УДК 517

К проблеме оцифровки кластерной переменной, согласованной с результатами post-hoc анализа

С.В. Дронов, Н.Н. Стрижов

Алтайский государственный университет (Барнаул, Россия)

On the Cluster Variable Quantification Consistent with the Results of Post-Hoc Analysis

S.V. Dronov, N.N. Strizhov

Altai State University (Barnaul, Russia)

Рассмотрим задачу изучения и классификации каких-либо объектов, заданных набором своих числовых (формирующих) показателей. Пусть наблюдаемые объекты разбиты на кластеры, а формирующие показатели упорядочены по степени их влияния на имеющуюся кластерную структуру, т.е. решена post-hoc задача кластерного анализа. Рассматривается задача оцифровки нечисловой кластерной переменной, которая для каждого из рассматриваемых объектов представляет собой обозначение того кластера, к которому данный объект принадлежит. При этом нашей целью является присвоение каждому из кластеров числовой метки таким образом, чтобы эти метки оказались наилучшим образом согласованы с post-hoc упорядочиванием формирующих показателей. В отличие от ранее предлагавшихся методов решения подобной задачи указанное присвоение происходит без применения каких-либо итерационных процедур, — формулы для меток получены в аналитическом виде. Обсуждаются различия предлагаемого и существовавших ранее методов, даются некоторые рекомендации по переводу получающихся меток в целочисленные. Высказываются предложения по возможному использованию результатов производимой оцифровки, в том числе и в задачах доказательной медицины. Приведен пример подобного использования к обработке реальных данных медицинского обследования.

Ключевые слова: оцифровка, кластерное разбиение, post-hoc анализ показателей.

Let us consider a task of studying some objects defined by a certain collection of their numerical (nominal) characteristics. These characteristics are said to be forming. We assume that the objects in view were already divided into clusters and the forming characteristics were ordered by the degree of their influence on the existing cluster structure of the objects. It means that so-called post-hoc problem for the cluster analysis is solved. We deal with a quantification problem for the cluster variable. Generally speaking, this variable is nonnumeric and represents some symbol for the cluster to which the object belongs. Our purpose is the assigning of some numerical label to each cluster in such a way that the labels assigned would be coordinated with the post-hoc ordering of the forming characteristics in the best possible way. In contrast to the existing methods of solving such a problem, the assignment of the labels is organized in a direct way, no iterative procedures of any kind are used. Exact formulae for the best possible labels are obtained and theoretically confirmed. We discuss the differences between the proposed and preexisting methods. Some recommendations for the transfer of the resulting labels to integer ones are provided. Suggestions have been made on the possible use of the results produced by the quantification, including problems in evidence-based medicine. An example of such processing of the real data of medical examination is included and discussed as well.

Key words: quantification, cluster partition, post-hoc analysis of variables.

DOI 10.14258/izvasu(2015)1.2-19

Рассмотрим n объектов $A_1, ..., A_n$, заданных p числовыми показателями $X_1, ..., X_p$. Заранее известно разбиение множества объектов на m кластеров. Предполагается, что это разбиение строится в соответствии со значениями $X_1, ..., X_p$. Сначала присвоим последовательные номера от 1 до m этим кластерам произвольным образом. Пусть $N(j)$ — множество номеров объектов, попавших в j-й кластер. Допустим, что $N(j)$ состоит из n_j элементов, $j = 1, .., m$.

Теперь поставим каждому объекту в соответствие номер того кластера, которому он принадлежит. Этим определено отображение f из $\{1, ..., n\}$ на множество всех кластеров. Фактически у каж-

дого изучаемого объекта появилась новая нечисловая характеристика (номера кластеров здесь определяют не числовые их значения, а только порядок расположения). У j-го объекта эта характеристика равна $f(j)$, и мы называем ее кластерной переменной.

Займемся оцифровкой этой переменной. Это означает выбор неких действительных чисел $\alpha_1, ..., \alpha_m$ так, чтобы можно было считать, что $f(j) = \alpha_i,\ j \in N(i),\ i = 1, ..., m$. Для корректной постановки задачи нам нужен некоторый критерий качества получающегося решения. Обратимся, например, к медицинской интерпретации рассматриваемой задачи. Пусть производится дифференциация n пациентов по m диагнозам на основании значений p показателей. Требуется упорядочить имеющиеся диагнозы по величине некоторого нового признака f. Естественным предположением в этой ситуации является то, что эксперту-практику известно, какой из рассматриваемых показателей более, а какой — менее важен для правильного диагностирования. В соответствии с этим предположим, что априори задан некоторый экспертный порядок предпочтения показателей, т. е. каждому из признаков приписан некий ранг важности. Без ограничения общности будем считать, что номера показателей соответствуют рангам их важности.

Отметим, что в самом кластерном анализе имеется естественный способ ранжирования формирующих кластеры характеристик, называемый post-hoc анализом. При этом характеристики ранжируются по степени их влияния на кластерную структуру. Подробная постановка этой задачи и подходы к решению приводятся в [1–2]. В [3], где были начаты исследования решаемой задачи, изучался подход к оцифровке кластерной переменной, при котором метки представляли собой некоторую перестановку натуральных чисел от 1 до m, а их оптимальные значения выбирались путем перебора всех таких перестановок. После фиксации какой-либо перестановки формирующие показатели ранжировались по убыванию $|\rho(X_j, f)|$ — чем больше это число, тем более важным является показатель X_j с точки зрения текущей перестановки. При этом перестановка считалась тем лучшей, чем более полученное ранжирование формирующих показателей оказывалась похожим на априорное экспертное ранжирование.

Подойдем к задаче построения меток кластеров с более просто формализуемой точки зрения. Откажемся от поиска меток в виде натуральных чисел и снабдим каждый из квадратов коэффициентов корреляции в целевой функции тем большим весом, чем более важной является соответствующий показатель в экспертном ранжировании. Точнее, если $\rho_j = \rho(X_j, f),\ j = 1, ..., p$ — коэффициенты корреляции между кластерной переменной и показателями, то, зафиксировав некоторый масштабный множитель $b \geq 1$, мы предлагаем искать $\alpha_1, ..., \alpha_m$ из условия

$$F_1(\alpha_1, ..., \alpha_m) = \sum_{j=1}^{p} b^{p-j} \rho_j^2 \ \to \max_{\alpha_1, ..., \alpha_m}. \qquad (1)$$

Впоследствии мы сможем путем подбора найти удобное значение для b. Конечно же, при работе с экспериментальными данными объект A_j задан p-мерным вектором $(x_{1,j}, ..., x_{p,j})$, а вместо теоретических коэффициентов корреляции рассматриваются их эмпирические варианты. Заметим, что без ограничения общности можно предполагать

$$\overline{X_k} = \frac{1}{n} \sum_{j=1}^{n} x_{k,j} = 0;$$

$$\overline{X_k^2} = \frac{1}{n} \sum_{j=1}^{n} {x^2}_{k,j} = 1, \qquad k = 1, ..., p. \qquad (2)$$

Дополнительно потребуем

$$\sum_{i=1}^{m} n_i \alpha_i = 0; \qquad (3)$$

$$\sum_{i=1}^{m} n_i \alpha_i^2 = n \qquad (4)$$

для искомых $\alpha_1, ..., \alpha_m$. Обозначим

$$S_k^{(i)} = \sum_{j \in N(i)} x_{k,j}; \quad i = 1, ..., m, \ \ k = 1, ..., p.$$

Тогда (2)–(4) позволяют нам переписать целевую функцию (1) в виде

$$F_1(\alpha_1, ..., \alpha_m) = \sum_{k=1}^{p} b^{p-k} \left(\sum_{i=1}^{m} \alpha_i S_k^{(i)} \right)^2. \qquad (5)$$

Рассматривая (3)–(4) как ограничения на $\alpha_1, ..., \alpha_m$, построим функцию Лагранжа

$$L(\alpha_1, ..., \alpha_m, \lambda, \mu) = F_1(\alpha_1, ..., \alpha_m) - \lambda \sum_{i=1}^{m} n_i \alpha_i - \\ - \mu \left(\sum_{i=1}^{m} n_i \alpha_i^2 - n \right).$$

Тогда для любых $t = 1, \ldots, m$

$$\frac{\partial L}{\partial \alpha_t} = 2 \sum_{k=1}^{p} b^{p-k} \left(\sum_{i=1}^{m} \alpha_i S_k^{(i)} \right) S_k^{(t)} - \lambda n_t - 2\mu n_t \alpha_t.$$

Приравнивая все эти производные к нулю, просуммируем получившиеся равенства. Привлекая (3), получаем

$$2\sum_{t=1}^{m}\sum_{i=1}^{m}\left(\sum_{k=1}^{p} b^{p-k} S_k^{(i)} S_k^{(t)}\right)\alpha_i - \lambda n = 0.$$

Но, имея в виду (2),

$$\sum_{t=1}^{m} S_k^{(t)} = n\overline{X_k} = 0, \tag{6}$$

откуда $\lambda = 0$. Итак, для $\alpha_1, ..., \alpha_m$ и μ мы имеем систему уравнений

$$\sum_{i=1}^{m}\left(\sum_{k=1}^{p} b^{p-k} S_k^{(i)} S_k^{(t)}\right)\alpha_i = \mu n_t \alpha_t, \quad t = 1, ..., m. \tag{7}$$

Эта система означает, что числа $\alpha_1, ..., \alpha_m$ являются координатами собственного вектора матрицы G с элементами

$$G_{i,j} = \frac{1}{n_i}\sum_{k=1}^{p} b^{p-k} S_k^{(i)} S_k^{(j)}, \quad i, j = 1, ..., m, \tag{8}$$

отвечающего ее собственному числу μ.

Если справедливо (7), то, складывая эти равенства по t, получаем

$$\mu\sum_{t=1}^{m} n_t\alpha_t = \sum_{i=1}^{m}\sum_{k=1}^{p}\left(b^{p-k} S_k^{(i)}\alpha_i \cdot \left(\sum_{t=1}^{m} S_k^{(t)}\right)\right) = 0$$

(использовано (6)). Поэтому, если $\mu \neq 0$, то (3) является следствием (7).

Заметим, что если мы умножим t-е уравнение системы (7) на α_t и сложим их, то, привлекая (4), получим

$$\sum_{t=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{m} b^{p-k} S_k^{(i)} S_k^{(t)}\alpha_i\alpha_t = n\mu,$$

или, если принять во внимание (5),

$$\mu = \frac{1}{n}F_1(\alpha_1, ..., \alpha_n). \tag{9}$$

Поэтому видно, что случай $\mu = 0$ приводит к минимальному значению целевой функции, а значит, мы можем его не рассматривать. Более того, (9) означает, что нам нужно максимальное собственное число G, и искомые $\alpha_1, ..., \alpha_m$ являются координатами собственного вектора $\vec{\alpha}$, который ему соответствует.

Но нужно также добиться, чтобы было выполнено (4). Для этого достаточно умножить каждую из координат полученного собственного вектора на один и тот же коэффициент. Точнее, следует положить

$$\widehat{\alpha}_t = \frac{\alpha_t\sqrt{n}}{\sqrt{\sum_{i=1}^{m} n_i\alpha_i^2}}, \quad t = 1, ..., m. \tag{10}$$

Подобное преобразование меток не влияет на справедливость (3) или (7). Мы пришли к теореме.

Теорема. *Наилучший набор цифровых меток для кластеров задается формулой (10), где $\alpha_1, ..., \alpha_m$ — координаты единичного собственного вектора матрицы G с элементами (8), отвечающего ее наибольшему собственному числу.*

Масштабный множитель b используется для более тонкой настройки метода. При его увеличении коэффициенты корреляции с меньшими номерами получают все более заметный весовой приоритет, поэтому таким путем может быть достигнуто post-hoc ранжирование показателей, совпадающее с априорным.

Рассмотрим пример обработки реальных медицинских данных 18 пациентов, разбитых на 4 кластера по степени тяжести тромбоза глубоких вен нижних конечностей. Нумерация показателей соответствует их рангам важности. В последних двух колонках таблицы 1 указан номер кластера, к которому принадлежит объект и метка этого кластера, полученная с помощью (10) при значении масштабного множителя $b = 3$.

Таблица 1

18 пациентов

№	X_1	X_2	X_3	X_4	кл-р	метка
1	57	27	9,7	43,3	1	1,177
2	25,8	11,4	11,8	26,3	1	1,177
3	26,2	11,6	5,5	29,8	1	1,177
4	53	25	6,4	40	2	−0,189
5	29,4	13,2	11	23,9	2	−0,189
6	31	14	9,7	30,9	2	−0,189
7	31	14	10,9	27,9	3	−1,483
8	43,4	20,2	6,8	50	2	−0,189
9	33	15	6,6	31,9	1	1,177
10	42,2	19,6	10,6	28,2	3	−1,483
11	29	13	17,4	13	4	−0,491
12	34,6	15,8	6,5	29	1	1,177
13	30,2	13,6	6,4	29	3	−1,483
14	27	12	9,2	40,6	3	−1,483
15	47,8	22,4	8,6	37,2	1	1,177
16	35	16	5,7	23,3	4	−0,491
17	43	20	11,1	33	2	−0,189
18	29	13	5,1	29,7	4	−0,491

Значение $b = 3$ было выбрано после нескольких попыток подбора этого множителя. Некоторые другие возможности, испробованные во время подбора, приводятся в таблице 2.

Таблица 2

Подбор масштабного множителя

b	$\rho(X_1, f)$	$\rho(X_2, f)$	$\rho(X_3, f)$	$\rho(X_4, f)$
1	−0,08 (2)	−0,08 (3)	−0,04 (4)	0,16 (1)
2	0,25 (1)	0,25 (2)	−0,15 (4)	0,18 (3)
3	0,20 (1)	0,19 (2)	−0,16 (3)	0,14 (4)
4	0,19 (1)	0,18 (2)	−0,16 (3)	0,12 (4)

В скобках у каждого значения коэффициента корреляции указан его текущий ранг. Начиная с $b = 3$ все текущие ранги совпадают с априорными. Поэтому принято решение остановиться на этих значениях. Кластеры при этом получили метки $1,177$; $-0,189$; $-1,483$ и $-0,491$ соответственно (последний столбец таблицы 1).

Если их заменить натуральными числами, руководствуясь порядком от большего к меньшему, то получим 1; 2; 4; 3. При применении к тем же данным алгоритма из [3] они получают метки 2; 4; 1 и 3, а текущие наилучшие ранги показателей при этом составляют 4; 2; 3 и 1. Наш результат явно лучше, а значит и метки кластеров, их порядок по степени тяжести заболевания и величины расстояний между ними лучше соответствуют реальной картине.

Библиографический список

1. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия Алт. гос. ун-та. — 2011. — № 1/2 (69).

2. Dronov S.V., Dementjeva E.A. A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. — 2012. — Vol. 7, № 1.

3. Бобрышева М.С., Дронов С.В. Разметка кластеров, согласованная с post-hoc ранжированием формирующих показателей // Труды молодых ученых Алтайского госуниверситета : материалы XL науч. конф. студентов, аспирантов и учащихся лицейных классов. — Барнаул, 2013. — Вып. 10.