УДК 519.259

Новый алгоритм выявления и квантификации латентных классов

С.В. Дронов, А.Ю. Шеларь

Алтайский государственный университет (Барнаул, Россия)

A New Algorithm for Identifying and Quantifying of Latent Classes

S.V. Dronov, A.Yu. Shelar

Altai State University (Barnaul, Russia)

Работа с большими объемами данных может быть существенно упрощена, если эти данные разбиты на примерно однородные группы. Разбиение на такие группы — задача кластерного анализа. Однако вопрос о построении объективного, естественного разбиения на кластеры пока остается открытым. Рассматривается современный подход к поиску такой объективной кластерной структуры путем выделения из совокупности задающих объекты показателей общей существенной части. При ее фиксации формирующие показатели становятся независимыми или близкими к таковым. Получающиеся независимые остатки интерпретируются как своеобразный информационный шум, а латентная кластерная переменная, та общая фиксированная часть, которая обеспечивает такое превращение, — как причина объективного объединения объектов в кластеры. Предлагается новый алгоритм формирования кластерного разбиения на основе близости или совпадения значений латентной кластерной переменной с одновременной квантификацией ее значений. Алгоритм основан на целенаправленном переборе разбиений, переходе от стартового к разбиению, все более близкому к объективному. Предлагаемый алгоритм может быть просто перенесен на случай задания объектов нечисловыми категорированными показателями.

Ключевые слова: объективное разбиение, кластерный анализ, латентный анализ классов, большие данные.

Processing large amounts of data can be greatly simplified if this data is divided into approximately homogeneous groups. Splitting into such groups is the task of cluster analysis. However, the question of constructing an objective, natural partition into clusters remains open. The paper considers a modern approach to the search for such an objective cluster structure by highlighting the indicator of a common essential part from the set of characteristics that define objects (we call them the forming ones). When this indicator is fixed, the remains of the forming characteristics become independent or close to such. The resulting independent residuals are interpreted as a kind of information noise, and the latent cluster variable, the common fixed part that provides such a transformation, can be a reason for the objective integration of objects into clusters. A new algorithm for the formation of a cluster partition based on the proximity or coincidence of the values of a latent cluster variable with the simultaneous quantification of its values is proposed. The algorithm is based on the targeted search of partitions, the transition from the start one to the partition, more close to the objective. The algorithm proposed in the paper can be easily modified to the case of non-numeric categorized characteristics.

Key words: objective partitioning, cluster analysis, latent class analysis, big data.

DOI 10.14258/izvasu(2020)4-12

1. О выявлении латентных классов. В условиях возможности накопления и обработки все большего количества данных об объектах наблюдения произвольной природы (в наши дни это называют big data [1–3]) все большую востребованность приобретают методы предварительной классификации или кластеризации данных, см., например, [4, 5]. При этом методы кластеризации (разбиения множества объектов на дизъюнктные группы) производятся традиционными кластерными алгоритмами по степени близости наборов некоторых показателей, которыми характеризуются наблюдаемые объекты. Мы называем такие показатели Формирующими, имея в виду, что именно с помощью них Формируются те кластеры, к изучению которых исследователь впоследствии переходит.

В последнее десятилетие возникло понимание того, что, как бы удачно ни подбирались формирующие показатели, существенные особенности объектов, по близости которых их следовало бы объединять в кластеры, классы похожих друг на друга объектов, чаще всего остаются недоступными для наблюдений, или иначе — латентными. Такие разделы интеллектуального анализа данных, как факторный анализ (см. [6]), были созданы с надеждой, что требуемые показатели могут быть оценены через комбинации формирующих, если потребовать, чтобы эти комбинации в наибольшей возможной степени объясняли бы попарные корреляции между исходными формирующими показателями.

Но давайте заметим, что объяснение корреляций даже наилучшим образом, тем не менее, непосредственно не связано с построением объективного, естественного разбиения объектов на кластеры. Конечно же, возможно сначала выделить упомянутые латентные факторы, а затем именно по ним произвести кластерное разбиение, используя их в качестве формирующих. Иногда это приводит к очень хорошим, легкоинтерпретируемым результатам, как было сделано одним из авторов в [7]. Но (и это довольно обычная в приложениях ситуация), чаще последовательное применение двух приближенных методов приводит к существенному искажению исходной картины.

Поэтому можно попытаться, что называется, пройти два этапа в один проход алгоритма, совместить построение объективных классов с выделением определяющих их показателей. Такие алгоритмы получили названия алгоритмов выявления латентных классов. В результате их работы возникает новая качественная переменная, указывающая для каждого объекта тот класс, к которому он должен быть объективно отнесен. Подобным образом строящаяся переменная в [8] была названа кластерной переменной.

Решение задачи формирования латентных классов, строго сформулированная в [9], тем самым, формализует поиск качественной латентной кластерной переменной (ЛКП), каждое значение которой является обозначением того кластера, в который попадает объект. При этом значения ЛКП должны содержать в себе фактически всю полезную информацию о некотором общем свойстве, первоначально содержавшемся в значениях формирующих показателей объектов, по совокупности которых кластеры строятся.

Предполагается, что все исходные формирующие показатели достаточно тесно связаны с этим свойством и, помимо полезной для создания объективных классов информации, вероятно, содержат некоторый, предположительно небольшой, информационный шум. Идея алгоритмов латентного анализа классов основана на том, что если мы успешно «уберем» значение ЛКП из каждого формирующего показателя, то в результате останется лишь шум. Соберем значения информационного шума в многомерный вектор невязок, каждая координата которого представляет собой значение шума для одного из показателей. Критерием степени успешности полученного решения является то, насколько координаты полученного вектора окажутся близки к статистически независимой совокупности. Действительно, если это не так, то остатки формирующих показателей содержат как бы общую для всех них часть, которая также может быть перенесена в значения ЛКП. Итак, ЛКП надо построить так, чтобы совместное условное при фиксации ЛКП распределение вектора невязок оказалось бы как можно ближе к распределению вектора с независимыми координатами (ср. [10], где эта мысль изложена гораздо менее строго).

Таким образом, объективное разбиение множества рассматриваемых объектов на кластеры должно производится так, чтобы внутри каждого из кластеров — при фиксировании соответствующего значения ЛКП, — формирующие факторы были бы как можно более похожи на независимые.

2. Постановка основной задачи. Исходными данными в нашей задаче будут, таким образом, наборы формирующих показателей, задающих множество n изучаемых объектов. Требуется разбить это основное множество на дизъюнктные подмножества так, чтобы внутри каждого из построенных подмножеств (которые мы будем называть латентными кластерами) формирующие показатели были бы как можно более похожи на независимые. Хотя можно оставить неопределенным и подлежащим подбору также и количество строящихся латентных кластеров, но мы предположим, что число этих кластеров k заранее известно. Это позволит существенно ускорить работу алгоритма, грубо говоря, просто сократив число возможных вариантов построения кластеров. К тому же в практических задачах число кластеров, которое должно в итоге получиться, известно. Это, например, так в одном из наиболее популярных сегодня кластерных алгоритмов — методе k-средних (современные варианты этого алгоритма и проблемы, возникающие при его применении, обсуждаются в [11]). Приведем сначала постановку задачи в наиболее общем виде.

Разобьем основное множество объектов на k попарно непересекающихся подмножеств-кластеров. Пусть R_j — числовая оценка степени взаимной независимости совокупности формирующих показателей по их наборам внутри j-го из кластеров, $j = 1, ..., k$. Будем считать, что эта характеристика тем меньше, чем более показатели в этом кластере похожи на независимые. Напри-

мер, этой характеристикой может служить сумма квадратов всех попарных оценок коэффициентов корреляции формирующих показателей, или она может быть построена на основе статистики χ^2 Пирсона. Последнее делает возможным применение предлагаемых далее методик и при обработке нечисловых категорированных показателей.

Требуется построить такое разбиение исходного множества, чтобы критерий

$$R = \sum_{j=1}^{k} R_j \qquad (1)$$

принимал бы свое минимальное значение.

Как уже было сказано, построив такое разбиение, мы создадим новый качественный показатель, ЛКП, который для каждого из объектов указывает тот кластер, к которому относится в итоге этот объект. Параллельно поставим задачу оцифровки, или, иначе, квантификации ЛКП, то есть замены ее качественных значений на соответствующим образом подобранные числовые значения или числовые метки латентных классов.

Задача квантификации ЛКП имеет особое значение, поскольку признак, по которому группируются объекты в этом случае, является латентным, ненаблюдаемым, хотя по сути служит причиной, приводящей к необходимости признания объектов близкими. Более того, можно сказать, что это «объективная» причина, объясняющая истинное положение вещей. Поэтому очень важно придать ей числовую форму для дальнейшей статистической обработки.

При создании алгоритма нами рассмотрен только простейший случай, когда объекты заданы лишь двумя формирующими показателями X, Y, а число кластеров, которые мы хотим получить, равно 4. Это число практически всегда появляется при исследовании двух категорированных показателей. Например, для медицинских данных можно изучать наличие или отсутствие заболевания при наличии или отсутствии воздействия фактора курения пациента или его контактов с носителями коронавируса. При этом естественно возникает 4 группы пациентов. На этом примере видно, что установление взаимосвязей между подобными показателями и определение скрытых причин такой связи (ЛКП) есть, безусловно, важная задача прикладного анализа.

Итак, алгоритм должен получить на вход две связанные выборки объема n $X = (x_1, ..., x_n)$ и $Y = (y_1, ..., y_n)$. При этом критерием качества (1) выбрана сумма квадратов оценок коэффициентов корреляции между X и Y внутри текущих кластеров

$$R = \sum_{j=1}^{4} R_j^2(X, Y).$$

3. Алгоритм направленного улучшения разбиения. Самое очевидное решение поставленной задачи — перебрать всевозможные разбиения множества изучаемых объектов на подмножества, например, с помощью алгоритма из [12], для каждого из разбиений вычислить R и выбрать то, которое приводит к наименьшему его значению. Для нашего случая можно сократить полный перебор всех разбиений, поскольку в приложениях не рассматриваются кластеры менее чем из 3 элементов, к тому же мы ограничили рассмотрение случаем, когда множество разбивается ровно на 4 кластера. Тем не менее ясно, что простой перебор может и должен быть сокращен. Можно придать некоторую направленность перебору, т.е. организовать его так, чтобы каждое новое разбиение было не хуже предыдущего. Для нас это означает уменьшение (или, по крайней мере, неувеличение) R.

Мы примем на веру тезис о том, что нужная система кластеров может быть получена за конечное число шагов из произвольной стартовой системы кластеров путем последовательного перемещения элементов между кластерами так, что каждый раз перемещается ровно один элемент. Следующие технические леммы позволят нам проследить изменение величины R при перемещении какого-то одного объекта между кластерами. Они доказываются аккуратным отслеживанием изменений в обычной формуле оценки выборочного коэффициента корреляции при таком перемещении.

Лемма 1. *При добавлении к двумерной выборке (X, Y) объема n нового элемента (x, y) значение нового выборочного коэффициента корреляции ρ_{new} связано с имевшимся ранее ρ соотношением*

$$\rho_{new} = \frac{\rho + a_x a_y}{\sqrt{(1 + a_x^2)(1 + a_y^2)}}, \qquad (2)$$

где

$$a_x = \frac{x - \overline{X}}{S_X\sqrt{n+1}}, \quad a_y = \frac{y - \overline{Y}}{S_Y\sqrt{n+1}},$$

а $\overline{X}, S_X, \overline{Y}, S_Y$ – выборочные средние и среднеквадратические отклонения координат исходной двумерной выборки.

Лемма 2. *При исключении из двумерной выборки (X, Y) объема n одного из элементов (x, y) значение нового выборочного коэффициента корреляции ρ_{new} связано с имевшимся ранее ρ соотношением*

$$\rho_{new} = \frac{\rho - b_x b_y}{\sqrt{(1 - b_x^2)(1 - b_y^2)}}, \qquad (3)$$

где

$$b_x = \frac{x - \overline{X}}{S_X\sqrt{n-1}}, \quad b_y = \frac{y - \overline{Y}}{S_Y\sqrt{n-1}},$$

а $\overline{X}, S_X, \overline{Y}, S_Y$ – выборочные средние и среднеквадратические отклонения координат исходной двумерной выборки.

Простое сравнение ρ и ρ_{new} показывает, что, если

$$\rho \cdot (x - \overline{X})(y - \overline{Y}) \ < \ 0, \qquad (4)$$

то при перемещении элемента с показателями (x, y) внутрь текущего кластера получаем уменьшение модуля выборочного коэффициента корреляции: $\rho^2_{new} < \rho^2$.

Теперь можно описать предлагаемый алгоритм построения оптимального разбиения, элементы которого являются объективными латентными классами. Предполагается, что исходное множество содержит не менее 12 объектов, иначе разбить его требуемым образом просто не получится.

Исходные данные некоторым произвольным образом разбиваются на 4 кластера (в каждом не менее 3 элементов). Примем это разбиение за текущее.

1. Вычисляем оценку коэффициента корреляции ρ_i по набору объектов, формирующих i-й кластер текущего разбиения, $i = 1, 2, 3, 4$.
2. Выбираем один кластер (A_i). Во всех остальных кластерах ищем потенциальные к перемещению в выбранный кластер объекты (x, y), удовлетворяющие условию (4).
3. Для каждого такого объекта, если его первоначальный кластер A_j содержал не менее 4 элементов, используя формулы (2) и (3), вычисляем величину, на которую уменьшится критерий R при перемещении

$$Q_{i,j}(x, y) = \rho_i^2 + \rho_j^2 - \rho^2_{new,i} - \rho^2_{new,j}$$

и запоминаем ее. Проделываем это для каждого из текущих кластеров.

4. Среди всех пар i, j выбираем такую, для которой $Q_{i,j}(x, y)$ максимально.
5. Если найденное максимальное значение отрицательно, то перемещать больше нечего — конец алгоритма. Текущее разбиение и есть оптимальное.
6. Иначе производим перемещение (x, y) из A_i в A_j и объявляем полученное разбиение текущим. Возвращаемся к шагу 1.

4. Квантификация латентной кластерной переменной. Два примера. Будем считать, что в результате работы описанного в предыдущем разделе алгоритма объекты разбиты на непересекающиеся классы, которые мы назвали латентными кластерами. На примерах отчетливо видно, что объекты собираются в эти кластеры, непосредственно не согласуясь со значениями пар значений своих формирующих показателей — на поле корреляции этих показателей латентные классы не образуют выделяющихся компактных групп. Поэтому даже упорядочивание этих кластеров и присваивание им последовательных натуральных меток невозможно произвести, руководствуясь только интуитивными соображениями.

Тем не менее согласно самой постановке задачи, латентные кластеры формируются в соответствии со значениями той общей, латентной части формирующих показателей, которая и представляет собой квинтэссенцию главного свойства, обеспечивающего получившееся объективное кластерное разбиение. Чтобы представить себе правильный порядок следования полученных кластеров, оценить степень их различия между собой, да и просто придать некое числовое значение латентной переменной, прибегнем к одному из методов квантификации категорированных показателей.

Мы использовали метод так называемой внутренней квантификации, предложенный в [8], хотя возможно применение произвольных методов оцифровки кластерной переменной, например, описанных в главе 14 монографии [13].

Вторым автором настоящей работы была написана компьютерная программа на языке Си, реализующая предложенный в предыдущем разделе алгоритм.

Рассмотрим два примера работы этого алгоритма на практических данных. Первый пример построен на данных переписи населения России 1897 года, которые ранее считались утраченными, и только недавно были обнаружены историками нашего университета. Данные взяты из [7].

В качестве формирующих показателей выбраны процент городского и процент мужского населения в каждом из 55 сибирских уездов. После обработки выделились латентные кластеры, обозначенные A, B, C, D, которые содержат 18, 20, 10 и 7 уездов соответственно. После квантификации эти кластеры получили числовые метки -0,86; 0,89; -1,16 и 1,33. Это означает, что естественный порядок кластеров таков — C, A, B, D, а наиболее сильно среди соседних отличаются друг от друга кластеры A и B. Значение критерия (1) здесь $R = 0,327$.

Второй пример связан с изучением LIC — концентрации железа в печени пациентов, вычисляющимся по данным МРТ и T2* — времени релаксации протонов на МРТ. Данные 22 пациентов Санкт-Петербургского МРТ-центра были любезно предоставлены врачом-радиологом А.М. Титовой. Здесь численный состав пациентов в кластерах оказался таким: A — 6 пациентов, B — 8, C — 5 и D — 3 человека. Метки кластеров: -0,28; 1,20; -0,59 и -1,65 соответственно. Тем самым, здесь порядок кластеров D, C, A, B. Здесь подсчеты дали $R = 2,73$. Был найден и конкретный состав каждого из кластеров в обоих примерах.

5. Обсуждение результатов. Есть также некоторые дополнительные возможности, которые дает возможность построения «естественного» разбиения объектов на кластеры. После построения латентных кластеров мы можем, например, дать численную оценку степени адекватности системы имевшихся формирующих показателей для подобного построения. Это может быть сделано с помощью следующего коэффициента:

$$Adq(X,Y) \;=\; 1 - \frac{R}{R_{max}},$$

где R – величина критерия (1), рассчитанная по построенной алгоритмом латентной кластерной структуре, а R_{max} – теоретически возможное максимальное значение этого критерия. Назовем его коэффициентом адекватности.

Для нашей постановки задачи $R_{max} = 4$ достигается тогда, когда внутри каждого из латентных кластеров между показателями имеется линейная связь, т.е. они предельно сильно зависимы. Значение $Adq(X,Y) = 1$ соответствует некоррелированности показателей внутри каждого из кластеров, а значит, полную адекватность системы показателей X, Y, т.е. полную возможность построить по их значениям систему объективных кластеров. Чем меньше значение предложенного коэффициента, тем хуже подходят имеющиеся показатели для объективной кластеризации. Малые значения Adq должны побуждать исследователя к поиску дополнительной информации об объектах, привлечению других показателей к построению объективных кластеров.

В первом из рассмотренных выше примеров $Adq(X,Y) = 0,918$, что позволяет сделать вывод о высокой адекватности системы выбранных показателей и близости полученного кластерного разбиения к объективному. Во втором же $Adq(X,Y) = 0,317$, что указывает на практическую невозможность построить 4 объективных кластера. Более подробный анализ показывает, что два предложенных показателя сильно, почти функционально зависимы, что и служит причиной выявленного обстоятельства.

6. Заключение. Нами предложен и реализован один из возможных алгоритмов построения латентных кластеров, который, несмотря на относительную сложность, работает значительно быстрее полного перебора. Его вычислительная сложность в нашей задаче составляет $O(n^2)$ по сравнению с экспоненциальной сложностью полного перебора разбиений. С помощью результатов работы возможно также численно оценить те скрытые причины, по которым строятся латентные кластеры и определить степень адекватности системы показателей для правильного построения «объективного» кластерного разбиения.

Библиографический список

1. Johnson J.M., Khoshgoftaar T.M. Survey on deep learning with class imbalance // J. Big Data. 2019. Vol. 6, 27. DOI 10.1186/s40537-019-0192-5.

2. Wu J., Dong M., Ota K., Li J. and Guan Z. Big Data Analysis-Based Secure Cluster Management for Optimized Control Plane in Software-Defined Networks // IEEE Transactions on Network and Service Management. 2018. Vol. 15. DOI: 10.1109/TNSM.2018.2799000.

3. Chen M., Mao S., Zhang Y., Leung V. Big Data. Related Technologies, Challenges, and Future Prospects. Spinger, 2014. DOI: 10.1007/978-3-319-06245-7.

4. Romesburg H.Ch. Cluster analysis for researchers. Lulu Press, 2007.

5. Chance B.L., Rossman A.J. Investigating statistical concepts, applications, and methods. Duxbury Press, 2013.

6. Mulaik S.A. Foundations of Factor Analysis. Boca Raton, 2009.

7. 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. № 7. DOI: 10.17516/1997-1370-2016-9-7-1651-1660.

8. Dronov S.V., Sazonova A.S. Two approaches to cluster variable quantification // Model Assisted Statistics and Applications. 2015. Vol. 10.

9. Vermunt J.K., Magidson J. Latent class cluster analysis // Applied latent class analysis. 2002. Vol. 11.

10. Rindskopf D. Latent Class Analysis. The SAGE Handbook of Quantitative Methods in Psychology. N.Y., 2009.

11. Gribel, D., Vidal T. HG-means: A scalable hybrid metaheuristic for minimum sum-of-squares clustering // Pattern Recognition. 2019. 88 (1). arXiv: 1804.09813.

12. Федоряева Т.И. Комбинаторные алгоритмы: учебное пособие. Новосибирск, 2011.

13. Дронов С.В. Методы и задачи многомерной статистики. Барнаул, 2015.