УДК 519.23

Геометрический подход в post-hoc задаче кластерного анализа

Дронов С.В., Еськов С.Ю.
Алтайский государственный университет, г. Барнаул
dsv@math.asu.ru, eskovslava13@gmail.com

Аннотация

В работе предложен единый подход к нескольким вариантам решения задачи о квантификации кластеров уже имеющегося кластерного разбиения конечного множества. В результате применения любого из предлагаемых подходов каждый кластер получает, вообще говоря, векторные метки. Для этого применяется методика, близкая к анализу латентных классов: каждый объект или каждый признак в рамках кластера отождествляется с некоторым вектором, а из полученных векторов геометрическим методами извлекается некая общая часть, вектор, в наибольшей степени близкий к каждому из построенных векторов. Этот вектор и объявляется меткой кластера.

Ключевые слова: Кластерная переменная, квантификация кластеров, post-hoc задача кластерного анализа, латентный анализ классов.

1. Постановка основной задачи

Разбиение некоторого конечного множества U объектов на непересекающиеся части, состоящие из схожих между собой объектов, известное, как кластеризация основного множества, применяется очень часто, особенно для больших по объему данных. После удачного построения кластеризации можно продолжать исследование на основе этого решения. Подобные проблемы (по уже готовым предварительным результатам) принято называть post-hoc задачами.

Разумеется, класс разнообразных post-hoc задач очень широк. Мы будем далее работать в рамках так называемой задачи оцифровки или квантификации кластеров, а точнее, заниматься приданием полученной кластеризации числового характера, отождествляя каждый из кластеров с некоторым числом (иногда вектором), которое будем называть меткой кластера. Конечно, делать это надо так, чтобы набор полученных в итоге числовых или векторных меток отражал бы структуру имеющейся системы кластеров.

Будем считать, что каждый из изучаемых объектов задан набором своих числовых показателей $X_1, \ldots, X_p$, и далее условимся не различать его и точку в p-мерном евклидовом пространстве, координатами которой являются значения показателей, или вектором в этом пространстве, имеющим те же, что и точка, координаты. Поскольку кластеризация, очевидно, проводится исключительно на основе значений $X_1, \ldots, X_p$, то далее будем называть эти показатели формирующими. Учитывая то, что в результате кластеризации каждому объекту оказался поставлен в соответствие его кластер, мы, следуя [1], будем считать, что, таким образом, на множестве всех объектов определена кластерная переменная, и мы имеем дело с проблемой ее квантификации в расширенном смысле, ведь мы допускаем появление не только числовых, но и векторных значений этой переменной.

При решении такой задачи заранее понятно, что значения кластерной переменной должны каким-то образом отражать значения формирующих показателей или объектов

внутри каждого из кластеров. Один из современных методов интеллектуального анализа данных, получивший названия анализа латентных классов (см., например, [2,3]) как раз и решает задачи выделения тех общих черт, которые присущи представителям некоторого класса, но непосредственно не наблюдаются. В попытке применить по сути ту же методику, сформулируем цель настоящей работы.

Основная задача – отождествляя объекты и показатели внутри каждого из сформированных кластеров с некоторыми векторами, построить множество векторов, каждый из которых в определенном смысле наиболее близок ко всем векторам в одном из кластеров. Найденный вектор мы и объявим векторной меткой соответствующего кластера.

Поскольку близость векторов можно понимать по-разному, далее рассмотрим две следующие экстремальные задачи. Пусть $\vec{G}_1, \ldots, \vec{G}_n \in R^p$ – известные векторы.

Первая задача. Найти такой p-мерный вектор $\vec{F}$, что

$$Q_d(\vec{F}) \;=\; \sum_{i=1}^{n} ||\vec{G}_i - \vec{F}||^2 \;\to\; \min\,. \tag{1}$$

Вторая задача. Обозначим φ_i угол между искомым вектором и $\vec{G}_i$, а затем будем подбирать этот вектор так, чтобы

$$Q_t(\vec{F}) \;=\; \sum_{i=1}^{n} \cos^2\varphi_i \;\to\; \max\,. \tag{2}$$

Предлагаемые критерии (1), (2) по мнению авторов наиболее наглядно конкретизируют понятие вектора, $\vec{F}$ близкого к набору векторов $\vec{G}_1, \ldots, \vec{G}_n$.

2. Основные теоремы. Техническая база алгоритмов

Приступим к решению задачи (1). Используя обозначения для координат векторов

$$\overrightarrow{G}_i = (G_1^{(i)}, \ldots, G_p^{(i)}); \quad \overrightarrow{F} = (f_1, \ldots, f_p)\,,$$

перепишем критерий в этих координатах и вычислим частные производные. Далее приравняем их к нулю. Получим систему уравнений

$$\frac{\partial Q_d}{\partial f_j} \;=\; 2\sum_{i=1}^{n}\left(G_j^{(i)} - f_j\right) \;=\; 0, \quad j = 1, \ldots, p. \tag{3}$$

Решение системы (3) задается формулами

$$f_j \;=\; \frac{1}{n}\sum_{i=1}^{n} G_j^{(i)},\; j = 1, ..., p. \tag{4}$$

Учитывая, что критерий Q_d не имеет максимального значения (его величина неограниченно увеличивается при неограниченном увеличении координат $\vec{F}$), нами доказана

Теорема 1. *Решением экстремальной задачи* (1) *является вектор* $\vec{F}$*, координаты которого определены формулой* (4).

Теперь перейдем к решению задачи (2). Расписывая косинусы углов через координаты векторов, видим, что для упрощения формул удобно считать, что все входящие в задачу

векторы имеют единичную длину. В любой задаче с углами это никак не нарушает общности, поскольку каждый вектор можно заменить просто его направлением. В таком случае задача в координатах приобретает вид

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{p} G_j^{(i)} f_j\right)^2 \to \max_{f_1,\ldots,f_p}; \quad \sum_{j=1}^{p} f_j^2 = 1.$$

Используем метод неопределенных множителей Лагранжа. Вычисляя частные производные функции

$$L(f_1,\ldots,f_p,\lambda) = \sum_{i=1}^{n}\left(\sum_{j=1}^{p} G_j^{(i)} f_j\right)^2 - \lambda\left(\sum_{i=1}^{p} f_i^2 - 1\right),$$

получаем, приравнивая 0 каждую из них,

$$2\sum_{i=1}^{n}\sum_{j=1}^{p} G_i^{(j)} G_k^{(j)} f_i - 2\lambda f_k = 0, \quad k = 1,\ldots,p.$$

Нетрудно заметить, что получившуюся систему p уравнений с p неизвестными можно в матричной форме записать в виде

$$G\vec{F} = \lambda\vec{F}, \tag{5}$$

где G – $p \times p$-матрица с элементами

$$G_{i,k} = \sum_{j=1}^{p} G_i^{(j)} G_k^{(j)}.$$

Таким образом, мы пришли к задаче на собственные числа и единичные собственные векторы матрицы G. Из сделанных выше замечаний вытекает, что

$$Q_t(\overrightarrow{F}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p} G_j^{(i)} G_k^{(i)} f_j f_k = \left\langle G\overrightarrow{F}, \overrightarrow{F}\right\rangle,$$

Здесь $\langle \cdot, \cdot \rangle$ обозначено скалярное произведение. Поэтому, если выполнено (5) и вектор $\vec{F}$ имеет единичную длину, то

$$Q_t(\overrightarrow{F}) = \lambda\left\langle \overrightarrow{F}, \overrightarrow{F}\right\rangle = \lambda.$$

Тем самым, доказана

Теорема 2. *Решением экстремальной задачи (2) среди векторов единичной длины является собственный вектор матрицы $G = A^t A$ отвечающий ее наибольшему собственному числу, где A – $n \times p$-матрица с элементами*

$$A_{i,j} = G_j^{(i)}, \quad i = 1,\ldots,n;\ j = 1,\ldots,p.$$

Отметим, наконец, что задачу (1) можно рассматривать как упрощенный вариант задачи (2), Действительно, максимизируя сумму квадратов косинусов некоторых углов мы минимизируем сумму квадратов их синусов. Но, если принять сделанные при доказательстве теоремы 2 допущения, то косинусы углов φ_i представляют собой длины проекций искомого вектора $\vec{F}$ на направления остальных векторов. В то же время длины векторов

$\overrightarrow{G}_i - \overrightarrow{F}$, сумму квадратов которых мы минимизируем в задаче (1), как правило, не совпадают с соответствующими синусами. Но в основной массе практически важных случаев их длины будут, видимо, к этим значениям близки. При этом, перейдя к форме задачи (1), мы существенно упростим решение.

3. Два новых метода квантификации кластерной переменной

Сначала введем нужные обозначения. Пусть основное множество U состоящее из n пронумерованных объектов, каждый из которых задан набором p своих показателей $X_1, \ldots, X_p$, разбито на m кластеров. Для $i = 1, \ldots, m$ через $N(i)$ обозначим множество номеров тех объектов из множества U, которые попали в i-й кластер. Ясно, что

$$\bigcup_{i=1}^{m} N(i) = \{1, ..., n\}, \quad \sum_{i=1}^{m} n_i = n,$$

где $n_i = |N(i)|$ – число элементов i-го кластера. Будем считать, что все показатели центрированы и нормированы

$$\overline{X^{(i)}} = \frac{1}{n}\sum_{j=1}^{n} X_j^{(i)} = 0, \tag{6}$$

$$D^{(i)} = \frac{1}{n}\sum_{j=1}^{n} \left(X_j^{(i)}\right)^2 = 1, \ i = 1, \ldots, m. \tag{7}$$

Для каждого из показателей определим

$$S_j^{(i)} = \sum_{k \in N(i)} X_k^{(i)}, \quad j = 1, \ldots, m; \ i = 1, \ldots, p -$$

суммы значений этого показателя на всех объектах соответствующего кластера. Тогда из (6) вытекает, что

$$\sum_{j=1}^{m} S_j^{(i)} = 0, \quad i = 1, \ldots, p. \tag{8}$$

Значения кластерной переменной f, одинаковое на всех объектах j-го кластера, обозначим $f_j, j = 1, \ldots, m$. Будем также считать их центрированными и нормированными:

$$\overline{f} = 0 \ \Rightarrow \ \sum_{j=1}^{m} n_j f_j = 0; \quad Df = 1 \ \Rightarrow \ \sum_{j=1}^{m} n_j f_j^2 = n. \tag{9}$$

Первый из предлагаемых нами методов мы назвали латентно-показательным. В основе его лежит представление о том, что внутри каждого кластера значения всех формирующих показателей должны быть близки, а то общее, что есть в наборе этих значений внутри кластера и может считаться (векторным) значением кластерной переменной на нем.

Для квантификации этим методом в i-м кластере рассмотрим p n_i-мерных векторов, за координаты которых приняты значения одного из формирующих показателей для каждого из объектов, попавших кластер, $i = 1, \ldots, m$. Тогда вектор, наиболее близкий к построенным p векторам, будет рассматриваться как векторная метка соответствующего кластера. В зависимости от того, в каком смысле понимается близость векторов, для поиска этих меток используются теоремы 1 или 2.

Отметим, что смысл каждого такого вектора-метки в том, что в нем собраны значения нового, латентного или, иначе, универсального показателя, который выделен как коррелят

Таблица 1

Два двумерных кластера

объект	X_1	X_2	Кластер	объект	X_1	X_2	Кластер
1	0.698	-0.801	1	7	0.970	-0.712	1
2	0.948	-0.990	1	8	-0.814	1.082	2
3	-0.980	0.974	2	9	-0.855	0.996	2
4	-0.910	1.722	2	10	-1.729	0.493	2
5	0.839	-0.985	1	11	1.157	-0.889	1
6	0.676	-0.889	1				

(общая часть) формирующих факторов. Следовательно, построив такой вектор, мы свели задачу к одномерной – каждый объект связан с соответствующей координатой нашего построенного вектора. Поэтому задача присвоения меток кластерам перешла в ее одномерный вариант – у каждого объекта в каждом из кластеров заданы значения его (одномерной) координаты. Далее годится любой метод получения одномерных меток. Например, можно взять в качестве метки кластера среднее значение координат вектора, выделенного в этом кластере. Или, считая координаты значениями единственного формирующего показателя, построить систему меток, наиболее согласованную с этими значениями. методом внутреннего согласования из статьи [1].

Второй метод назовем латентно-объектным методом. Идея здесь фактически та же, но общие черты выделяются не у значений показателей, а у самих объектов, составляющих кластер. Векторной меткой кластера с этой точки зрения будет типичный объект кластера, его центр в определенном смысле. В этом методе объекты, попавшие в i-ый кластер, рассматриваются, как вектора p-мерного пространства. На их основе ищется вектор, близкий к ним с помощью теорем 1 или 2. Латентно-объектный метод формирует новый (средний) объект в каждом кластере, который и объявляется меткой кластера. При построении одномерных меток фактически всегда производится проецирование векторных меток (или, иначе, искусственных объектов) на некоторую прямую и объявлении положения точки-проекции каждого из искусственных объектов на этой прямой одномерной меткой. Известно, что проекции точек наиболее разбросаны на направление первой главной компоненты данных. Поэтому по координатам полученных векторных меток далее предлагается построить их ковариационную матрицу размером $p \times p$, а затем найти наибольшее собственное число этой матрицы и направление соответствующего собственного вектора. Далее числовой меткой каждого кластера объявляем проекцию соответствующей векторной метки на это направление.

Рассмотрим числовой пример, данные для которого заимствованы из [1] (см. таблицу 1), Здесь приведены уже нормированные значения двух формирующих показателей 11 объектов, и у каждого указан номер одного из двух кластеров, к которому он относится.

Рассчитаем векторные метки двумя методами. Сначала используем латентно-объектный метод. Двумерные метки кластеров, полученные по теореме 1, равны (0,881; -0,878) и (-1,058; 1,053). Таким образом, принимая за одномерные (числовые) метки кластеров средние значения координат каждого из векторов, мы присваиваем первому кластеру метку 0,002, а второму -0,002.

Теперь используем теорему 2 для поиска тех же меток. Матрицы G для кластеров имеют вид

$$G_1 = \begin{pmatrix} 4.83 & -4.64 \\ -4.64 & 4.68 \end{pmatrix}, \quad G_2 = \begin{pmatrix} 6.17 & -5.11 \\ -5.11 & 6.32 \end{pmatrix},$$

их наибольшие собственные числа 9.4 (собственный вектор, он же метка первого кластера (-0,71; 0,70)) и 11,35 (векторная метка второго кластера (0,70;- 0,71)). Собственный век-

тор, отвечающий максимальному собственному числу ковариационной матрицы координат равен (-0,7074; 0,7069). Числовые метки кластеров теперь вычисляются как скалярные произведения последнего векторы на их ранее полученные векторные метки. Результатом являются метки, равные 1 и -1 соответственно.

Перейдем к латентно-показательному методу. Используя способ расчета меток, предложенный в теореме 1, получаем векторные метки для кластеров: для первого (0.346,0.447,0.42,0.361,0.388,0.472), а для второго (-0.41,-0.554,-0.398,-0.389,-0.464).

Вычислив средние значения координат каждого из вычисленных векторов, приходим к числовым меткам 0.41 и -0,44.

4. Обсуждение результатов и выводы

В работе предложены два новых способа придания числовых или векторных значений кластерной переменной, т.е. решений задачи квантификации кластеров. При этом для каждого из методов теоремы 1 и 2 работы дают альтернативные варианты организации процесса вычисления этих методов, первая из них обладает очевидным преимуществом с точки зрения простоты вычислений. Разумеется, результат при ее применении получается менее точным, но сравнение значений меток на примере показывает, что погрешность при этом незначительна.

Сравнение двух предложенных способов на примере из [1] показывает, что принципиальной разницы между полученными результатами нет, иным является лишь масштаб – величина расстояний между значениями кластерной переменной на первом и втором кластере.

Заметим также, что оба предложенных метода можно рассматривать как обобщение метода статьи [1], – там формулы нашей теоремы 2 использовались однократно ко всему имеющемуся набору данных. Мы же предлагаем пользоваться этими формулами отдельно в рамках каждого из имеющихся кластеров. Метки кластеров (значения кластерной переменной), полученные при этом в цитированной статье, будут еще более близки к полученными нами, если обратить внимание на следующее обстоятельство: в [1] утверждалось, что ограничение (9) однозначно определяет длину вектора, координатами которого являются значения кластерной переменной. Но, в действительности, оно приводит лишь к тому, что вектор $(f_1\sqrt{n_1/n}, \ldots, f_m\sqrt{n_m/n}$ имеет единичную длину. Результатов [1] это не нарушает, но, полученные в статье метки для i-го кластера нужно в итоге умножить на $\sqrt{n_i/n},\ i = 1, \ldots, m$.

По мнению авторов, примененный при обосновании предложенных методов геометрический подход, близкий к популярной сегодня технике анализа латентных классов, может оказаться полезным и при решении других post-hoc задач кластерного анализа.

Список литературы

1. Dronov S.V., Sazonova A.S. Two approaches to cluster variable quanti-fication // Model Assisted Statistics and Applications. — 2015. — Vol. 10. — P. 155–162.

2. Rindskopf D. Latent Class Analysis // The SAGE Handbook of Quantitative Methods in Psychology. — N.Y. : Sage, 2009. — P. 226–244.

3. Дронов С.В., Шеларь А.Ю. Латентный кластерный анализ для случая двух кластеров // МАК: “Математики - Алтайскому краю”: сборник трудов всероссийской конференции по математике. — Барнаул : Изд-во Алтайского госуниверситета, 2018. — С. 23–26.