Группировка признаков на основе оптимальной последовательности миноров корреляционной матрицы*

С. Д. Двоенко, Д. О. Пшеничный

dsd@tsu.tula.ru; denispshenichny@yandex.ru

[1]Тульский государственный университет, Россия, г. Тула, пр. Ленина, 92

При решении задачи группировки возникает проблема содержательной интерпретации полученных факторов и групп признаков. Тем не менее факторы групп является синтетическими признаками, интерпретация которых может быть затруднена, поэтому после выделения групп признаков и построения соответствующих им факторов в каждой группе обычно определяется ее представитель как наиболее сильно коррелирующий с фактором группы признак. Тогда оказывается возможным содержательно интерпретировать результат группировки прямо в терминах исходных признаков. Предложен новый подход для выбора подмножества признаков, способных адекватно представить скрытые факторы, без определения собственных или центроидных направлений в качестве промежуточных преобразований. Данный подход основан на построении оптимальной последовательности значений главных миноров корреляционной матрицы признаков. В начале такой оптимальной последовательности расположены наименее коррелированные друг с другом и с остальными признаки, а к ее концу выстраиваются все более коррелированные с остальными признаки, выбранные в последнюю очередь. Показано, что предложенный подход позволяет формировать начальное решение для других алгоритмов группировки и также может применяться самостоятельно для оценки числа групп и построения содержательных группировок.

Ключевые слова: *группировка; кластер; метрика; корреляция; собственное число; собственный вектор; детерминант*

DOI: 10.21469/22233792.2.2.1

1 Введение

Считается, что задачи анализа данных в первую очередь возникают на ранних этапах исследования изучаемого явления, когда еще не построена его модель и еще рано говорить о задаче ее идентификации. В этом случае необходимо накопить и изучить как можно больше разнородной информации о наиболее существенных свойствах изучаемого явления. Вынужденность и противоречивость такого подхода вполне очевидна: вообще говоря, неизвестно, какие свойства наиболее существенны, и, как следствие, неизвестно, какие сведения накапливать.

Таким образом, интеллектуальные методы анализа данных должны устранить указанное выше противоречие, сконцентрировав в описании изучаемого явления наиболее существенную и адекватную информацию о нем. В основе такого подхода лежат достаточно естественные предположения, сформулированные в виде так называемых гипотез. Таких гипотез, по сути, всего две: компактности и скрытых факторов.

*Работа выполнена при частичной финансовой поддержке РФФИ, проекты №№ 15-07-02228, 15-07-08967, 14-07-00527 и 14-07-00964.

В интеллектуальном анализе данных обычно предполагается, что экспериментальные сведения об изучаемом явлении представлены как результаты измерений в виде матрицы данных $X(N, n)$, где N — число измерений; n — число измеряемых характеристик. Каждый акт измерения характеристик изучаемого явления рассматривается как объект $\omega_i \in \Omega$, который процессом измерения помещен в n-мерное признаковое пространство и представлен в нем вектором-строкой $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$, $i = 1, \ldots, N$. Матрица данных представляет собой множество из N строк $X(N, n) = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$, расположенных друг под другом.

Согласно гипотезе компактности, предполагается, что объекты образуют K локальных сгущений (классы, кластеры, таксоны), которые следует выделить (отделить друг от друга), так как они, предположительно, характеризуют различные состояния изучаемого явления.

С другой стороны, совокупность измерений одной характеристики образует вариационный ряд, т. е. признак, представленный наблюдениями $X_j = (x_{1j}, \ldots, x_{Nj})^{\mathrm{T}}$. Тогда матрица данных представляет собой множество из n вариационных рядов-столбцов $X(N, n) = (X_1, \ldots, X_n)$.

Согласно гипотезе скрытых факторов, считается, что их поведение определяет соответствующие «глубинные» свойства объекта исследования, которые проявляются через измеренные признаки как его реакции на внешние воздействия. Факторы проявляются через измеряемые признаки и различным образом влияют на эти признаки. Зависимость признаков от некоторого фактора определяет похожесть их поведения, т. е. похожесть изменений значений соответствующих вариационных рядов. Предполагается, что существует L таких факторов F_i, которым должны соответствовать группы признаков G_i, $i = 1, \ldots, L$.

Очевидно, что объективная закономерность, скрыто присутствующая в изучаемом явлении, обязательно проявится в результатах обработки различными методами и алгоритмами, основанными на различных предположениях о нем. Комплексирование таких сведений позволит в итоге адекватно решить поставленную задачу [1].

Таким образом, необходимо поддерживать и расширять разнообразие интеллектуальных методов обработки данных. В данной работе предпринята такая попытка. Актуальность таких попыток вполне очевидна, особенно в связи с накоплением больших объемов экспериментальных данных и развитием методов обработки данных, представленных парными сравнениями.

2 Задача группировки признаков

Задача группировки признаков имеет самостоятельное значение и может решаться разными способами.

Относительно факторов делается важное предположение, что в идеале они независимы. Статистический смысл независимости факторов означает, что соответствующие вариационные ряды наблюдений, будучи построенными, окажутся некоррелированными. Это означает, что такие скрытые признаки могут быть представлены наблюдениями $F_i = (f_{1i}, \ldots, f_{Ni})^T$, $i = 1, \ldots, L$, которые в соответствующем пространстве формируют систему ортогональных векторов.

Если сначала определяются факторы, то потом определяются признаки, подверженные их влиянию в наибольшей степени (задача факторного анализа и проблема вращения для определения факторных нагрузок и получения так называемой «простой» факторной структуры).

Известная проблема заключается в том, что ортогональное вращение факторов не совсем адекватно решает проблему получения простой структуры, поэтому приходится применять косоугольное вращение, что усложняет модель факторного анализа, так как факторы уже не являются независимыми [2, 3].

Если же сначала выделять группы сильно коррелирующих признаков, где признаки из разных групп почти не коррелируют, то потом можно построить представляющие эти группы факторы. При таком решении, в частности, проблема простой факторной структуры для косоугольной системы факторов решается автоматически, хотя сами факторы несколько отличаются от классической факторной модели. В этом случае решается, например, известная задача экстремальной группировки [4]. Следует отметить, что данная задача решается также и для центроидных направлений.

Отметим, что в обоих вариантах задачи группировки возникает проблема содержательной интерпретации полученных факторов или соответствующих групп признаков. Опыт показывает, что признаки, объединяемые в группы, часто можно совместно содержательно интерпретировать. С другой стороны, фактор группы все-таки является синтетическим признаком, интерпретация которого может быть затруднена, поэтому часто применяется следующий прием.

После выделения групп признаков и построения соответствующих им факторов в каждой группе определяется так называемый «представитель» группы как наиболее сильно коррелирующий с фактором группы признак. Далее рассматривается только множество таких признаков-представителей.

Очевидно, что в этом случае задача группировки также решает и другую известную задачу сокращения размерности признакового пространства. Эта задача также имеет самостоятельное значение. В данном случае получается сокращенное и содержательно интерпретируемое признаковое пространство. Важное свойство такого подпространства очевидно: эти реальные признаки коррелируют между собой в наименьшей степени и лучше всего могут представить скрытые факторы. Совсем упрощая, их даже часто рассматривают как факторы.

Легко увидеть, что при таком подходе все преобразования, выполняемые в соответствии с факторной моделью, являются промежуточными, так как в итоге выбираются некоторые исходные признаки.

Можно ли предложить другой подход, который позволит выбрать подмножество из исходных признаков, обладающих аналогичными свойствами, не требуя построения собственных направлений (а также и центроидных) в качестве промежуточного этапа преобразований? Ниже рассмотрен один из возможных подходов.

3 Метричность конфигурации элементов и ее нарушения

Следует отметить, что задача группировки (выделения факторов) решается для матрицы взвешенных скалярных произведений признаков X_j, $j = 1, \ldots, n$, т. е. для матрицы $R(n, n)$ корреляций вариационных рядов наблюдений. Для определения свойств факторов сами наблюдения $X(N, n)$ уже не нужны. Именно поэтому в факторном анализе оценка значений факторов как восстановленных наблюдений является отдельной и дополнительной задачей.

Это замечание особенно актуально в связи с развитием современных подходов, опирающихся на данные об объектах исследования, представленных только лишь в виде парных сравнений. В этом случае предполагается, что существует гипотетическая система признаков или что реальные признаки существуют, но для измерения уже недоступны.

Считается, что от измеренных признаков остались лишь матрица расстояний $D(N,N)$ или скалярных произведений $C(N,N)$ между объектами и матрица корреляций $R(n,n)$ между признаками.

Развитие этих методов показывает, что нужно обеспечить вложенность экспериментальных наблюдений в соответствующее метрическое (евклидово) пространство признаков и предложить модификации алгоритмов кластер-анализа и группировки, не требующих явного наличия матрицы данных X.

Проблема метричности конфигурации элементов известна и рассматривается, например, в задаче шкалирования [5]. Ее конечной целью является восстановление хорошо интерпретируемых признаков в явном виде как представленных соответствующими измерениями. Если этого не требуется, то известные задачи кластеризации и группировки можно решить и без непосредственного восстановления собственно значений признаков. В частности, такой подход позволяет для решения задач кластеризации и группировки применять одни и те же алгоритмы, рассматривая объекты или признаки просто как элементы множества, погруженные в соответствующее метрическое пространство [6].

Если элементами множества являются признаки, то исходя из смысла похожести вариационных рядов рассматривают модули или квадраты коэффициентов корреляций в матрице $R(n,n)$. Кроме того, если изначально рассматривается некоторая функция парных сравнений, имеющая смысл близости $s_{ij} \geqslant 0$; $i,j = 1,\ldots,n$, то ее можно рассматривать как положительные вариации (или корреляции, если они нормированы).

На практике часто в полученных конфигурациях элементов имеются метрические нарушения. Причины этого различны. Поэтому одной из актуальных задач современного анализа данных является восстановление метричности данных. Именно в этом случае применение упомянутых выше алгоритмов является математически корректным.

Известно, что нарушения метричности конфигураций приводят к появлению отрицательных собственных чисел в матрице скалярных произведений между элементами множества. В случае множества признаков это относится к матрице корреляций $R(n,n)$. Если ее собственные числа упорядочить по убыванию $\lambda_1 > \cdots > \lambda_n$, то можно считать, что пространства размерностей, соответствующих отрицательным собственным числам, не существуют в том смысле, что в них для наблюдений не выполняется, например, теорема Пифагора или, в общем случае, теорема о косинусах, могут быть нарушены неравенства треугольника и т. д. Это ведет к тому, что интуитивно понятные аналогии нашему обычному трехмерному пространству, полезные в анализе данных, становятся некорректными. Но тогда и результаты обработки, вообще говоря, следует признать недостаточно корректными, где уровень некорректности определяется математической некорректностью результата.

Как известно, величина дисперсии данных — это размерность n пространства признаков. Для устранения в матрице $R(n,n)$ отрицательных собственных чисел обычно можно применить известное дискретное разложение Карунена–Лоэва, а именно: из всех элементов матрицы $R(n,n)$ «послойно» исключить вклады собственных векторов (направлений), соответствующих отрицательным собственным числам (также будем говорить, что это — «вклады» собственных чисел).

Заметим, что в факторном анализе матрица так называемых «остаточных» корреляций $R_q(n,n)$ определяется после послойного устранения вкладов первых q собственных векторов, соответствующих собственным числам, упорядоченным по убыванию. Естественно, что $\det R_q(n,n) = 0$.

Удобно также применить этот термин к результату устранения вкладов q отрицательных собственных чисел, которые оказываются последними в упорядочении. Очевидно, у такой матрицы остаточных корреляций $R_{q-}(n,n)$ также $\det R_{q-}(n,n)=0$.

Легко увидеть, что после устранения вкладов q отрицательных собственных чисел матрица остатков $R_{q-}(n,n)$ становится ненормированной (и более того, некорректной), где $r_{ii}>1,\ \ i=1,\ldots,n$. Но тогда, строго говоря, в данных «ниоткуда» появляется добавочная дисперсия, так как $\sum_{i=1}^{n} r_{ii}>n$. Формально из $R_{q-}(n,n)$ можно получить корректную корреляционную матрицу с единичной главной диагональю, просто пронормировав ее.

Очевидно, что нормировка в этом случае уничтожает сведения о доле внесенной дисперсии, поэтому в общем случае такой процесс появления новой дисперсии в данных после нормировок матрицы корреляций уже невозможно проконтролировать. Это нежелательно, например, когда решается задача группировки признаков.

По-видимому, эта проблема не столь принципиальна при наличии признакового пространства, т. е. матрицы данных $X(N,n)$. В этом случае можно построить матрицу так называемых «вычисленных признаков» $Y(N,m)$, где $m=n-q<n$ и $\lambda_1>\cdots>\lambda_m>0$, как проекций векторов-объектов из X на m первых собственных направлений. В пространстве вычисленных признаков наблюдения-строки $\mathbf{y}_i=(y_{i1},\ldots,y_{i\,m})$ образуют метрическую конфигурацию, что позволяет корректно решать задачи группировки, кластеризации, визуализации и т. д.

С другой стороны, в линейной факторной модели существует известная проблема определения общностей (вкладов общих факторов в дисперсию данных). Например, в методе главных факторов после редукции $R(n,n)$ с целью устранения дисперсий характерных факторов редуцированная матрица $\bar{R}(n,n)$ оказывается ненормированной, так как $r_{ii}<1,\ \ i=1,\ldots,n$, из-за уменьшенных значений ее диагональных элементов. Это известная в факторном анализе проблема определения общностей, теоретического решения которой не предлагалось. Есть лишь эмпирические рекомендации по оценке величины общностей.

Здесь следует отметить следующее. Следование эмпирическим рекомендациям часто приводит к появлению отрицательных собственных чисел в редуцированной матрице $\bar{R}(n,n)$, т. е. к нарушению метричности конфигурации элементов множества. Чтобы избежать этого, при построении главных факторов придется лишь «слегка» редуцировать диагональные элементы корреляционной матрицы, обычно в значительно меньшей степени, чем по эмпирическим рекомендациям. Вообще-то, это означает, что доля дисперсии в данных, объясняемая общими факторами, очень высока. Как в этом случае интерпретировать соотношения общностей и характерностей с точки зрения факторной модели — это другая проблема.

Следует также отметить, что проблема общностей возникает и при построении центроидных факторов. Центроидные направления отличаются от собственных направлений, но эмпирический принцип выбора общностей также приводит к появлению отрицательных собственных чисел, т. е. к нарушению метричности конфигурации элементов множества.

4 Оптимальная последовательность признаков и выбор их подмножества

В отличие от процедуры Карунена–Лоэва авторами был предложен другой метод так называемой «индивидуальной» корректировки лишь некоторых (или всех) парных сравнений некоторых элементов множества с остальными элементами для восстановления нарушенной метрической конфигурации, при котором сохраняется дисперсия данных [7, 8].

В данном методе наличие собственных чисел в матрице $S(n,n)$ взвешенных скалярных произведений, где $s_{ii}=1,\ i=1,\ldots,n$, связывается не с послойным ее разложением на вклады соответствующих собственных векторов (чисел), а с индивидуальными вкладами самих элементов множества. В качестве такой матрицы можно взять, например, матрицу $R(n,n)$ корреляций, модулей или квадратов корреляций признаков.

Пусть дана симметричная нормированная матрица $S(n,n)$. Согласно критерию Сильвестра [9], матрица квадратичной формы положительно определена, если все ее главные миноры $S_k=S(k,\ k),\ k=1,\ldots,n$, положительны: $\det S_k>0$, где $S_1=S(1,1)=s_{11}=1$. Согласно следствию из закона инерции Сильвестра, число q отрицательных собственных чисел совпадает с числом смен знаков детерминантов в последовательности $S_0=1,\ S_1,\ S_2,\ldots,S_n=S(n,n)$. Легко увидеть, что значения главных миноров (их детерминанты) в нормированной S убывают, начиная с единицы. При наличии отрицательных собственных чисел последовательность главных миноров оказывается знакопеременной, где значения главных миноров постепенно уменьшаются по модулю.

Известно, что одновременная перестановка двух строк и двух соответствующих столбцов в S не изменяет ее собственных чисел. Такая перестановка соответствует перестановке двух элементов множества. Определим такой порядок элементов множества, чтобы смены знаков значений главных миноров в последовательности $S_k,\ k=1,\ldots,n$, происходили в ее конце. Если матрица $S(n,n)$ ранга n имеет q отрицательных собственных чисел, то тогда в идеальном случае главный минор S_{n-q+1} впервые окажется отрицательным: $\det S_{n-q+1}<0$, а знаки последующих миноров будут чередоваться.

Естественно считать, что именно в этот момент: $k=n-q+1$ очередной элемент множества ω_k, представленный своими парными сравнениями $s_{ki}=s_{ik},\ i=1,\ldots,n$, с остальными, внес метрическое нарушение в уже построенную конфигурацию. Нарушение можно устранить одним из предложенных нами ранее способов коррекции его парных сравнений, получив положительное значение текущего главного минора S_k [7, 8]. Следующий минор S_{k+1} снова окажется отрицательным и потребует исправления. Всего потребуется скорректировать парные сравнения для q элементов множества. В этом смысле отрицательные собственные числа оказываются связанными с конкретными элементами множества или, другими словами, оказываются «локализованными» в матрице парных сравнений.

Рассмотрим процедуру, которая позволит получить оптимальную последовательность элементов множества. Известно, что определитель матрицы $S(n,n)$ равен произведению ее собственных чисел. Если он отрицателен, то количество собственных чисел нечетно, если положителен, то четно.

Рассмотрим главные миноры $S_k,\ k=n,\ldots,1$, в обратном порядке. Определим в матрице S_k такую строку и столбец i, что значение дополнительного минора $(S_k)_i^i$, $1\leqslant i\leqslant k$, образованного при их удалении, сменит знак по сравнению с S_k и окажется максимальным по модулю. Если знак не изменяется, то просто найдем такой дополнительный минор без смены знака. Пусть u — общее число таких шагов без смены знака дополнительного минора до локализации всех q смен знаков главных миноров.

Последовательность поочередно отброшенных строк и столбцов формирует оптимальную последовательность главных миноров $S_k,\ k=1,\ldots,n$ (и элементов множества, последовательно формирующих текущие миноры), в которой впервые отрицательный минор встретится не ранее, чем в момент $n-q-u+1$. Это означает, что полученная перестановка формирует такую матрицу $S(n,n)$, у которой придется корректировать парные сравнения не более, чем у $q+u$ последних элементов множества в оптимальной последовательности.

В общем случае при неоптимальной последовательности элементов приходится корректировать значительно большее число элементов множества, так как каждая очередная коррекция обычно порождает шлейф дополнительных коррекций.

Легко увидеть, что при отсутствии метрических нарушений будет получена локально оптимальная последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где их значения, оставаясь неотрицательными, убывают наиболее медленно (почти).

Рассмотрим матрицу корреляций $R(n, n)$. Можно заметить, что значение $\det R$ зависит от степени «ортогональности» конфигурации системы признаков: чем «ортогональнее» система признаков, тем ближе значение детерминанта к единице, и к нулю — в противном случае. Для $n = 2$ это очевидно, так как $\det R = 1 - r^2$. Для $n = 3$ в этом нетрудно убедиться, так как $\det R = 1 + 2r_{12}r_{13}r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2$, рассмотрев возможные значения парных коэффициентов корреляций, которые соответствуют конфигурациям без метрических нарушений, и т. д. С увеличением размерности n это эмпирическое свойство преимущественно сохраняется в целом, но, естественно, появляются возможности для взаимной компенсации достаточно высоких корреляций в усложняющихся формулах вычислений детерминантов, тем более для корреляций со знаками.

В этих условиях оказывается, что для метрически корректной матрицы $R(n, n)$ оптимальная последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где $S_1 = 1$ и $S_n = R(n, n)$, определяет локально оптимальную последовательность вложенных подмножеств «наиболее ортогональных» признаков. В начале такой оптимальной последовательности расположены «наиболее ортогональные» друг к другу и к остальным признаки, а к концу последовательности выстраиваются все «менее ортогональные» к остальным признаки, выбранные в последнюю очередь.

Корреляционная матрица $R(n, n)$ имеет статистический смысл, поэтому будем говорить об оптимальной последовательности вложенных подмножеств наименее коррелированных признаков. Отсюда легко увидеть, что первые m признаков в оптимальной последовательности должны образовать наименее коррелирующее подмножество из всех признаков, которое содержательно удобно интерпретировать как множество представителей m групп признаков.

Таким образом, процедура построения локально-оптимальной последовательности признаков позволяет решить задачу группировки на m групп (редукции размерности) без построения собственных направлений (для квадратов корреляций) или без построения центроидных направлений (для модулей корреляций).

5 Проблема начального разбиения в алгоритмах группировки

При решении задач группировки авторами было показано, что известный алгоритм экстремальной группировки на модулях коэффициентов корреляций («модуль») эквивалентен алгоритму k-средних, который представлен в модифицированной форме для обработки близостей [6]. В свою очередь, было показано, что такая модификация эквивалента классическому алгоритму k-средних для матрицы данных X в том смысле, что «внезапное» погружение элементов множества в пространство признаков не изменит результат разбиения. Такая кластеризация решает задачу построения центроидных факторов в задаче экстремальной группировки признаков (алгоритм «модуль»).

В свою очередь, алгоритм экстремальной группировки на квадратах коэффициентов корреляций («квадрат») решает задачу построения первых главных компонент для каждой группы сильно коррелирующих признаков. Тем самым решается известная задача факторного анализа как задача построения главных компонент или главных факторов

для, соответственно, нередуцированной $R(n,n)$ или редуцированной $\bar{R}(n,n)$ матриц корреляций.

Как и все процедуры кластер-анализа и группировки, процедура построения оптимальной последовательности также является локальной. Эксперименты показывают, что локальность процедуры построения оптимальной последовательности того же свойства, что и у процедур кластер-анализа и группировки. В частности, ожидаемым свойством оптимальной последовательности признаков обычно является устойчивое выделение от двух до пяти наименее (почти) коррелирующих признаков.

Таким образом, в силу локальности свойств известных процедур экстремальной группировки процедура построения оптимальной последовательности имеет самостоятельное значение в задаче группировки признаков.

Известно, что в процедурах с локальными свойствами важной проблемой является поиск начального решения (разбиения). Например, в задаче экстремальной группировки считается, что центроидные решения являются хорошим началом для группировок по собственным направлениям. Поэтому оптимальная последовательность признаков может рассматриваться как другой способ получения начального решения для алгоритмов экстремальной группировки, которое для заданной матрицы $R(n,n)$ является единственным. Это свойство представляется нам наиболее интересным.

6 Эксперименты

6.1 Программа экспериментов

Пусть L — число групп признаков G_i, $i = 1,\ldots,L$, где $|G_i| = n_i$, $\sum_{i=1}^{L} n_i = n$ и $r(X_j,F_i)$ — корреляция фактора $F_i = (f_{1i},\ldots,f_{Ni})^{\mathrm{T}}$ с признаком $X_j = (x_{1j},\ldots,x_{Nj})^{\mathrm{T}}$. Для количественной оценки качества группировок рассмотрим известные критерии I_Q для алгоритма «квадрат» и I_M для алгоритма «модуль»:

$$I_Q = \sum_{i=1}^{L}\sum_{j\in G_i} r^2(X_j,F_i)\,; \quad I_M = \sum_{i=1}^{L}\sum_{j\in G_i} |r(X_j,F_i)|\,.$$

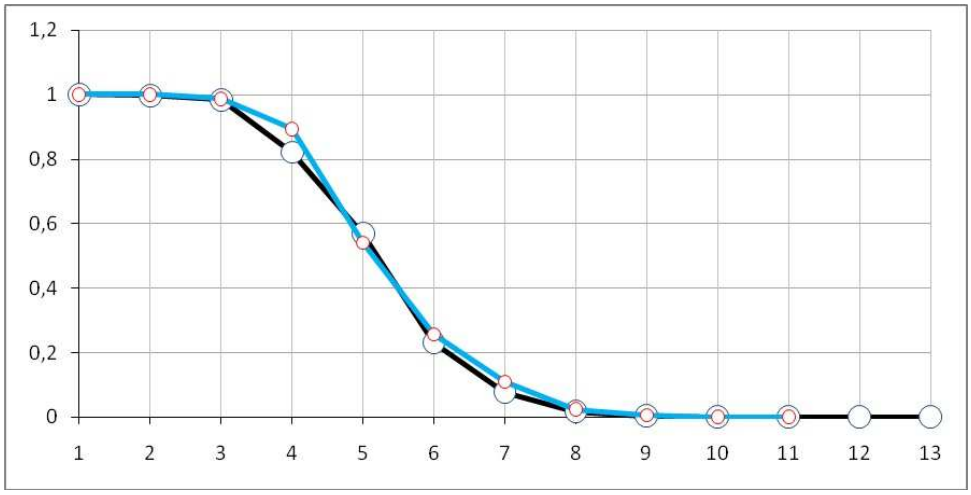
Алгоритмы экстремальной группировки имеют следующий вид.

Начальный шаг. Для L групп найти начальное разбиение каким-либо способом.
Шаг k.

1. В каждой группе G_i, $i = 1,\ldots,L$, образующей подматрицу $R(n_i,n_i)$, $i = 1,\ldots,L$, построить фактор $F_i = (f_{1i},\ldots,f_{Ni})^{\mathrm{T}}$ как главный или центроидный фактор или как первую главную компоненту.
2. Просмотреть все признаки и перенести каждый из них в ту группу, с фактором которой он коррелирует сильнее всего: $X_j \in G_p$, если $s(X_j,F_p) > s(X_j,F_i)$, $i = 1,\ldots,L$. Здесь $s(X_j,F_p) = |r(X_j,F_i)|$ или $s(X_j,F_p) = r^2(X_j,F_i)$.
3. Повторять шаги 1 и 2 до тех пор, пока группы не перестанут изменяться.

В данной работе программа экспериментов была направлена на исследование свойств оптимальной последовательности признаков, представляющей как начальное решение для алгоритмов экстремальной группировки, так и применяемой самостоятельно.

Рассматривались следующие начальные решения: первые L признаков в оптимальной последовательности; L минимально коррелирующих признаков как явная классическая альтернатива им; просто первые L признаков; случайно отобранные L признаков. Эти начальные решения порождали начальные разбиения для алгоритмов «квадрат» и «модуль»



Оптимальные последовательности значений главных миноров корреляционных матриц: 13 экономических показателей и 11 электрокардиограмм

в смысле критериев I_Q и I_M. Очевидно, что первые два начальных решения являются хорошими, вторые два — плохими. Эксперименты это подтвердили, поэтому здесь далее рассматриваются результаты только для хороших начальных решений.

Определение числа кластеров K является хорошо известной проблемой кластер-анализа. Одним из известных эвристических приемов для его выбора является определение границы, начиная с которой убывание критерия кластеризации до нуля (средневзвешенная дисперсия кластеров) при изменении K от 1 до N резко замедляется, где N — число объектов. Аналогично рассуждают и в задаче группировки. Очевидно, что при изменении числа групп $L = 1, \ldots, n$ критерии группировки I_Q и I_M возрастают до n, где n — число признаков. В этом случае также рассматривают границу, начиная с которой возрастание этих критериев резко замедляется.

Рассмотрим оптимальную последовательность главных миноров $S_k,\ k = 1, \ldots, n$, где их значения, оставаясь неотрицательными (если потребовалось, то после коррекции), убывают наиболее медленно. Оказывается, что в этом случае график изменения их значений приобретает характерный вид (см., например, рисунок).

В этом случае аналогичное эвристическое предположение о числе групп признаков предполагает рассмотрение области резкого падения значений главных миноров.

Также следует предположить, что при оптимальном числе групп должны получиться хорошо содержательно интерпретируемые факторы, характеризующие их.

Здесь представлены результаты экспериментов с некоторыми массивами данных.

6.2 Экономические показатели

Первый массив представляет собой данные Организации экономического сотрудничества и развития (Organization for Economic Cooperation and Development, OECD) [10] из сводного отчета за 2013 г. (Fastbook Country Statistical Profiles, 2013 edition) по 13 экономическим показателям 13 стран мира: Австралия, Франция, Германия, Италия, Япония, Корея, Мексика, Турция, США, Китай, Индонезия, Россия, ЮАР. Представлены следующие показатели:

1. Валовой внутренний продукт (ВВП) на душу населения (долл.)
2. Рост реального ВВП (%).
3. Прибыль, полученная в сельском хозяйстве, охоте и лесном хозяйстве, рыбалке (%).

4. Прибыль, полученная в промышленности, включая энергетические отрасли (%).
5. Прибыль, полученная в оптовых и розничных продажах, отелях, ресторанах, ремонте, транспорте (%).
6. Прибыль, полученная в финансовом посредничестве, недвижимости, арендных и деловых услугах (%).
7. Реальная прибыль, полученная в сельском хозяйстве, охоте и лесном хозяйстве, рыбалке (%).
8. Реальная прибыль, полученная в промышленности, включая энергетические отрасли (%).
9. Реальная прибыль, полученная в оптовых и розничных продажах, отелях, ресторанах, ремонте, транспорте (%).
10. Реальная прибыль, полученная в финансовом посредничестве, недвижимости, арендных и деловых услугах (%).
11. Общее потребление энергии (ТВт-ч).
12. Электричество, производимое ядерной энергетикой (ТВт-ч).
13. Доля электричества, производимого ядерной энергетикой, от общего объема (%).

Значения показателей прибыли в различных сферах активности представлены как с учетом общего уровня цен (реальная прибыль) в результате процессов инфляции-дефляции, так и без учета общего уровня цен (прибыль). Другие показатели связаны с уровнем ВВП и потреблением энергии. Статистические связи между уровнем ВВП, прибылью и энергетическими затратами представлены корреляционной матрицей $R(13, 13)$.

Как было сказано выше, на рисунке видно, что для экономических показателей число предполагаемых групп составляет 4–5.

Оптимальная последовательность признаков, построенная по матрице квадратов корреляций экономических показателей, имеет вид: [8, 5, 13, 9, 11, 4, 7, 12, 3, 1, 10, 6, 2].

Оптимальная последовательность признаков, построенная по матрице модулей корреляций экономических показателей, имеет вид: [10, 8, 5, 13, 12, 4, 11, 9, 2, 7, 3, 6, 1].

Результаты группировок показаны в табл. 1 и 2. Для каждого числа групп показаны начальные решения как представители, выбранные по разным принципам (минимально коррелирующие признаки и первые признаки из оптимальной последовательности). Также показаны результирующие группы и их представители. Изменение хотя бы одного начального представителя после перегруппировки означает, что начальное разбиение было улучшено. Если представители не изменились, то начальное разбиение не улучшилось. Номера признаков-представителей выделены жирным шрифтом.

Таблица 1 Группировки экономических показателей по критерию I_Q

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
	7	7	**7** 8 13	8	2	**2 8**
3	11	11	5 **11** 12	5	5	**5** 11
	6	10	1 2 3 4 **6** 9 **10**	13	3	1 **3** 4 6 7 9 10 12 **13**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
5	7	7	**7** 13	13	13	7 12 **13**
	6	10	1 2 3 4 **6** 9 **10**	9	10	1 2 3 4 6 **9 10**
	11	11	**11** 12	11	11	**11**

Таблица 2 Группировки экономических показателей по критерию I_M

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
3	6	10	1 2 3 4 **6** 9 **10**	10	1	**1** 2 3 4 6 9 **10** 12 13
	7	7	**7** 8 3	8	7	**7 8**
	11	11	5 **11** 12	5	5	**5** 11
4	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	7	7	**7** 8 13	8	8	**8**
	5	5	**5**	5	5	**5** 11
	11	11	**11** 12	13	13	7 12 **13**
5	6	10	1 2 3 4 **6** 9 **10**	10	10	1 2 3 4 6 9 **10**
	8	8	**8**	8	8	**8**
	5	5	**5**	5	5	**5**
	7	7	**7** 13	13	7	7 **13**
	11	11	**11** 12	12	11	**11 12**

Рассмотрим табл. 1. Для разбиения на три группы признаков результат по критерию I_Q неудовлетворителен, а именно: для начального разбиения по минимальным корреляциям признаки 2 и 8 после перегруппировки попали в разные группы. В то же время для начального разбиения по оптимальной последовательности эти два признака после перегруппировки оказались вместе в одной отдельной группе. Такая же ситуация сохраняется и для четырех групп признаков (в таблице не показана).

Но для пяти групп признаков результаты двух группировок практически одинаковы, а именно: представители разных групп для обоих вариантов начального разбиения обязательно входят в состав разных групп и после перегруппировки. В частности, признаки 2 и 8 также входят в составы разных групп в обеих группировках. В обоих случаях начальные группировки были улучшены, причем начальные представители остались в своих группах, даже если для них после перегруппировки были выбраны новые представители. Сами результирующие группировки минимально отличаются друг от друга признаком 12.

Такой результат хорошо соответствует ранее сделанному формальному предположению о пяти группах экономических показателей. Состав полученных групп позволяет содержательно интерпретировать их следующим образом (в порядке перечисления в табл. 1 показаны признаки, присутствующие в составе соответствующих групп одновременно в обоих разбиениях):

1) прибыль в промышленности с учетом энергозатрат (8);
2) прибыль в торговых и транспортных услугах (5);
3) прибыль в производстве натуральной продукции с учетом энергозатрат (7, 13);
4) ВВП и прибыль во всех сферах активности (1, 2, 3, 4, 6, 9, 10);
5) общее потребление энергии (11).

Рассмотрим табл. 2. Для разбиения на три группы по критерию I_M результаты похожи в том смысле, что все представители разных групп находятся в разных группах до и после перегруппировки. Для разбиения на четыре группы результат неудовлетворителен, так как в одной группировке признаки 5 и 11 представляют разные группы, а в другой группировке признаки 5 и 11 располагаются вместе и образуют отдельную группу.

Для пяти групп признаков разбиения полностью совпадают там, где признак 12 находится в одной группе вместе с признаком 11. Таким образом, и в этом варианте под-

Таблица 3 Качество группировок экономических показателей по критерию I_Q

Число групп	Минимальные корреляции		Оптимальная последовательность	
	Начальное разбиение	Результат	Начальное разбиение	Результат
3	5,1537	7,1308	3,5312	6,4304
5	7,0055	8,8924	**7,1196**	8,8002

Таблица 4 Качество группировок экономических показателей по критерию I_M

Число групп	Минимальные корреляции		Оптимальная последовательность	
	Начальное разбиение	Результат	Начальное разбиение	Результат
3	9,0739	9,0739	8,5066	8,8205
4	9,7296	9,7296	9,9426	9,9426
5	10,521	10,521	**10,521**	10,521

твердилась ранее предложенная интерпретация групп признаков (порядок перечисления групп соответствует табл. 1).

Рассмотрим качество группировок. Таблица 3 показывает, что для пяти групп начальное разбиение, полученное по оптимальной последовательности признаков, лучше, чем по минимальным корреляциям. Этот результат имеет самостоятельное значение, если экстремальная группировка по критерию «квадрат» не применяется. С другой стороны, здесь продемонстрирована локальность рассматриваемых процедур, заключающаяся в том, что лучшее по качеству начальное решение (оптимальная последовательность) не всегда обеспечивает лучший результат.

Таблица 4 также показывает, что для 4 и 5 групп экстремальная группировка по критерию «модуль» не улучшила начальное разбиение. В этом случае разбиение, полученное по оптимальной последовательности признаков, также имеет самостоятельное значение, так как сразу формирует окончательную группировку. Отметим, что в данном случае оптимальная последовательность формирует множество признаков, наиболее адекватно соответствующих предположению о наименьшей коррелированности. В целом можно отметить более сложное поведение критерия I_Q.

6.3 Электроэнцефалограммы

Второй массив представляет собой электроэнцефалограммы (ЭЭГ) биоритмов головного мозга, представленные корреляционной матрицей статистических взаимосвязей между энергетическими свойствами биоритмов для 11 частот: три частоты (1–3) представляют тета-ритмы, две частоты (4, 5) представляют низкочастотные (НЧ) альфа-ритмы, две частоты (6, 7) представляют высокочастотные (ВЧ) альфа-ритмы, четыре частоты (8–11) представляют бета-ритмы. Электроэнцефаллограммы были получены В. Д. Небылицыным в ходе исследований по эффекту навязывания ритма при светослуховом воздействии [11] на испытуемых.

Для ЭЭГ биоритмов головного мозга получается аналогичный график (см. рисунок) изменения значений главных миноров корреляционной матрицы $R(11, 11)$ биоритмов в оптимальной последовательности. Оказалось, что число предполагаемых групп также со-

Таблица 5 Группировки ЭЭГ по критериям I_Q и I_M

Число групп	Минимальная корреляция	Представители	Группы	Оптимальная последовательность	Представители	Группы
4	7	6	**6 7**	7	6	**6 7**
	2	3	1 **2 3**	2	3	1 **2 3**
	4	4	**4** 5	4	4	**4** 5
	10	10	8 9 **10** 11	8	10	**8** 9 **10** 11

Таблица 6 Качество группировок ЭЭГ по критериям I_Q и I_M

Критерий	Число групп	Минимальные корреляции		Оптимальная последовательность	
		Начальное разбиение	Результат	Начальное разбиение	Результат
Квадрат	4	6,5210	7,8048	6,1681	7,8048
Модуль	4	9,0942	9,0942	**9,0942**	9,0942

ставляет 4–5. Также очевидно, что содержательно ЭЭГ биоритмов должны быть представлены четырьмя группами ритмов разных типов (тета-, НЧ альфа-, ВЧ альфа- и бета-).

Оптимальная последовательность признаков, построенная по матрице корреляций ЭЭГ, имеет вид: [7, 4, 8, 1, 3, 11, 5, 9, 6, 10, 2].

Оптимальная последовательность признаков, построенная по матрице для квадратов корреляций ЭЭГ, имеет вид: [7, 2, 4, 8, 1, 5, 11, 3, 6, 9, 10].

Оптимальная последовательность признаков, построенная по матрице для модулей корреляций ЭЭГ, имеет вид: [7, 2, 4, 8, 1, 5, 11, 3, 10, 6, 9].

Отметим, что оптимальные последовательности признаков во всех случаях сразу же выделяют в качестве представителей четырех групп по одной частоте каждого типа. В качестве представителей пятой и т. д. групп выделяются частоты уже ранее представленного типа. В соответствии со смыслом критериев I_Q и I_M использовались оптимальные последовательности признаков для квадратов и модулей корреляций.

Результаты группировок показаны в табл. 5. Разбиения на четыре группы при разных способах получения начальных разбиений, а также по обоим критериям оказались полностью идентичными. Здесь сразу же показаны результаты только для четырех групп, так как для меньшего числа групп они не соответствуют содержательному смыслу, а для большего числа групп просто происходит расщепление содержательных групп ритмов разных типов.

Рассмотрим качество группировок. Таблица 6 показывает, что для 4 групп группировка по критерию «модуль» не улучшила начальное разбиение. В этом случае разбиение, полученное по оптимальной последовательности признаков, наиболее адекватно соответствует предположению о наименьшей коррелированности и также имеет самостоятельное значение, так как сразу формирует окончательную группировку.

В свою очередь, группировки по критерию «квадрат» были улучшены для обоих видов начальных разбиений, сформировав одну и ту же группировку. Таким образом, разбиение, полученное по оптимальной последовательности признаков, также адекватно соответствует предположению о наименьшей коррелированности признаков в искомой группировке. И снова продемонстрирована локальность рассматриваемых процедур, заключающаяся

в том, что лучшее по качеству начальное решение (минимальные корреляции) не всегда обеспечивает лучший результат. В целом также можно отметить более сложное поведение критерия I_Q.

7 Заключение

При решении задачи группировки возникает проблема содержательной интерпретации полученных факторов и групп признаков. Признаки, объединенные в группы, обычно поддаются содержательной совместной интерпретации. Тем не менее факторы групп являются синтетическими признаками, интерпретация которых может быть затруднена.

Часто после выделения групп признаков и построения соответствующих им факторов в каждой группе определяется ее представитель как наиболее сильно коррелирующий с фактором группы признак. Если рассматривать только множество таких признаков-представителей, то оказывается возможным содержательно интерпретировать результат группировки прямо в терминах исходных признаков.

При таком подходе все преобразования, выполняемые в соответствии с факторной моделью, являются промежуточными, так как в итоге выбираются представители из исходных признаков.

В данной работе предложен подход, позволяющий выбрать подмножество из исходных признаков, способных адекватно представить скрытые факторы, не требуя построения собственных и центроидных направлений в качестве промежуточного этапа преобразований. Данный подход основан на построении оптимальной последовательности признаков. В начале такой оптимальной последовательности расположены наименее коррелированные друг с другом и с остальными признаки, а к концу последовательности выстраиваются все более коррелированные с остальными признаки, выбранные в последнюю очередь.

Показано, что предложенный подход позволяет формировать начальное решение для известных алгоритмов группировки и также может применяться самостоятельно для оценки числа групп и построения содержательных группировок.

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. Т. 33. С. 5–68.

[2] *Lawley D.N., Maxwell A.E.* Factor analysis as a statistical method. — 2nd ed. — London: Butterworth, 1971. 117 p.

[3] *Harman H.H.* Modern factor analysis. — 3rd ed. — University of Chicago Press, 1976. 508 p.

[4] *Lumelskii V. Ya.* Parameter grouping on the basis of the square coupling matrix // Automat. Rem. Contr., 1970. No. 1. P. 117–127.

[5] *Cox T.F., Cox M.A.A.* Multidimensional scaling. — 2nd ed. — Chapman and Hall/CRC, 2000. 328 p.

[6] *Двоенко С. Д.* Кластеризация множества, описанного парными расстояниями и близостями между его элементами // Сиб. журн. индустр. матем., 2009. Т. 12. № 1. С. 61–73.

[7] *Двоенко С. Д., Пшеничный Д. О.* Устранение метрических нарушений в матрицах парных сравнений // Известия Тульского государственного университета. Технические науки, 2013. № 2. С. 96–104.

[8] *Двоенко С. Д., Пшеничный Д. О.* О локализации отрицательных собственных значений в матрицах парных сравнений // Известия Тульского государственного университета. Технические науки, 2013. № 9(2). С. 94–102.

[9] *Гантмахер Ф.Р.* Теория матриц — М.: Наука, 1988. 552 с.

[10] OECD statistics. OECD, 2013–2014. http://stats.oecd.org/.

[11] *Небылицын В.Д.* Основные свойства нервной системы человека — М.: Просвещение, 1966. 384 с.

Поступила в редакцию 19.08.2016

Feature grouping based on the optimal sequence of correlation matrix minors*

S. D. Dvoenko and D. O. Pshenichny

dsd@tsu.tula.ru; denispshenichny@yandex.ru

Tula State University, 92 Lenina pr., Tula, Russia

Background: It is known that data analysis problems usually arise in early stages of investigations, when a model of a phenomenon in researching has not been developed yet. Hence, it is too early to introduce a problem of a model identification. It needs to collect and study a lot of miscellaneous information about the most significant characteristics of a phenomenon under investigation in this case. Such a situation forces one to use inconsistent approach, since it is unknown what characteristics are important and what knowledge needs to be collected. Therefore, data analysis methods should resolve the contradiction and focus on the correct description of the phenomenon. The problem of informal interpretation of factors and groups arises in the grouping problem. Factors are synthetic features and difficulties can arise in informal interpretation of them. Therefore, groups and corresponding factors have been built, the representative usually is defined for each group as a feature, the most correlated with the group factor. As a result, it is possible to name groups informally as such initial features.

Methods: The new approach to specify a feature subset is proposed to represent correctly hidden factors. In this approach, it does not need to define eigenvectors or centroid ones as intermediate transformations. It is based on the optimal sequence of correlation matrix minors, since the less correlated features are placed at the beginning of the sequence and the more correlated ones are placed closer to the end of it.

Results: As it is shown, the proposed approach can produce initial partitioning for other grouping algorithms and additionally can be used to evaluate a number of groups and to get informal partitions.

Concluding Remarks: As it is evident, the natural hidden regularity in the phenomenon under investigation appears undoubtedly because of processing data by different techniques and algorithms targeted to uncover it. All such results as a whole will support the correct result. Therefore, it needs to support and develop the diversity of data processing intelligent methods. In this paper, an attempt to do it is presented. It is the relevant attempt since large volume of experimental data has been collected and methods for pairwise comparisons have been developed.

Keywords: *grouping; cluster; metrics; correlation; eigenvalue; eigenvector; determinant*

DOI: 10.21469/22233792.2.2.1

*The research was partially supported by the Russian Foundation for Basic Research (grants 15-07-02228, 15-07-08967, 14-07-00527, and 14-07-00964).

References

[1] Zhuravlev, U. I. 1978. Ob algebraicheskom podhode k resheniyu zadach raspoznavaniya ili klassifikatsii [About algebraic approach to solving problems of recognition or classification]. *Problemy kibernetiki* [Cybernetics Problems] 33:5–68.

[2] Lawley, D. N., and A. E. Maxwell. 1971. *Factor analysis as a statistical method*. 2nd ed. London: Butterworth. 117 p.

[3] Harman, H. H. 1976. *Modern factor analysis*. 3rd ed. University of Chicago Press. 508 p.

[4] Lumelskii, V. Ya. 1970. Parameter grouping on the basis of the square coupling matrix. *Automat. Rem. Contr.* 1:117–127.

[5] Cox, T. F., and M. A. A. Cox. 2000. *Multidimensional scaling*. 2nd ed. Chapman and Hall/CRC. 328 p.

[6] Dvoenko, S. D. 2009. Clustering and separating of a set of members in terms of mutual distances and similarities. *Trans. Machine Learning Data Mining* 2(2):80–99.

[7] Dvoenko, S. D., and D. O. Pshenichny. 2013. Ustranenie metricheskikh narusheniy v matritsakh parnykh sravneniy [The removing of metric violations in matrixes of pair comparisons]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering Sciences] 9(2):96–104.

[8] Dvoenko, S. D., and D. O. Pshenichny. 2013. O lokalizatsii otritsatel'nykh sobstvennykh znacheniy v matritsakh parnykh sravneniy [On localization of the negative eigenvalues for matrices of pairwise comparisons]. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering Sciences] 9(2):94–102.

[9] Gilbert, G. T. 1991. Positive definite matrices and Sylvester's criterion. *Am. Math. Mon.* 98(1):44–46. doi: 10.2307/2324036.

[10] OECD. 2013–2014. OECD statistics. Available at: `http://stats.oecd.org/` (accessed August 15, 2016).

[11] Nebylitsyn,V. D. 1966. *Osnovnye svoystva nervnoy sistemy cheloveka* [Main characteristics of the nervous system of a person]. 1966. Moscow: Prosveshchenie. 384 p.

Received August 19, 2016