Функциональное уравнение (4) на множество двухточечных инвариантов группы преобразований (6) двумерного многообразия $M_2 \subset R^2$ запишется в следующем виде:

$$f(ax_i + c,\quad by_i + d,\quad ax_j + c,\quad by_j + d) = f(x_i, y_i, x_j, y_j), \tag{7}$$

решение которого сводится к последовательному дифференцированию по координатам соответствующих точек и решению системы функционально-дифференциальных соотношений.

Теорема. Каждый двухточечный инвариант трехпараметрической группы преобразований двумерного многообразия $M_2 \subset R^2$

$$x' = ax + c,\quad y' = by + d, \tag{8}$$

где $a^m b^m = 1\ (m, n \in \mathbb{R}^n, m \neq 0, n \neq 0, m \neq n)$, совпадает с точностью до гладкого преобразования с метрической функцией симплициальной плоскости и задает на нем феноменологически симметричную ранга 4 двумерную геометрию.

В работе установлено, что каждый двухточечный инвариант группы движений симплициальной плоскости с точностью до гладкого преобразования $\psi(f) \to f$ совпадает с метрической функцией.

Библиографический список

1. Богданова Р.А. Группа движений симплициальной плоскости как решение функционального уравнения // Вестник Томского государственного университета. Математика и механика. – 2014. – № 4(30). – С. 5–13.

2. Михайличенко Г.Г. О групповой и феноменологической симметриях в геометрии // Докл. АН СССР. – 1983. – Т.269, № 2. – С. 284 – 288. (Michailichenko, G.G. On group and phenomenological simmetries in geometry / G.G. Michailichenko // Soviet Math. Dokl. – 1983. – V.27, №2. – P. 325–326.)

3. Михайличенко Г.Г. Двумерные геометрии. – Барнаул: Изд-во БГПУ, 2004.

4. Кулаков Ю.И. Теория физических структур. – М.: Доминико, 2004.

5. Кулаков Ю.И. Геометрия пространств постоянной кривизны как частный случай теории физических структур // Докл. АН СССР. – 1970. Т. 193, №5, С. 985–987.

6. Богданова Р.А. Группа движений симплициальной плоскости как решение функционального уравнения // Вестник Томского государственного университета. Математика и механика. – 2014. – №4(30). – С. 5–13.

УДК 519.23

Критические точки распределения ледж-коэффициента

И.Ю. Бойко, С.В. Дронов
АлтГУ, г. Барнаул

Имея в наличии бинарный и числовой показатели, хочется сделать вывод о наличии связи между ними, а также ее силе.

Такие связи на практике встречаются в медицине, где бинарная переменная указывает наличие или отсутствие заболевания, а числовая – медицинский показатель, например, уровень лейкоцитов в крови. То есть, пока числовая переменная находится в определенных границах *[a,b]* пациент здоров, иначе болен. Идеальная картина связи – это индикатор отрезка *[a,b]*.

Такой вид связи мы называем связь типа «ступенька». Он впервые изучался в [1–2].

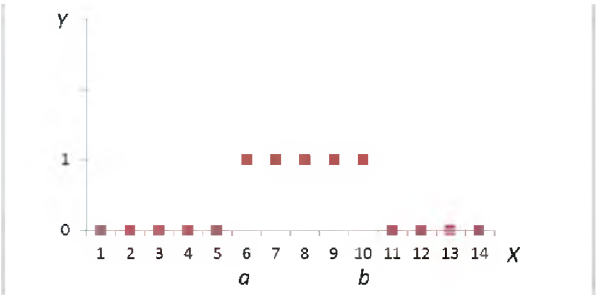

Рисунок 1 – Идеальная картина связи типа «ступенька»

Эта функция в силу своей нелинейности не выявляется при использовании классических методов (не аппроксимируется прямой линией).

В предшествующих работах [1, 3] был введен коэффициент *L*E, позволяющий численно оценить силу такого типа связи. Пусть X, Y – связанные выборки объема $n = k + m$, причем Y состоит из

k нулей и m единиц. Будем считать, что они упорядочены по возрастанию элементов X. Это позволяет заменить эти элементы их рангами и изучать только значения Y.

Критерием качества связи является так называемое число ошибок (то есть численная мера того, насколько наблюдаемая картина отличается от графика индикатора)

$$S(Y) = \min_{a,b} \sum_{j=1}^{n} (y_j - Y_{a,b}(j))^2 ,$$

где $Y_{a,b}(X) = \begin{cases} 1, & a \le X \le b, \\ 0 & \text{иначе.} \end{cases}$ – индикатор отрезка [a,b].

Определение. *Коэффициент L_E, вычисляемый по формуле*

$$L_E(X,Y) = 1 - \frac{S(Y)}{\overline{S}},$$

где $\overline{S} = \begin{cases} k-1, & k < m+1, \\ m, & k \ge m+1, \end{cases}$ назовем ледж-коэффициентом (от ledge – ступенька).

Этот коэффициент по заданной цепочке нулей и единиц позволяет оценивать силу связи типа «ступенька».

Нас интересует, насколько большим должен быть ледж-коэффициент для конкретной бинарной цепочки, чтобы мы могли говорить о наличии статистически значимой связи. Предположим, что связи рассматриваемого типа нет вообще, то есть бинарная цепочка формируется случайным образом. При этом может оказаться так, что ледж-коэффициент, рассчитанный по ней, окажется большим, что приведет к ложному заключению о наличии связи.

Чтобы этого не произошло, предполагая, что цепочка случайна, находят некоторое критическое значение $L_E(\varepsilon, n)$, которое может быть превышено лишь с малой вероятностью ε. Тогда, согласно обычным статистическим процедурам, если по выборочным данным LE окажется больше критического $L_E(\varepsilon, n)$, то наличие связи признается статистически подтвержденным на уровне доверия 1– ε.

Приступим к поиску такого критического значения для ледж-коэффициента. Для него должно выполняться

$$P(L_E > L_E(\varepsilon, n)) \le \varepsilon.$$

Примем $\varepsilon = 0.05$ и далее ограничимся этим случаем.

Приведем алгоритм нахождения критического значения L_E.

Зададим объем выборки n. Рассматривая последовательно каждую из $N = 2^n$ бинарных цепочек, считаем для нее ледж-коэффициент. Затем упорядочиваем этот набор по убыванию значений L_E. Выделим в нем подмножество $F(n)$ из первых $\lfloor \varepsilon \cdot N \rfloor$ элементов (где символ $\lfloor\ \rfloor$ означает округление вниз до ближайшего целого). То есть в нашем случае, мы выбираем 5% цепочек с самой сильной связью.

Значение последнего (т.е. минимального) элемента, попавшего во множество $F(n)$ и будет искомым критическим значением ледж-коэффициента $L_E(\varepsilon, n)$ для заданного n и ε.

Существует и другой способ поиска критического значения L_E. Рассмотрим его на примере цепочек без ошибок. Их количество обозначим $Y_0(n)$. Для них $L_E = 1$. Такие цепочки точно попадут в $F(n)$, и если выполнено

$$Y_0(n) \ge F(n), \qquad (1)$$

тогда при заданных n и ε $L_E(\varepsilon, n) = 1$. Довольно понятно, что при заданных k, m $Y_0(k,m) = m+1$. Просуммируем $Y_0(k,m)$ по всем m.

Таким образом, получим формулу по нахождению числа всех возможных цепочек без ошибок при заданном n.

$$Y_0(n) = \sum_{m=0}^{n-1} (m+1) = \left(\frac{n+1}{2} \right) \cdot n ,$$

Решая неравенство $F(n) \le Y_0(n)$, получим, что $L_E(\varepsilon, n) = 1$ для $0 < n \le 10$.

Можно использовать этот способ нахождения критического значения ледж-коэффициента и далее, но получение формул для количества бинарных цепочек с r ошибками $Y_r(k,m)$ при $r \ge 2$, оказалось достаточно трудоемким, поэтому от дальнейшего исследования было решено отказаться.

На основе изложенного алгоритма была разработана компьютерная программа на языке Java, вычисляющая критические значения ледж-коэффициента $L_E(\varepsilon, n)$. Она, в том числе, подтвердила полученный выше теоретический результат для $0 < n \le 10$. С помощью этой программы создана таблица.

Таблица – Критические значения ледж-коэффициента для малых n

n	$L_E(\varepsilon, n)$	n	$L_E(\varepsilon, n)$	n	$L_E(\varepsilon, n)$	n	$L_E(\varepsilon, n)$	n	$L_E(\varepsilon, n)$
1	1.0	6	1.0	11	0.86	16	0.73	21	0.67
2	1.0	7	1.0	12	0.83	17	0.73	22	0.67
3	1.0	8	1.0	13	0.8	18	0.7	23	0.67
4	1.0	9	1.0	14	0.78	19	0.69	24	0.65
5	1.0	10	1.0	15	0.75	20	0.67	25	0.65

Библиографический список

1. Дронов С.В., Петухова Р.В. Один вид связи между номинальной и бинарной переменными // Известия АлтГУ. – 2010.– Вып. 1/2 (65).– С. 34–36.

3. Мирмоминов Р.М. Исследование крайних случаев для оценки степени связи типа «ступенька» // Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования : сб. научных статей международной конференции, Барнаул, 11–14 ноября 2014 г. – Барнаул: Изд-во Алт. ун-та, 2014. – С. 171–173.

3. Дронов С.В. , Бойко И. Ю. Метод оценки степени связи бинарного и номинального показателей // ПДМ. – 2015. – № 4(30). –С. 109–119.

УДК 519.25

Post-hoc оценка силы кластерной связи

Е. А. Евдокимов
АлтГУ, г. Барнаул

Интересной проблемой для специалиста любой области науки, где проводится кластерный анализ или классификация, может оказаться выявление неявных связей между показателями, задействованными в исследовании. Обнаружение таких связей может способствовать получению неожиданных результатов в своей области исследования, а также привести к понижению размерности задачи, то есть уменьшению количества параметров, которые стоит подробно изучать: если, например, удалось выявить сильную связь между какими-то двумя или несколькими показателями, то можно заменить их одним, универсальным. Такой подход позволит упростить сбор данных в дальнейшем, ускорить работу алгоритма кластеризации и сделать итог его работы в той или иной мере нагляднее. При этом кластерная структура множества изучаемых объектов после сокращения размерности не должна существенно измениться.

Понятие различия кластерных разбиений и коэффициент их сходства

Вслед за [1] определим расстояние между двумя кластерными разбиениями G_A , G_B одного и того же множества объектов X формулой:

$$d(G_A, G_B) = \sum_{x \in X} |A_x \Delta B_x|, \quad (1)$$

где $|A_x \Delta B_x|$ – число элементов симметрической разности тех кластеров, в которые отнесен элемент x в этих двух разбиениях.

В качестве меры сходства K_{G_A,G_B} кластерных разбиений G_B и G_A для множества из n элементов примем коэффициент, определенный как

$$K_{G_A,G_B} = 1 - \frac{d(G_A, G_B)}{n(n-1)}.$$

Этот коэффициент тем больше, чем более похожи разбиения, и лежит в диапазоне {0:1}.