УДК 519.25

Об обобщении метода латентного кластерного анализа

Шеларь А.Ю., Дронов С.В.
Алтайский государственный университет, г. Барнаул
shelaranton@gmail.com, dsv@math.asu.ru

Аннотация

В работе предложен способ развития методов латентного кластерного анализа для задачи числовой оценки степени справедливости некоторого достаточно произвольного совместного свойства L двух показателей. Рассмотрено приложение этих методов для оценки степени независимости показателей на примере реальных социологических данных. Обсуждаются отличия нового подхода от традиционных методов и его преимущества.

Ключевые слова: латентный кластерный анализ, четырехпольные таблицы, оценка степени зависимости, корреляция, множественная регрессия.

1. Вводные замечания

Алгоритмы кластерного анализа широко применяются в последнее время. С их помощью множество объектов наблюдения разбивается на группы относительно схожих между собой (кластеры). Такая схожесть определяется наборами некоторых показателей. Эти показатели в данном контексте условимся называть формирующими. Однако, довольно часто оказывается, что имеется некоторое свойство, признак объектов, которое может заменить в построении кластеров все формирующие показатели. Этот признак является, как правило, не наблюдаемым непосредственно, и может быть назван латентным кластерным признаком.

Раздел интеллектуальной обработки данных, возникший в последнее десятилетие и рассматривающий методы, связанные с выявлением и изучением латентного кластерного признака, называется латентным кластерным анализом или латентным анализом классов, следуя, например, [1]. Более того, при решении задачи латентного кластерного анализа одновременно и кластеры могут строиться так, чтобы они хорошо разделялись при правильном выборе латентного кластерного признака. Тогда фактически одновременно решается как задача кластерного, так и задача дискриминантного анализа, т.е. латентный кластерный анализ как бы соединяет в одно целое эти две ранее решаемых отдельно задачи.

Иногда поиск латентного кластерного признака фактически означает оцифровку (квантификацию) некоторого свойства, ранее рассматриваемого, как исключительно качественное. Например, объекты разбиты по признаку пригодный / непригодный. Если мы сумеем выделить латентный признак, значимо связанный с этим разбиением, то мы оцифруем свойство “пригодности” и каждому объекту поставим в соответствие число, характеризующее степень этой пригодности. Поскольку проблема перевода качественных (нечисловых) данных в цифровую форму все более востребована в связи с массовым внедрением во все сферы нашей жизни компьютерной техники, то, таким образом, любая задача в области латентной кластеризации может рассматриваться, как весьма актуальная.

Далее будем изучать только случаи, когда латентный кластерный признак можно задать при помощи формулы, аналитически, как функцию от признаков формирующих. Эту функцию будем называть латентной кластерной переменной (ЛКП), поскольку именно она

и показывает числовые значения латентного кластерного признака. Таким образом, под ЛКП мы понимаем оцифровку латентного кластерного признака. В нашей системе определений это фактически будет означать постоянство значения ЛКП на всех объектах одного кластера. Если это не так, то, определив границы значений, которые может принимать построенная переменная в рамках каждого кластера, можно заменить их все, например, некоторым средним значением. Такой подход позволяет взглянуть на решаемые задачи как на частный случай построения оцифровки (квантификации) кластерной переменной, которая, видимо, впервые была поставлена в [2] и подробно обсуждается в [3].

2. Постановка основной задачи

Использующаяся сегодня методика латентного кластерного анализа предполагает, что, после фиксирования значений ЛКП, все формирующие признаки должны оказаться независимыми, – по этому поводу см., например, [4]. Это означает, что ЛКП содержит в себе практически всю информацию, сколько-нибудь важную для объективно правильного построения кластеров, а те составляющие формирующих показателей, которые не связаны с ЛКП, на фоне кластерного разбиения представляют собой просто статистически незначимый "шум". Собственно, указанное обстоятельство и принимается в [4] за определение ЛКП. После этого извлечение латентной кластерной переменной представляет собой просто процесс её оценки тем или иным статистическим методом. Фактически именно этот метод был применен в [5].

Таким образом, исследование ЛКП оказывается тесно связанным с изучением независимости формирующих показателей, а точнее, с возможностью их преобразования так, чтобы результат как можно более походил на независимые величины. Именно этот процесс мы ниже в основном и будем иметь в виду, хотя главная задача настоящей работы состоит в приспособлении метода латентного кластерного анализа для оценки степени соответствия совокупности изучаемых показателей некоторому достаточно произвольному свойству, лишь в определенной степени похожему на независимость. При этом здесь мы ограничимся случаем двух показателей.

Пусть есть некоторое свойство L, в определении которого участвуют два числовых показателя. Предположим, что L таково, что, если изучаемые показатели обладают этим свойством, то множество всех объектов можно некоторым "идеальным образом" разбить на 4 кластера, факт попадания объекта в каждый из которых характеризует способ участия его показателей в формировании свойства L. Иначе – это деление как бы наиболее явно указывает на наличие у изучаемых показателей свойства L. Назовем такое деление идеальным 4-разбиением. Договоримся также, что идеальное 4-разбиение всегда строится по типу четырехпольной таблицы, и, тем самым, порождает некоторые деления множеств значений каждого из числовых показателей на две части.

Способ деления множества значений каждого из показателей, порождаемый идеальным 4-разбиением, будем считать 4-идеальным. Таким образом, 4-идеальный способ представляет собой пару алгоритмов деления (для множества значений 1-го и 2-го показателя соответственно), которые должны работать в определенном смысле параллельно.

Конкретизируем задачу.

Исходными данными будут являться две связанные выборки, в каждой из которых i-м элементом выступает значение соответствующего числового показателя i-го объекта. Требуется оценить, в какой степени пара показателей X и Y удовлетворяют свойству L.

Для решения поставленной задачи предлагается создать "идеальную" четырехпольную таблицу, которая могла бы получиться на практике для данных, наилучшим возможным способом удовлетворяющих свойству L и сформулировать алгоритм ее построения. Назовем этот алгоритм L-идеальным. По разработанному L-идеальному алгоритму для имеющихся выборочных данных построим реальную четырехпольную таблицу. Степень

отличия этой таблицы от "идеальной" и будет являться оценкой того, насколько свойство L можно считать справедливым для наших данных.

Ниже, в примере, как чаще всего и случается в приложениях, под свойством L будем понимать свойство независимости показателей, что полностью соответствует стандартной методике латентного кластерного анализа. В этом случае весь спектр значений каждого из показателей можно разделить на "верхнюю" и "нижнюю" части, а кластеры на множестве объектов в итоге формируются по тем же принципам, что и стандартные четырехпольные таблицы. "Идеальной" таблицей в этом случае следует признать диагональную.

3. Метод решения. Обсуждение алгоритма

Таким образом, одной из важных проблем при изучении произвольного свойства L является задача построения идеальной четырехпольной таблицы. Для этого применим методику латентного кластерного анализа и начнем с построения латентной кластерной переменной. После (или непосредственно в процессе) ее построения кластеры связываются с определенными постоянными ее значениями или принадлежностью ее значений каким-либо заданным интервалам, – если, например, она приобрела значения от a до b, то это объект первого кластера, если от b до c, то второго и т.д.

В общем случае латентную кластерную переменную следует строить так, чтобы при переходе от реальных данных к "огрубленным" по ЛКП, в полученных данных в наибольшей степени наблюдалось бы свойство L. Под огрублением в данном случае мы понимаем замену всей информации об объекте значением его ЛКП, т.е. номером или нечисловым обозначением его кластера.

Мы предлагаем для построения L-идеальной четырехпольной таблицы рассмотреть значения каждого из двух наших показателей по отдельности и, перебирая все возможные способы образования из этих числовых рядов новых пар, сформировать новые пары показателей так, чтобы искусственные объекты, соответствующие этим парам, наилучшим возможным образом удовлетворяли L. Затем по "идеальной" четырехпольной таблице, содержащей эти искусственные объекты, построим L-идеальный алгоритм ее формирования. Далее этот алгоритм применяем уже к реальным выборочным данным.

Вследствие этого у нас получится две четырехпольных таблицы – "идеальная" и реальная. Произвольная заранее выбранная мера схожести этих таблиц Q_0 определяет степень справедливости свойства L на изучаемых данных.

Способ выбора меры схожести четырехпольных таблиц зависит, собственно, от рассматриваемой задачи. Это может быть, например, любая матричная мера разности двух таблиц, или произвольная кластерная метрика, оценивающая различия между идеальным и реальным кластерными разбиениями.

В качестве приложения предлагаемого метода рассмотрим случай, когда L означает независимость двух числовых показателей. Пусть были даны две связанные выборки X и Y объема n. Упорядочим каждую из них, например, по возрастанию ее элементов. В результате получим две новых выборки – X' и Y'. В упорядоченных множествах элементов этих выборок проводим границы так, чтобы по обе стороны каждой из границ было приблизительно равное количество элементов. Затем, соединяя пары значений X‘ и Y‘ согласно полученному их порядку в новые искусственные объекты, по проведённым границам строим идеальную четырёхпольную таблицу сопряжённости. Идеальной она окажется хотя бы потому, что у искусственных объектов возрастанию первого показателя строго соответствует возрастание второго, а следовательно, она будем максимально похожа на диагональную, если только исходные выборочные данные имели именно такую тенденцию. Следовательно, при сортировке исходных данных далее нам будет нужен предварительный анализ тенденции в них – например, определение знака коэффициента

корреляции. Если он окажется положительным, то для идеальной четырёхпольной таблицы большему X должен соответствовать больший Y,и выборки нужно сортировать по возрастанию элементов. В случае, если коэффициент корреляции отрицателен, то порядок следования значений одного из признаков необходимо инвертировать.

Теперь можно перейти к нахождению латентной кластерной переменной.

Для этого воспользуемся множественной регрессией, точнее, как это принято в задачах латентного кластерного анализа, будем одновременно подбирать метки четырех кластеров $a_i,\ i = 1, 2, 3, 4$, соответствующих четырем клеткам таблицы и коэффициенты некоторого многочлена f двух формирующих показателей, решая регрессионными методами следующую оптимизационную задачу:

$$\sum_{j=1}^{4} \sum_{i \in N(j)} (f(x'_i, y'_i) - a_j)^2 \quad \to \min_{a_i, i=1,\ldots,4,\ f \in F}, \tag{1}$$

где $N(j)$ – множества номеров идеальных объектов, попавших в j-й кластер, а в качестве класса допустимых преобразований F мы выбрали класс многочленов степени, не выше 3, хотя, конечно же, возможны и иные решения.

При нахождении ЛКП можно столкнуться со следующими проблемами.

• Пустой кластер в реорганизованных данных. Это часто встречающаяся проблема, поскольку “идеальная” четырехпольная таблица должна быть похожа на диагональную. Ее можно решить небольшим сдвигом границы в каком-либо из упорядоченных наборов данных.

• При получении меток кластеров регрессионными методами, возможно получение совпадающих (в частности, нулевых) меток. Здесь можно наложить некоторое дополнительное условие на эти метки. В произведенных ниже расчетах в качестве такого условия было использовано

$$\sum_{j=1}^{4} a_j^2 = 1.$$

4. Пример применения к социологическим данным

Рассмотрим пример работы предложенного алгоритма на реальных данных. Они взяты из результатов первой переписи населения России 1897 г. и заимствованы из [6]. Использовались данные по 55 уездам Российской империи. Пусть X – плотность населения в уезде, а Y – процент городского населения в нём. В выборке X минимальное значение оказалось равным 1, а максимальное - 1199. В выборке Y минимальное – 0, а максимальное – 59.

После сортировки исходных данных получилась идеальная четырёхпольная таблица, приведенная в таблице 1. Для ее построения граница, разделяющая множество значений X, была выбрана равной 96, а для Y равной 5. Заметим, что при сортировке исходных данных мы инвертировали порядок следования Y, так как выборочный коэффициент корреляции $r(X, Y) = -0.171$ оказался отрицательным.

Таблица 1

“Идеальная” четырёхпольная таблица T'

	Y	X
X	26	9
Y	2	18

Далее, с помощью алгоритма расчёта оптимальных внутренних меток кластеров, представленном в статье [3], который в нашем случае совпал методом решения (1), были найдены метки четырех образовавшихся кластеров $a_i,\ i = 1, 2, 3, 4$ и одновременно построена $f(x, y)$. Метки получились равными -0.594, 0.125, 0.008, 0.795, а ЛКП имеет вид

$$f(x, y) = -0.588 + 0.0042x - 0.0121y - 0.0000028x^2 + 0.00032xy + 0.00022y^2.$$

Теперь для каждого из кластеров "идеальной" таблицы определим границы значений функции $f(x, y)$ и "реальную" четырехпольную таблицу T.

$$\begin{aligned} 1 \to &\quad f(x, y) < -0.235; \\ 2 \to &\quad -0.235 \leqslant f(x, y) < 0.066; \\ 3 \to &\quad 0.066 \leqslant f(x, y) < 0.401; \\ 4 \to &\quad f(x, y) \geqslant 0.401. \end{aligned}$$

Таблица 2

"Реальная" четырёхпольная таблица T

	Y	X
X	20	9
Y	7	19

Для оценки степени отличия двух получившихся таблиц используем евклидову матричную норму $||A||$, которая равна квадратному корню из наибольшего собственного числа матрицы AA^t. Степень выполненности нашего свойства по имеющимся данным, таким образом, может быть оценена числом

$$K = 1 - \frac{||T - T'||}{||T'||} = 0.7305\ ,$$

что свидетельствует о достаточно высокой степени независимости формирующих показателей. Принимая во внимание достаточно близкое к нулю значение коэффициента корреляции $r(X, Y)$, приведенное выше, можно сказать, что мы с использованием предложенного метода подтвердили наличие свойства независимости между исходными X и Y , при этом слегка иначе оценив степень справедливости этого свойства численно.

5. Заключение и выводы

В работе предложен метод оценивания степени соответствия наблюдаемых данных достаточно произвольному взаимному свойству двух числовых показателей. Стремление оценить степень обладания статистическим свойством некоторым числом, а не просто описать эту степень словами вроде "сильная" или "слабая" становится все более востребованным в связи с внедрением точных математических методов во все более широкие отрасли человеческой деятельности. Фактически предложенную методику можно понимать, как общие рекомендации по введению числовых коэффициентов, подобных коэффициенту корреляции в широкий круг задач статистики, в которых подобная формализация ранее отсутствовала.

При этом базой для оценивания выступает некоторое "идеальное" кластерное разбиение объектов, по значениям показателей которых решается поставленная задача. Это позволяет использовать соображения, отличные от классических даже при изучении свойства независимости, приобретая при этом новую интуицию и свежий взгляд на исследуемые свойства.

Список литературы

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

2. Дронов С.В., Герасимова А.С. К проблеме оцифровки кластерной переменной // Анализ, геометрия и топология. Труды Всероссийской молодежной школы-семинара. — Барнаул : ИП Колмогоров И.А., 2013. — С. 54–58.

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

4. Vermunt J.K., Magidson J. Latent class cluster analysis // Applied latent class analysis. — 2002. — Vol. 11. — P. 89–106.

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

6. Bryukhanova E.A., Chekryzhova O.I., Dronov S.V. Spatial Approach to the Analysis of the Employment Data in Siberia Based on the 1897 Census (the Experience of the Multivariate Statistical Analysis of the Districts Data) // Journal of Siberian Federal University. Humanities & Social Sciences. — 2016. — Vol. 7. — P. 1651–1660.