значенням коефіцієнтів кореляції, ніж порогове.

Назвемо незамкненим граф, для якого для довільних двох його кіл існує єдина траєкторія з його ребер, що з'єднує ці кола. Очевидно, що другий варіант методу кореляційних плеяд дає побудову лише незамкнених графів, що відсутнє у першому способі. Тому результати побудови графів, одержані різними методами, можуть не співпадати.

ПИТАННЯ ДЛЯ САМОПЕРЕВІРКИ

1. Поясніть природу евристичних методів зменшення ознакового простору
2. В чому сутність методу екстремального групування ознак? З якими іншими методами зменшенняознакового простору він має зв'язок?
3. В чому сутність методу кореляційних плеяд?