УДК 519.23

Иерархическая квантификация кластерной переменной

Болгов А.В., Дронов С.В.
Алтайский государственный университет, г. Барнаул
bolgov.c@gmail.com, dsv@math.asu.ru

Аннотация

В статье предлагается и обосновывается способ присвоения числовых меток (квантификация) кластерам, связанный с их построением на основе агломеративного кластерного алгоритма, рассматриваются проблемы, которые могут возникнуть при такой квантификации, в частности, возникновение числовых меток кластеров, значения которых противоречат их естественному порядку (инверсии). Предложен новый вариант алгоритма, при котором подобные инверсии не возникают.

Ключевые слова: квантификация, кластерная переменная, агломеративный кластерный алгоритм, метод Варда.

1. Постановка задачи

При работе с большими объемами данных часто оказывается удобным разбить множество всех изучаемых объектов на непересекающиеся группы (кластеры) объектов, относительно схожих между собой. При этом, как правило, можно ввести дополнительный искусственный показатель, постоянный в рамках каждой группы. В работе [1] этот показатель был назван кластерной переменной. Выбор значений такой переменной для каждой из групп принято называть задачей квантификации кластеров, а конкретные значения ее внутри групп – их числовыми метками. Задача квантификации кластерной переменной чаще всего решается уже по подготовленной заранее системе кластеров, т.е. по результатам работы некоторого кластерного алгоритма.

На сегодняшний день имеется достаточно широкое разнообразие алгоритмов кластеризации, см., например, [2–6]. Многие из них реализованы в основных статистических компьютерных пакетах, см. [7–12]. Основной алгоритм, используемый в данной работе – иерархический агломеративный кластерный алгоритм. Он подробно изложен, например, в [6, с. 51-60].

Решение задачи квантификации кластерной переменной становится все более востребованным в настоящее время, потому, что оно дает возможности:

1. Получить числовую оценку различий между кластерами.

2. Упорядочить кластеры, т.е. расположить их в том порядке, который определяет увеличение значений полученных числовых меток.

3. Включить в математическую модель кластерную переменную, что позволит в итоге сделать более адекватные выводы и даст возможность при необходимости объединить разные для каждой группы модели в единую (подобная задача решалась, например, в [13]).

Основной задачей работы является разработка нового метода квантификации кластерной переменной, связанного с агломеративным иерархическим алгоритмом. При этом, поскольку алгоритм производит построение сразу нескольких возможных кластерных разбиений, предлагается последовательное построение текущих кластерных меток непосредственно в процессе работы алгоритма.

2. Иерархический алгоритм как инструмент построения меток

В качестве основного кластерного алгоритма в данной работе возьмем агломеративный вариант иерархического алгоритма. Такое решение связано с тем, что он реализован практически в любом статистическом пакете (в частности, SPSS, Statistica, Stata, Origin, Python, R и др.). Опишем кратко классический вариант этого алгоритма.

Рассмотрим конечное множество объектов $U = \{x_1, ..., x_n\}$, обладающих набором признаков $(p_1, ..., p_m)$. Каждый объект $x \in U$ может быть представлен в виде точки или вектора $(p_1, ..., p_m)$ m-мерного евклидова пространства R^m, k-я координата которого равна значению признака p_k у этого объекта, $k \in \overline{1, m}$. Для определения степени близости объектов нам понадобится метрика. Обозначим ее d и будем называть метрикой первого уровня. Наиболее естественным, вероятно, является рассмотрение в качестве метрики d евклидовой метрики.

Предположим, что имеется кластерное разбиение $S = \{S_1, ..., S_r\}$, множества объектов U. Количество объектов, отнесенных к i-му кластеру этого разбиения, обозначим $|S_i|$. Центром кластера S_i назовем векторное среднее арифметическое его элементов:

$$Z_i = \frac{1}{|S_i|} \sum_{x \in S_i} x.$$

Нам потребуется еще одна метрика ρ, заданная на наборе всех кластеров (расстояние между произвольно выбранными двумя кластерами). Эту метрику, называемую далее метрикой второго уровня, можно определять по-разному (см., например, [2]). При этом совсем не обязательно она как-то должна быть связана с метрикой первого уровня d. Пока определим "временное" расстояние между кластерами как значение метрики первого уровня на паре их центров: $\rho(S_i, S_j) = d(Z_i, Z_j)$. Такое определение метрики второго уровня является одним из наиболее часто встречающихся. Отметим, что если каждый из кластеров состоит ровно из одного объекта, то ρ и d – одно и то же.

На старте работы алгоритма каждый элемент множества U рассматривается как отдельный кластер. Создается список всех имеющихся на данный момент кластеров. Затем происходит поиск двух наименее удаленных кластеров из списка (в соответствии с метрикой второго уровня) и формируется новый кластер, путем объединения элементов найденных кластеров. Далее, эти два кластера удаляются из списка. Описанный процесс повторяется до тех пор, пока все элементы U не оказываются в одном кластере.

В качестве простейшего способа квантификации предлагается присвоение кластеру метки, равной расстоянию между двумя образовавшими его кластерами или нулю, если этот кластер состоит из одного элемента (то есть это кластер, не менявшийся со старта алгоритма). Такой метод построения меток условимся называть стандартным относительно метрики ρ. Хотя ни в одном из известных авторам статистических пакетов описанный процесс не реализован, но он легко считывается из результатов обработки, выдаваемых используемым пакетом (под каждым из образующихся кластеров, как правило, подписано значение метрики второго уровня, при котором образуется этот кластер).

К сожалению, данный способ квантификации может приводить к нежелательным эффектам с точки зрения поставленной задачи, конкретнее, не всегда дает возможность правильно упорядочить кластеры. При использовании в качестве метрики второго уровня введенной выше "временной" метрики иногда метки двух кластеров превосходят метку кластера, полученного в результате их слияния. Это явление называется инверсией (см. [2, с. 235]). Строгое определение инверсии дается в следующем разделе.

3. Инверсии. Причины их образования

Пусть S_k, S_i, S_j – кластеры. Ситуация, когда $\rho(S_k, S_i \cup S_j) < \rho(S_i, S_j)$ называется инверсией. Если имеется инверсия, то метка кластера, образованного на более позднем шаге иерархического алгоритма, оказывается меньшей по величине, что в нашей задаче является нежелательным.

Причина возникновения инверсий кроется в том, что метрика второго уровня (которую мы временно определили как евклидово расстояние между их центрами), не вполне корректно взаимодействует с используемым алгоритмом, хотя ее использование на первый взгляд в этой задаче вполне естественно.

Рассмотрим в качестве примера инверсии три точки: $S_1 = (0, 0)$, $S_2 = (1, 1.9)$, $S_3 = (2, 0)$, составляющие основное множество U. Приближенные значения попарных расстояний между ними: $\rho(S_1, S_2) = 2.14$, $\rho(S_1, S_3) = 2$, $\rho(S_2, S_3) = 2.14$.

Три кластера на первом шаге состоят из отдельных точек. Тогда, если следовать алгоритму, сначала происходит слияние одноэлементных кластеров S_1 с S_3 и образуется новый двухэлементный кластер с центром (1, 0) с меткой 2, поскольку расстояние между ними равно 2 и меньше остальных попарных расстояний между кластерами. На втором шаге сливаются кластеры с центрами (1,0) и (1, 1.9), образуется кластер с центром (1, 0.95) с меткой 1.9, который содержит все три точки.

Таким образом, кластер, образованный на последнем шаге, имеет меньшую метку (1,9), чем ранее образованный (2), а, следовательно, возникла инверсия.

4. Метрика Жамбю-Миллигана

Тем не менее, существуют метрики, при использовании которых в качестве метрики второго уровня, инверсии не образуются. Рассмотрим так называемую метрику Жамбю-Миллигана, см. [3, с. 324]:

$$\rho(S_1, S_2) = \sum_{x \in S_1 \cup S_2} \|x - Z\|^2 - \sum_{x \in S_1} \|x - Z_1\|^2 - \sum_{x \in S_2} \|x - Z_2\|^2,$$

где S_1, S_2 – кластеры, а Z_1, Z_2, Z центры S_1, S_2 и $S_1 \cup S_2$ соответственно. В ранней работе [14], где был предложен новый метод построения кластеров, известный сегодня, как метод Варда, эта метрика фактически уже использовалась, но не была явно выделена и исследована. Таким образом, эта метрика напрямую связана с методом Варда.

Формула для вычисления метрики Жамбю-Милигана может быть записана в более коротком виде. Пусть S_1, S_2 – произвольные кластеры такие, что $|S_1| = n_1$ и $|S_2| = n_2$. Обозначим центры $S_1, S_2, S_1 \cup S_2$ через Z_1, Z_2, Z соответственно.

Лемма 1.

$$\sum_{x \in S_i} \|x - Z_i\|^2 = \sum_{x \in S_i} \|x\|^2 - n \|Z_i\|^2, \quad i = 1, 2.$$

На основании этой леммы можно получить упомянутую более простую формулу, привлекая следующее соотношение между центрами кластеров:

$$Z = \frac{n_1 Z_1 + n_2 Z_2}{n_1 + n_2}.$$

Теорема 1. *Пусть S_1, S_2 дизъюнктные подмножества основного множества $|S_1| = n_1$, $|S_2| = n_2$, Z_1, Z_2 – центры этих множеств, ρ – метрика Жамбю-Миллигана. Тогда*

$$\rho(S_1, S_2) = \frac{n_1 n_2 \|Z_1 - Z_2\|^2}{n_1 + n_2}.$$

Известен набор требований на метрику второго уровня, при выполнении которых инверсии не возникают. Приведем здесь соответствующую формулировку.

Теорема 2 ([4]). *Пусть для произвольных конечных множеств S_1, S_2, S имеет место формула*

$$\rho(S, S_1 \cup S_2) = \alpha_1\rho(S, S_1) + \alpha_2\rho(S, S_2) + \alpha_3\rho(S_1, S_2) + \\ + \alpha_4\nu(S) + \alpha_5\nu(S_1) + \alpha_6\nu(S_2) + \alpha_7|\rho(S, S_1) - \rho(S, S_2)|,$$

где $\nu(S_1)$, $\nu(S_2)$, $\nu(S)$ – числовые метки соответствующих множеств, построенные стандартным для метрики ρ методом. Если

1. $a_k \geq 0,\ k = 1,\ ... 6$;
2. $a_7 \geq -\min\{a_1, a_2\}$;
3. $a_1 + a_2 + a_3 \geq 1$,

то метрика второго уровня ρ не образует инверсий.

Опираясь частично на идеи, изложенные в [14], нам удалось дать новое простое доказательство подобной теоремы для метрики Жамбю – Миллигана.

Следующая лемма нетрудно проверяется непосредственно.

Лемма 2. *Для произвольных конечных множеств S_1, S_2, S с количествами элементов n_1, n_2, n и центрами Z_1, Z_2, Z соответственно, имеет место формула*

$$\rho(S, S_1 \cup S_2) = \alpha_1\rho(S, S_1) + \alpha_2\rho(S, S_2) + \beta\rho(S_1, S_2),$$

где

$$\alpha_i = \frac{n + n_i}{n + n_1 + n_2},\ i = 1, 2; \quad \beta = -\frac{n}{n + n_1 + n_2}.$$

При этом $\alpha_1 + \alpha_2 + \beta = 1$.

Теорема 3. *При использовании метрики Жамбю - Милигана в качестве метрики второго уровня инверсии не возникают, то есть*

$$\forall i, j,\ k:\ \rho(S_k, S_i \cup S_j) \geq \rho(S_i, S_j).$$

Доказательство. Без ограничения общности будем считать, что

$$\rho(S_k, S_i) \geq \rho(S_k, S_j).$$

Исходя из этого и используя лемму 2, получим:

$$\rho(S_k, S_i \cup S_j) = \alpha_1\rho(S_k, S_j) + \alpha_2\rho(S_k, S_j) + \beta\rho(S_i, S_j) \geq (\alpha_1 + \alpha_2)\rho(S_k, S_j) + \beta\rho(S_i, S_j).$$

На каждом шаге используемого алгоритма объединяются два ближайших кластера, следовательно, расстояние между i-м и j-м кластерами было минимальным среди расстояний для всех пар кластеров, откуда

$$\rho(S_k, S_j) \geq \rho(S_i, S_j).$$

Используя лемму 2, приходим к

$$\rho(S_k, S_i \cup S_j) \geq (\alpha_1 + \alpha_2 + \beta)\rho(S_i, S_j) = \rho(S_i, S_j),$$

что и требовалось доказать. □

5. Новый алгоритм квантификации кластерной переменной

Из теоремы, доказанной в предыдущем пункте, следует, что для успешной квантификации кластерной переменной в классическом иерархическом алгоритме достаточно заменить нашу временную метрику второго уровня на метрику Жамбю-Миллигана. При этом на каждом шаге модифицированного алгоритма значение этой метрики на двух объединяемых кластерах принимается за метку вновь образующегося кластера.

Проведем вычисления меток для точек из приведенного выше примера с инверсией в случае использования метрики Жамбю-Милигана на втором уровне. Кластеры на первом шаге: $S_1 = \{(0,0)\}$, $S_2 = \{(1,1.9)\}$, $S_3 = \{(2,0)\}$.

Сначала найдем три участвующих в вычислении центра.

$$Z = \frac{(0,0)+(1,1.9)}{2} = (0.5, 0.95), \quad Z_1 = (0,0), \quad Z_2 = (1,1.9).$$

Вычислим расстояние между S_1 и S_2 с помощью теоремы 1:

$$\rho(S_1, S_2) = \|(0,0) - (1,1.9)\|^2 = 2.305,$$

поскольку $n_1, n_2 = 1$. Аналогичным образом $\rho(S_1, S_3) = 2$, $\rho(S_2, S_3) = 2.305$.

Как видно, наименьшее расстояние (равное 2) получается между кластерами S_1 и S_3, значит, эти кластеры объединятся, и на втором шаге будет всего два кластера: $S_1 = \{(0,0),\ (2,0)\}$, $S_2 = \{(1,1.9)\}$.

Найдем новые центры кластеров:

$$Z = \frac{(0,0)+(2,0)+(1,1.9)}{3} = (1, 0.63), \quad Z_1 = \frac{(0,0)+(2,0)}{2} = (1,0), \quad Z_2 = (1,1.9).$$

Расстояние между этими кластерами равно $\rho(S_1, S_2) = 2.406$.

Видим, что метка последнего кластера при слиянии S_1 и S_2 (2,406) получилась больше, чем предыдущая метка (она равна 2), следовательно, в отличие от евклидовой метрики, инверсии нет.

6. Заключение

В работе предложена модификация широко используемого в практической статистике иерархического кластерного алгоритма, которая позволяет присвоить метки кластерам непосредственно в процессе их построения. При этом получающиеся метки упорядочивают кластеры именно в порядке их появления в алгоритме, что обеспечивает отсутствие так называемых инверсий. Невозможность их возникновения в новом алгоритме следует из доказанной теоремы 2. Первым автором написана компьютерная программа на языке Python, реализующая предложенную модификацию.

При всем многообразии используемых сегодня методов кластерного анализа изучение его возможностей и перспектив, видимо, еще далеко от завершения. В частности, решение задачи квантификации кластерной переменной методом, подстроенным под решение конкретной кластерной задачи, может давать почву для далеко идущих выводов и новой интуиции. Так, например, в работе [15], соображения, сходные с изложенными выше, использованы для решения задач распознавания образов, что дает повод надеяться на широкое применение полученных в настоящей работе результатов.

Список литературы

1. Дронов С.В., Герасимова А.С. К проблеме оцифровки кластерной переменной // Труды Всероссийской молодежной школы-семинара “Анализ, Геометрия и топология” Барнаул, 2-4 октября 2013 г. — Барнаул : ИП Колмогоров И.А., 2013. — С. 54–58. — Ч.2.
2. Romesburg H.C. Cluster Analysis for Researchers. — Morrisville, NC : Lulu.com, 2004.
3. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. — М. : Финансы и статистика, 1989.
4. Жамбю М. Иерархический кластер-анализ и соответствия. — М. : Финансы и статистика, 1988.
5. Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. Cluster Analysis. — 5th edition. — UK : John Wiley & Sons, Ltd, 2011.
6. Дронов С.В. Методы и задачи многомерной статистики: монография. — Барнаул : Изд-во Алт. ун-та, 2015.
7. IBM SPSS Statistics. — URL: http://www.ibm.com/ru-ru/products/spss-statistics.
8. StatSoft Statistica. — URL: https://statsoft.com.
9. Stata: Software for statistics. — URL: https://www.stata.com/.
10. Origin: Data Analysis and Graphing Software. — URL: https://www.originlab.com/.
11. The R Project for Statistical Computing. — URL: https://www.r-project.org/.
12. Python programming language. — URL: https://www.python.org/.
13. Жилин С.И. Решение задач дисперсионного и ковариационного анализа методом центра неопределенности // Известия АГУ. — 2011. — № 1-2(69). — С. 54–57.
14. Batagelj V. Note on ultrametric hierarchical clustering algorithms // Psychometrika. — 1981. — Vol. 46, no. 3. — P. 351–352.
15. Ward Jr. J.H. Hierarchical Grouping to Optimize an Objective Function // Journal of the American Statistical Association. — 1963. — no. 58. — P. 236–244.
16. Boutell M. R., Luo J., Shen X., Brown C.M. Learning multi-label scene classification // Pattern Recognition. — 2004. — no. 37(9). — P. 1757–1771.