

## МЕТОДЫ ОПРЕДЕЛЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ ПРИ КЛАССИФИКАЦИИ БЕЗ ОБУЧЕНИЯ

*Ирина Яцкив, Лада Гусарова*

*Институт транспорта и связи  
Ломоносова 1, Рига, LV 1019, Латвия.  
Ph: (+371)-7100650. Fax: (+371)-7100660. E-mail: ivl@tsi.lv*

Кластерный анализ — технология группирования объектов в неизвестные группы. Он отличается от дискриминантного тем, что неизвестны ни число кластеров, ни их характеристики. Проблема определения числа кластеров является одной из основных нерешенных до настоящего времени задач кластерного анализа (непараметрического случая классификации). Обратимся к двум наиболее используемым типам процедур кластерного анализа: иерархическим и итеративным. Для итеративных алгоритмов число кластеров является одним из входных параметров алгоритма. Для иерархических процедур характерен визуальный анализ дендрограммы и определение по ней наиболее предпочтительного числа кластеров.

По дендрограмме можно определить разницу между уровнями объединения. Уровнем будем считать очередной шаг алгоритма, на котором происходит объединение кластеров. Наибольшая разница по оси расстояний между соседними уровнями указывает на предпочтительное число классов (соответствующее уровню, от которого осуществляется переход к последующему). Однако этот подход неформализован и поэтому легко может быть подвергнут критике. Процедура полезна, но только как предварительный анализ результата разбиения.

Такой визуальный анализ дендрограммы чрезвычайно затруднен при:

- 1) большом количестве рассматриваемых объектов;
- 2) неявной выраженности структуры данных.

Кластерный анализ, прежде всего, решает задачу внесения структуры в данные, а если данные по сути своей однородны, то, не проанализировав корректность определенного числа кластеров, можно допустить грубую ошибку в дальнейших статистических исследованиях.

Существуют различные формальные подходы, облегчающие процедуры определения предпочтительного числа кластеров. Эти подходы называются правилами остановки. Миллиган и Купер (1985) исследовали более тридцати из них. Рассмотрим те, которые были определены как лучшие и продемонстрируем их на примере.

- 1) Duda и Hart (1973) предложили критерий, основанный на следующем отношении:

$$F_1 = \frac{w[2]}{w[1]}, \quad (1)$$

где  $w[2]$  — сумма квадратов внутрикластерных расстояний в случае, когда данные распределены по двум кластерам;

$w[1]$  - сумма квадратов внутрикластерных расстояний в случае одного кластера.

Гипотеза о существовании единого кластера однородных данных отвергается, если значение критерия  $F_1$  меньше, чем критическое значение, которое рассчитывается по следующей формуле

$$1 - 2/(\pi p) - z_{(1-\alpha)} \sqrt{2(1 - 8/(\pi^2/p)) / (np)}, \quad (2)$$

где  $n$  — число объектов для классификации;

$p$  — число признаков;

$z_{1-\alpha}$  — квантиль стандартного нормального распределения уровня  $(1-\alpha)$ .

Этот тест предназначен для ответа на вопрос, есть ли вообще структура в данных или они однородны.

2) Beale (1969) предложил использовать для решения этой же задачи другую статистику

$$F_2 = \left( \frac{w[1] - w[2]}{w[2]} \right) / \left( \frac{n-1}{n-2} \cdot 2^{2/p} - 1 \right), \quad (3)$$

где  $w[2], w[1], n, p$  — определены также как и критерия (1).

Данный критерий имеет  $F$  распределение с  $(p, p(n-2))$  параметрами. Основная гипотеза состоит в постулировании одного кластера, и она отвергается, если значение критерия больше критического уровня статистики.

Хотя оба эти критерия решают дилемму между 1 и 2 кластерами, они не обязательно применяются только для определения необходимости последнего шага иерархической агломеративной процедуры (объединения всех в один кластер). Данные критерии могут применяться и на предыдущих шагах агломеративной процедуры для обоснования объединения 2-х кластеров в один (относительно подмножества исходных данных).

3) Calinski и Harabasz (1974) предложили следующий критерий

$$F_3 = \frac{\text{trace}(\mathbf{B})/(k-1)}{\text{trace}(\mathbf{W})/(n-k)}, \quad (4)$$

где  $\mathbf{B}, \mathbf{W}$  — матрица межкластерных и внутрикластерных сумм квадратов расстояний;

$k$  — число кластеров.

Максимальное значение критерия указывает на наиболее вероятное число кластеров. При проведении иерархического кластерного анализа на каждом шаге определяется значение критерия (для  $k=1, n$ ) и находится число кластеров  $k$ , при котором  $F_3$  оптимально (максимально).

Данный критерий часто называют глобальным правилом остановки (Гордон(2001)). Предыдущие два критерия (1), (3) называют локальными правилами, так как они оценивают корректность остановки при  $k=1$ .

И те, и другие процедуры имеют ограничения и недостатки. Недостаток глобальных правил остановки в том, что они не определены при  $k=1$ , то есть не исследуется оптимальность при отсутствии разбиения, а значит, не решается вопрос — а должны ли вообще разбиваться данные?

Локальные правила остановки помимо ограниченности сравнений имеют еще недостаток — необходимость задания уровня значимости, а значит, постулируют возможность наличия ошибки и зависимость результата от неизвестных свойств набора данных.

Также можно отметить ограниченность приведенных выше критериев в применении только к иерархическим процедурам.

### **Пример применения правил остановки**

В качестве примера рассмотрим дендрограмму кластеризации методом полной связи 15 Европейских стран по уровню развития транспорта. Кластеризация проводилась в пакете STATISTICA/WIN [3].

Для классификации были использованы переменные, характеризующие уровень развития грузоперевозок различными видами транспорта (длина железнодорожных путей, количество перевезенных грузов за 1997 год и т. д.). [2].

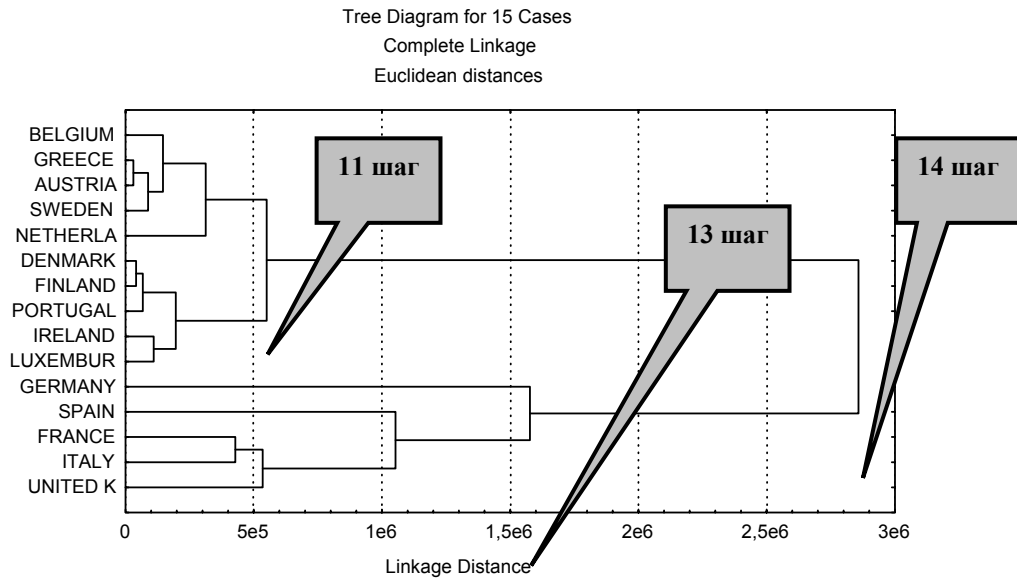


Рис. 1. Разбиение стран ЕС методом полной связи (Complete linkage)

Как видно из дендрограммы в результате кластеризации можно выделить две группы (два кластера) стран или три. В первый кластер вошли страны: Бельгия, Греция, Австрия, Португалия, Швеция, Нидерланды, Дания, Финляндия, Ирландия, Люксембург. Во второй кластер вошли: Германия, Испания, Франция, Англия, Италия. При выделении трех кластеров Германия образует отдельный кластер. Это может соответствовать реальной ситуации — уровень развития транспорта в Германии наиболее высок.

Для применения правил остановки были разработаны скрипты на языке STATISTICA BASIC для пакета STATISTICA/WIN.

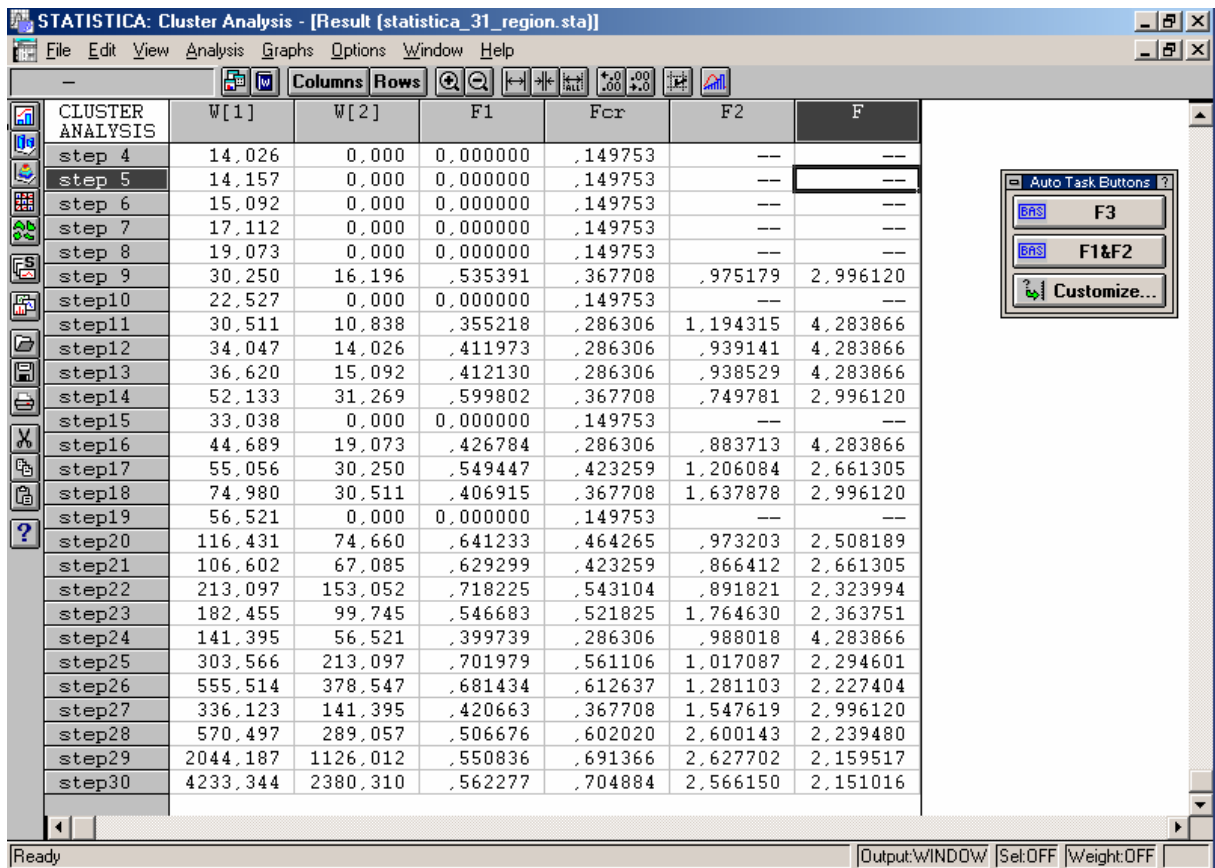


Рис. 2. Результаты работы скриптов в пакете STATISTICA/WIN

1) Рассмотрим последний шаг классификации (см. 14 шаг на дендрограмме). Этот шаг соответствует объединению двух кластеров в один. С помощью критерия Duda и Hart, рассчитываемого по формуле (1) оценим обоснованность одного кластера. Критерий равен  $F_1 = 0,2958$ , а его критическое значение, рассчитанное по формуле (2)  $F_{кр} = 0,7194$  при 5% уровне значимости. Это говорит об отвержении гипотезы о едином кластере (однородности 15 стран) и о предпочтительности гипотезы о разделимости их на два кластера.

Проанализируем теперь предпоследний шаг (13 шаг на дендрограмме). Он соответствует объединению Германии и кластера, состоящего из Испании, Франции, Англии и Италии. Опять с помощью критерия (1) оценим обоснованность этого объединения. Критерий равен  $F_1 = 0,6871$ , а его критическое значение при 5% уровне значимости равно  $F_{кр} = 0,5605$ . Такое соотношение говорит о принятии гипотезы о едином кластере (однородности 5 стран) и об обоснованности их объединения на этом в кластер.

Также интерес представляет 11 шаг на дендрограмме — объединение кластеров — первого, состоящего из стран: Бельгия, Греция, Австрия, Швеция, Нидерланды и второго, состоящего из стран: Дания, Финляндия, Португалия, Ирландия, Люксембург. Попробуем оценить правомочно ли это объединение? Критерий равен  $F_1 = 0,7079$ , а его критическое значение  $F_{кр} = 0,6706$ . Такое соотношение также говорит о предпочтительности гипотезы о едином кластере (однородности стран этих двух групп) при 5% уровне значимости и об обоснованности их объединения в кластер. На рис. 3 приведены значения критерия и его критического уровня для всего шагов дендрограммы. Конечно, смысл имеет рассмотрение только части из них при таком небольшом объеме исходных данных.

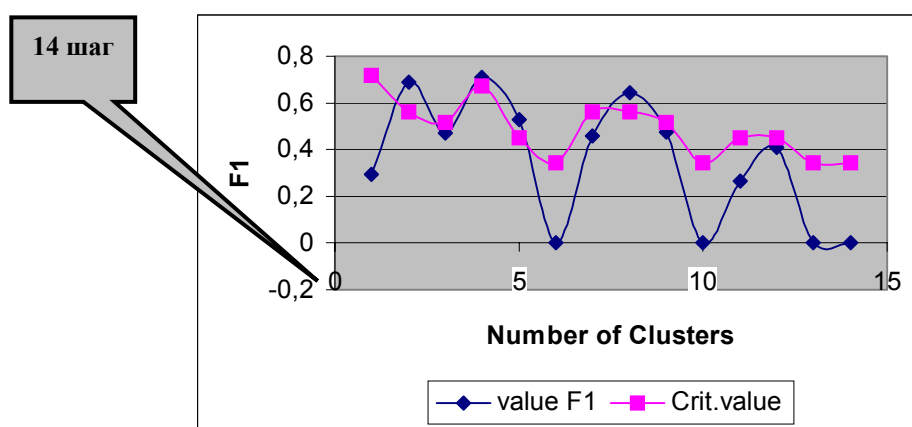


Рис. 3. Значения критерия  $F_1$  для дендрограммы, приведенной на рис.1

2) Рассчитаем для тех же шагов значение второго локального критерия  $F_2$ . Для последнего шага кластеризации значение критерия  $F_2 = 10,0433$ . Критическое значение определяется как квантиль распределения Фишера уровня 0,95 с числом степеней свободы 10 и 130  $F_{0,95}(10,130) = 1,9042$ . Такое соотношение говорит об отвержении основной гипотезы о едином кластере при 5% уровне значимости. Для предпоследнего 13 шага соотношение другое: значение критерия  $F_2 = 0,8566$  и критический уровень  $F_{0,95}(10,40) = 2,077$ . Этот результат подтверждает выводы, основанные на применении предыдущего критерия, — обоснованность объединения 5 стран в единый кластер.

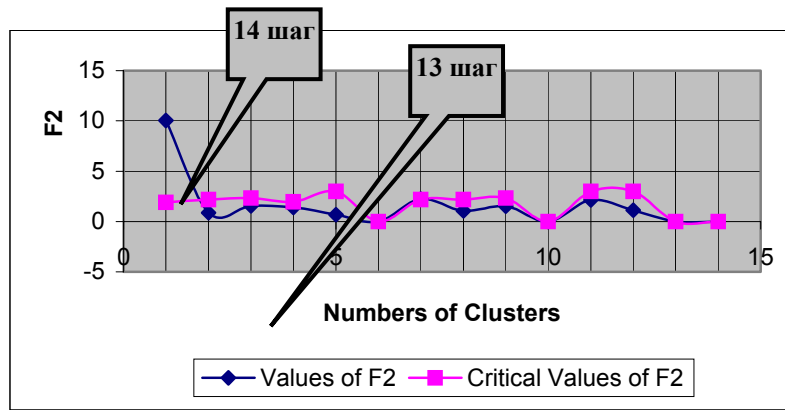


Рис. 4. Значения критерия  $F_2$  для дендрограммы, приведенной на рис.1

3) Рассчитаем значения третьего критерия — глобального по формуле (4). Так как данный критерий не имеет смысла применять для малого набора данных, то используем другой пример. На рис. 5 приведена классификация 31 региона России по уровню развития малого бизнеса. Использовались показатели развития региона: доля экономически активного населения, количество студентов Вузов, инвестиции, расходы бюджета на производство, капитальные вложения, количество банков в регионе и т. д.

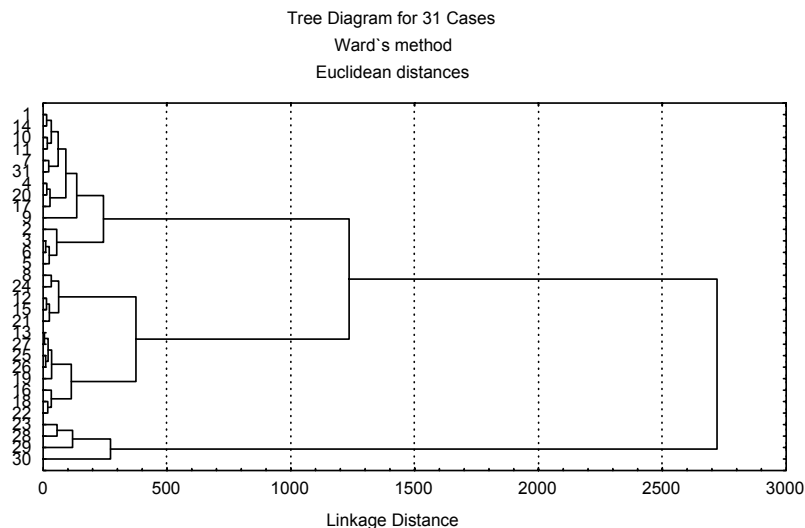


Рис. 5. Классификация регионов России по уровню развития малого бизнеса (метод Уорда)

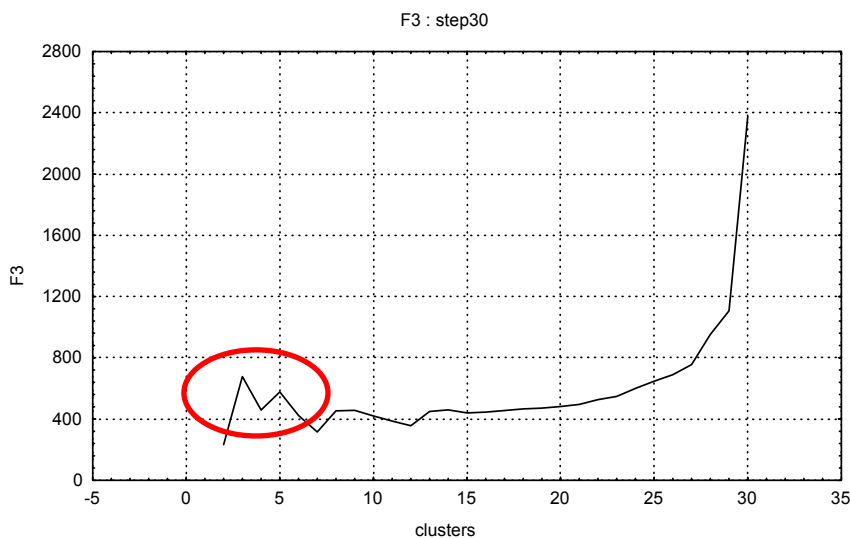


Рис. 6. График значений критерия  $F_3$  для дендрограммы, приведенной на рис.5

Как видно максимальное значение критерий имеет при числе кластеров равном трем  $F_3 = 677,39$ . (Значения для числа кластеров после 10 не рассматриваются). Если бы мы ограничились анализом только дендрограммы (рис. 5), то остановились бы скорее всего на двух кластерах, хотя решение с тремя кластерами тоже просматривалось как вероятное.

### **Основные результаты:**

1) Разработаны 3 скрипта в пакете STATISTICA/WIN, реализующие:

- процедуры расчета значений локальных правил остановки - критериев F1 и F2, их критических значений,
- процедуру расчета критерия F3 (глобального правила) и построения графика зависимости его значения от количества кластеров.

Данные скрипты могут быть применены для анализа результатов кластеризации любого набора данных в пакете STATISTICA/WIN.

2) Продемонстрировано применение правил остановки на примерах.

## **Литература**

1. Milligan G.W. and Cooper M.C. An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, 1985, 50, 159-179p.
2. The statistical guide to Europe. Data 1989 – 99 . EUROSTAT YEARBOOK 2001
3. STATISTICA ('99 Edition). Quick Reference, 1999. USA, StatSoft.
4. Afifi A.H., Clark V. "Computer Aided Multivariate Analysis". London: Chapman & Hall, 1996, 455 p.
5. Everitt B. "Cluster Analysis", 2<sup>nd</sup> edn. Wiley, NewYork, 1993, 283 p.
6. Gordon A.D. "Classification", 2<sup>nd</sup> edn. Chapman & Hall, NewYork, 1999, 256 p.