УДК 519.237.8

С.М. ЛАПАЧ[*]

ВИЗНАЧЕННЯ ОПТИМАЛЬНОЇ КІЛЬКОСТІ КЛАСТЕРІВ

[*]Національний технічний університет України «КПІ», Київ, Україна

Анотація. *Розглянута проблема визначення оптимальної кількості кластерів для випадку розривного факторного простору. Викладені наукова ідея і алгоритм формування кластерів, який базується на опису залежності значення критерію Q від кількості кластерів сплайн-регресією. Галузь застосування: виділення однорідних підобластей факторного простору при побудові регресійних моделей.*

Ключові слова: *кластерний аналіз, регресійний аналіз, факторний простір, сплайн-регресія.*

Аннотация. *Рассмотрена проблема определения оптимального количества кластеров в случае разрывного факторного пространства. Описаны научная идея и алгоритм формирования кластеров, который основан на описании зависимости изменения критерия Q от количества кластеров сплайн-регрессией. Область применения: выделение однородных подобластей факторного пространства при построении регрессионных моделей.*

Ключевые слова: *кластерный анализ, регрессионный анализ, факторное пространство, сплайн-регрессия.*

Abstract. *The problem of definition of the optimum quantity of clusters in case of discontinuous factor space was considered. The scientifically-based idea and cluster formation algorithms are described; the latter being based on the description of dependency of Q criteria changes from the quantity of spline regression clusters. Areas of use – allocation of homogeneous subareas of factor space at regression models creation.*

Keywords: *cluster analysis, regression analysis, factor space, spline regression.*

1. Вступ. Проблема і мета роботи

Однією із нерозв'язаних проблем у регресійному аналізі є побудова моделей для неоднорідних (розривних) областей факторного простору [1]. Виділення однорідних підобластей можна виконати за допомогою кластерного аналізу, відповідним чином підібравши його параметри [2]. Разом з тим при числі областей більше двох виникає проблема визначення a priory невідомої їх кількості. Вона пояснюється відсутністю показника якості розбиття на кластери, який дозволив би визначити оптимальну (істинну) кількість кластерів при розбитті. В кластерному аналізі для оцінки якості розбиття найбільш часто використовується показник Q [3–6]. Це показник внутрішньогрупового розсіювання, який розраховується за формулою $Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i} d^2(X_j,\overline{X}_i)$. Для випадку нечіткої кластеризації ця формула приймає вигляд $Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mu_{ij} d^2(X_j,\overline{X}_i)$. В ній залежність елементів вибірки до певного класу описується матрицею $\mu_{ij}\in[0,1], i=\overline{1,M}; j=\overline{1,k}$. Рядок і містить степінь належності об'єкта i до кластера j. При цьому $\sum_{j=1}^{k}\mu_{ij}=1; 0<\sum_{i=1}^{M}\mu_{ij}<M$. Як найкраще, розбиття рекомендується таке, яке мінімізує вказаний показник. На жаль, вказаний показник монотонно зменшується при збільшенні числа кластерів і дорівнює нулю (мінімум), коли кількість кластерів дорівнює кількості об'єктів. Тобто в використанні мінімуму цього показника для

ISSN 1028-9763. Математичні машини і системи, 2015, № 3

вибору оптимальної кількості кластерів немає практичного сенсу. У зв'язку з вищевказаним є необхідність розробити методику, яка дозволить визначати оптимальну кількість кластерів.

2. Базові припущення

Будемо вважати, що існує ідеальне розбиття, яке відповідає фактичній кількості кластерів. Тоді, незважаючи на монотонність зменшення показника Q, швидкість зміни його значення на різних етапах буде різна. Спочатку зменшення його відбувається за рахунок того, що об'єкти, які штучно об'єднані в один великий кластер, розкидаються по кількох різних. При цьому до моменту наближення до «істинної» кількості кластерів внутрішньогрупове розсіяння зменшується дуже швидко з кожним кроком. Потім, коли кожен «істинний» кластер розбивається на кілька штучних, швидкість зміни показника різко зменшується. Звичайно, на характер цієї залежності буде впливати те, наскільки чітко розділяються кластери. Чим більш чітке розділення, тим більш явно буде відбуватись зміна швидкості.

Оскільки ми не можемо шукати мінімум, то пропонується знайти точку, в якій відбувається різка зміна швидкості зменшення показника за допомогою сплайн-регресії [7]. Залежність описується двома відрізками. Точка перелому вибирається така, яка забезпечує мінімум залишкової дисперсії. Відповідно координати точки перелому і будуть відповідати «істинній» кількості кластерів.

3. Приклад і загальний алгоритм розбиття

Розглянемо застосування пропонованого підходу на тестовому прикладі. Приклад створений таким чином, щоб наочно показати, як відбувається визначення кількості кластерів.

Вихідні дані, які описують вибірку, представлені в табл. 1.

Таблиця 1. Вихідні дані для прикладу

Кластер	Об'єкт	X1	X2	Кластер	Об'єкт	X1	X2
1	1	3	13,5	3	17	11	5,4
1	2	4	13,7	3	18	10,2	5
1	3	3,5	12,75	3	19	9,5	4,5
1	4	2,5	12	3	20	11,3	4,3
1	5	3	11,2	3	21	10,3	3,8
1	6	4,1	11,4	3	22	9,3	3,5
1	7	5	12	3	23	10	2,5
2	8	11	11,5	4	24	4	2,3
2	9	11,75	11,6	4	25	2,3	2,5
2	10	11,5	10,5	4	26	4	3,5
2	11	10,3	10	4	27	4,8	3,5
2	12	10,2	9	4	28	3,1	3,6
2	13	11	8,5	4	29	2,5	4,8
2	14	12,4	9,2	4	30	3,5	5
2	15	13	10,5	4	31	5,2	4,6
3	16	9,9	5,7	4	32	4,3	5,5

У цій штучній задачі «істинна» кількість кластерів – 4. Якщо виконати розрахунки методом k середніх для різної кількості кластерів (від 2 до 8), то залежність внутрішньогрупового розсіяння від кількості кластерів буде мати вигляд, представлений на рис. 1 (ламана $Q_{екс}$). Як ми бачимо, характер зміни повністю відповідає висунутим припущенням. Указаний фрагмент даних досить добре описується експонентою. Чітко видно, що ніякого

мінімуму при чотирьох кластерах не існує, але зате є різке зменшення крутизни спаду. Опис цієї кривої експонентою хоч і є найкращим зі статистичної точки зору, але не дає нам ніякої інформації про зміну тенденцій.

Якщо цю залежність намагатись описати ламаною (сплайн-регресією), то найкращий по мінімуму залишкової дисперсії варіант буде з точкою перелому при значенні незалежної змінної, рівної 4. У табл. 2 наведені характеристики регресійної моделі, яка забезпечує мінімум залишкової дисперсії. Модель розрахована за спеціальною програмою [7].

Таблиця 2. Регресійна сплайн-модель і її статистичні характеристики

Множинний коефіцієнт кореляції, R	0,983207				
Частка, пояснювана моделлю, R^2	0,966696				
Розрахункове F-відношення для R	58,05198				
Критичне значення для F_R	6,944272	V_1=	2	V_2=	4
Залишкова дисперсія	827,3025				
Число обумовленості	389,5287				
Точки перелому					
Номер	1				
Координата	4				
Коефіцієнти регресії					
Ім'я	Значення				
0	668,8234				
1	-158,933				
2	157,3249				

Таким чином, ми переходимо від задачі пошуку неіснуючого в рамках поставленої задачі мінімуму функції Q до мінімізації залишкової дисперсії регресії сплайн-моделі (сформованої двома відрізками), яка описує залежність Q від кількості кластерів. Ця залежність має мінімум при кількості кластерів, яка відповідає найкращому розбиттю.

Звичайно, при великій розмірності простору і потенційно великій кількості можливих варіантів розбиття на кластери недоцільний або й неможливий їх повний перебір для побудови залежності Q. В цьому випадку послідовність визначення може мати такий вигляд (скорочено).

1. Побудова дендриту, який охоплює всі об’єкти вибірки.

2. Визначення набору критичних відстаней, наприклад, при заданні переліку значень $k(0, 1, 2...)$ у формулі $d_{kp} = d_{cep} + k\sigma_d$ [5].

3. Сформувати кластери відповідно до кожного варіанту критичної відстані.

4. Розрахувати Q для отриманих варіантів розбиття на кластери.

5. Побудувати сплайн-регресію залежності Q від кількості кластерів як сплайн-регресію і визначити оптимальну кількість кластерів аналогічно описаному вище.

6. Якщо коефіцієнти двох частин сплайн-регресії статистично не відрізняються один від одного, то це означає відсутність перелому (зміни швидкості). Тоді необхідно перейти в п.2 для зміни значення k і розширення інтервалу кількості кластерів.

7. Визначити критичну відстань, яка відповідає оптимальній кількості кластерів.

8. Сформувати кластери відповідно визначеній в п.7 критичній відстані.

9. Перерахувати Q і порівняти з теоретично оптимальним. У разі невідповідності – перейти в п.7 для коригування критичної відстані.

За обмежену кількість ітерацій задача розбиття буде розв’язана. Отримане розбиття буде відповідати розбиттю на однорідні підобласті факторного простору.

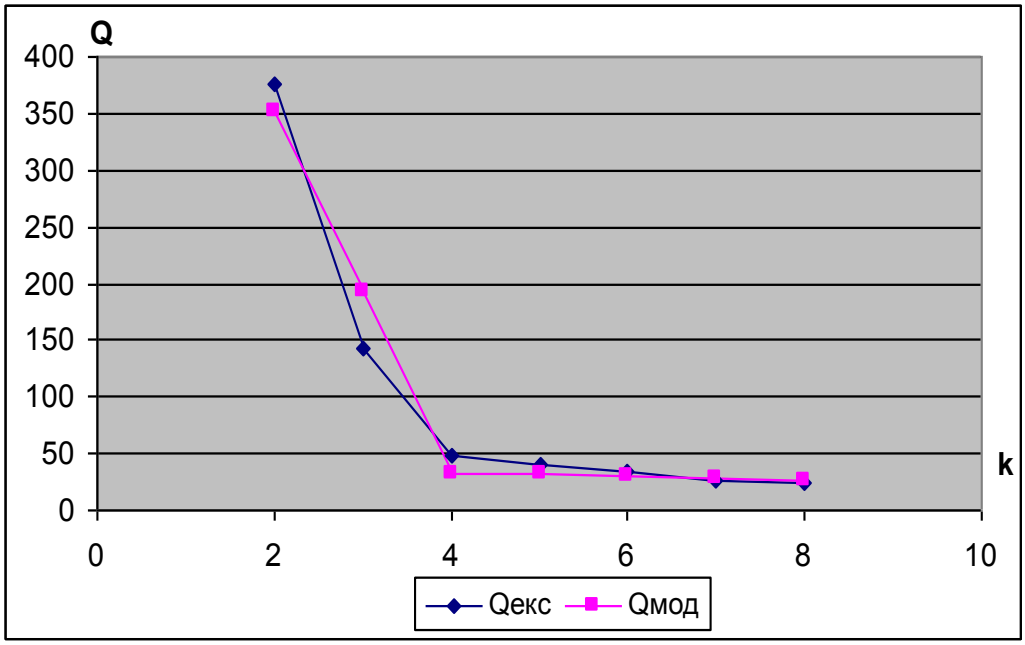

Рис. 1. Залежність внутрішньогрупового розсіяння Q від кількості кластерів і її опис сплайн-регресією

3. Висновки і рекомендації

Стаття продовжує цикл робіт по «закриттю прогалин» у класичній регресії [1, 2]. Рекомендований в літературі показник якості розбиття Q (внутрішньогрупове розсіяння) непридатний для визначення найкращої кількості кластерів шляхом його мінімізації, оскільки він монотонно зменшується до нуля при збільшенні кількості кластерів до числа об'єктів. У роботі пропонується знаходження цієї оптимальної кількості через опис залежності Q від числа кластерів сплайн-регресією і вибір точки перелому її як «істинної» кількості кластерів. Запропонований підхід і розроблені алгоритми дозволяють розв'язувати задачу розбиття факторного простору на однорідні підпростори у випадку наявності в ньому розривів, що має велике прикладне значення для багатофакторного регресійного аналізу.

СПИСОК ЛИТЕРАТУРИ

1. Лапач С.Н. Основные проблемы построения регрессионных моделей / С.Н. Лапач, С.Г. Радченко // Математичні машини і системи. – 2012. – № 4. – С. 125 – 133.
2. Лапач С.М. Кластерний аналіз при визначенні однорідних областей факторного простору в регресійному аналізі / С.М. Лапач // П`ятнадцята міжнародна конференція ім. академіка Михайла Кравчука, (Київ, 15–17 травня 2014 р.). – Т. 3: Теорія ймовірностей та математична статистика. – К.: НТУУ «КПІ», 2014. – С. 82 – 84.
3. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА / Енюков И.С. – М.: Финансы и статистика, 1986. – 232 с.
4. Мандель И.Д. Кластерный анализ / Мандель И.Д. – М.: Финансы и статистика, 1988. – 176 с.
5. Плюта В. Сравнительный многомерный анализ в эконометрическом моделировании / Плюта В.; пер. с польск. – М.: Финансы и статистика, 1989. – 175 с.
6. Штовба С.Д. Проектирование нечетких систем средствами MATLAB / Штовба С.Д. – М.: Горячая линия – Телеком, 2007. – 288 с.
7. Кузьмін В.М. Використання полігональної регресії в економічних дослідженнях / В.М. Кузьмін, С.М. Лапач // Економіка і управління. – 2004. – № 3. – С. 79 – 84.

Стаття надійшла до редакції 06.10.2014