УДК 519.25

Один вариант анализа соответствий для квантификации кластерной переменной

Дронов С.В.
Алтайский государственный университет, г. Барнаул
dsv@math.asu.ru

Аннотация

В случае, когда изучаемые объекты разбиты на кластеры, для построения более точных математических моделей удобно использовать искусственную переменную, которая каждому объекту ставит в соответствие его кластер. Эта кластерная переменная нуждается в переводе в числовую форму, т.е. в квантификации. Сегодня для решения этой задачи часто применяют алгоритм анализа соответствий. Он позволяет квантифицировать сразу пару нечисловых переменных по таблице их сопряженности. Но оказывается, метод перестает работать в случае, когда кластеры в задаче выделяются предельно четко, в частности, он склонен приписывать различным кластерам одинаковые метки. Поэтому актуальна задача его модификации. В работе обсуждаются несколько методов идентификации и последующего обхода формальных сбоев методики анализа соответствий для случая четко выделяющихся кластеров.

Ключевые слова: Кластерный анализ, оцифровка результатов кластеризации, квантификация кластерной переменной, анализ соответствий, проецирование точек многомерного пространства на прямую и плоскость

1. Введение. Задача оцифровки кластерной переменной

При обработке больших объемов данных часто возникает необходимость их кластеризации, т.е. разбиения исследуемых объектов на группы, объекты внутри которых можно считать похожими друг на друга, а объекты разных групп должны оказаться в большей степени различными. Если это в каком-то смысле так, группы называют кластерами. Таким образом, на интуитивном уровне задача построения кластеров поставлена, но при ее формализации возникает много проблем, поэтому, насколько известно автору, математически строгого определения кластера пока не предложено.

Тем не менее, предположим, что задача кластеризации успешно решена. Для каждого из образованных кластеров можно, например, строить свою математическую модель развития в нем каких-то процессов. Но, без сомнения, более удобной является ситуация, когда модель является общей для них всех. Нетрудно привести примеры, когда такая универсальная модель существует, но оказывается весьма неточной, хотя «частные» модели в рамках каждого из кластеров дают удовлетворительные практические результаты.

Можно попытаться решить эту проблему путем введения во все частные модели искусственной переменной, которая на объектах каждого из кластеров принимает уникальное значение, а при смене этого значения одна из частных моделей переходит в другую. На самом деле, после построения кластеризации подобная переменная возникает автоматически. Она представляет собой обозначение или условную метку того кластера, к которому отнесен данный объект, и, следовательно, обычно является нечисловой категорированной переменной. Для нее в [1] было предложено название кластерная переменная. И, конечно

же, для включения в модель этой переменной крайне желательно, чтобы она принимала числовые значения.

Задача придания кластерной переменной числовых значений или, иначе, построение числовых меток кластеров, называется задачей квантификации кластерной переменной. Для решения задачи такого типа нужен критерий качества выбираемых меток. Чаще всего при его построении принято привлекать ряд интуитивных соображений, например, необходимость естественного следования кластеров в порядке возрастания их меток или относительные величины различия между кластерами (расстояния между ними), которые после выбора меток оцениваются абсолютными величинами их разностей. Впрочем, в ситуации полной неопределенности возможен и противоположный подход: построим числовые метки кластеров из каких-то формальных соображений, а порядок следования и степени различия кластеров определим уже по готовым меткам.

Но, в любом случае, величины вводимых меток кластеров должны быть согласованы со значениями числовых показателей объектов, составляющих кластер или напрямую, или через призму строящейся математической модели. Некоторые из таких подходов были исследованы в [2], хотя можно подойти к этой задаче и совсем по-другому, см., например, [3].

2. Классический анализ соответствий как инструмент оцифровки

Широко применяется способ квантификации кластерной переменной, основанный на алгоритме анализа соответствий. Он, в частности, реализован в компьютерном статистическом пакете IBM SPSS Statistics [4], и может быть описан в первом приближении следующим образом.

Пусть на входе алгоритма задано кластерное разбиение изучаемого множества объектов на m кластеров, что, в частности, задает кластерную переменную, пока в нечисловой форме. Допустим, мы ставим своей задачей произвести квантификацию так, чтобы образом согласовать ее результат с некоторым числовым показателем Y, значения которого также известны для каждого из объектов. Это может быть как один из тех показателей, по которым строилось кластерное разбиение, так и их совокупность, или какой-то внешний показатель, полезный для последующего построения адекватной математической модели.

Произведем группировку значений Y на наших объектах, сообразуясь с близостью, похожестью этих значений. Алгоритмов для этого имеется достаточно большое количество (см. [5]), но нам далее конкретный способ не будет важен. Пусть получилось k групп объектов. Таким образом, с каждым наблюдаемым объектом оказались связаны два номинальных категорированных показателя – его кластер и группа, в которую попало значение Y для этого объекта. Рассматривая их значения как координаты клеток таблицы из k строк и m столбцов, заполним эту таблицу, называемую обычно таблицей сопряженности, указав в каждой из клеток число объектов, ей соответствующей. Алгоритм, известный под названием анализа соответствий, по таблице сопряженности конструирует метки для ее столбцов и строк, учитывающие степень их согласованности в этой таблице.

На самом деле этот алгоритм позволяет построить не только числовые, но и векторные метки для строк и столбцов таблицы в некотором общем искусственном пространстве, имеющем размерность не выше, чем $\min\{k-1, m-1\}$. В качестве базиса пространства выбираются, например, собственные векторы матрицы рассеивания T_1 нормированных профилей строк матрицы сопряженности, отвечающие ее достаточно большим собственным числам. Возможность изображения в том же пространстве также и столбцов связана с тем, что у матрицы рассеивания профилей столбцов T_2 набор ненулевых собственных чисел оказывается тем же, что у T_1. Это позволяет считать, что соответственные собственные векторы этих двух матриц связаны специальным линейным преобразованием. Построение меток столбцов осуществляется путем применения этого преобразования к векторам их

нормированных профилей в пространстве, натянутом на собственные векторы матрицы T_2.

При этом в силу нормированности профилей как строк, так и столбцов, данные всегда оказываются организованы так, что максимальные собственные числа обеих матриц рассеивания равны 1, и попытка в описанном построении задействовать собственные вектора, отвечающие этому собственному числу, приводит к тривиальным, совпадающим для всех рядов таблицы сопряженности, меткам (все профили ортогональны одному и тому же вектору, и проекции их концов на него все падают в одну и ту же точку). Именно этим объясняется невозможность построения меток полных размерностей – одно измерение, связанное с собственным числом 1, теряется.

Появление собственных векторов матриц рассеивания в алгоритме связано с тем, что, как это принято чаще всего в алгоритмах оцифровки и визуализации, он пытается построить в каком-то смысле наиболее удаленные друг от друга метки строк (и отдельно, столбцов) матрицы сопряженности.

В настоящей работе рассматривается ситуация, когда переменная Y предельно сильно связана с кластерным разбиением. Это значит, что каждый из объектов, значения Y на котором попадают в некоторые границы, обязательно оказывается элементом строго определенного кластера. Такое предположение, разумеется, не описывает основной случай, но нельзя сказать, что в конкретных практических задачах он не встречается.

К чему же приведет наличие такого случая при попытке применить методику анализа соответствий? После очевидной перенумерации кластеров таблица сопряженности перейдет в диагональную, – до перенумерации в каждой ее строке может располагаться лишь один ненулевой элемент, остальные элементы строки равны 0, а обе матрицы рассеивания превратятся в единичные матрицы (в одной из них может потребоваться исключение рядов, состоящих лишь из нулей). При этом, очевидно, все ненулевые собственные числа и T_1, и T_2 окажутся равными 1. Тем самым, классический анализ соответствий в этой ситуации исключит все собственные векторы обеих матриц, а значит, не сможет работать. Заметим, что все профили рядов каждой из матриц рассеивания в этом случае пропорциональны координатам базисных векторов соответствующего пространства, а, следовательно, после нормировки будут располагаться в концах соответствующих базисных векторов, отложенных от начала координат.

Если не отказываться от построения меток путем должным образом организованного проецирования, то мы приходим к следующей задаче: заданы точки $A_1, \ldots, A_p$, лежащие на концах единичных базисных векторов в p-мерном евклидовом пространстве. Это – вершины основания p-мерного единичного симплекса.

Основная задача. Разработать алгоритм оптимальной визуализации вершин основания p-мерного симплекса в пространствах размерности 1 и 2.

Наилучшей визуализацией с точки зрения решаемой задачи будет такая, когда изображения точек наиболее сильно удалены друг от друга. Это, кроме обычных доводов задач визуализации, оправдывается тем, что в нашей ситуации кластеры оказываются четко разделенными, и, следовательно, максимально различными. Итак, требуется спроецировать вершины основания симплекса на некоторую прямую (задача одномерной визуализации) или плоскость (двумерная визуализация) так, чтобы полученные проекции были максимально удалены друг от друга с точки зрения некоторого заранее выбранного критерия их взаимной несхожести ("неединости", если принять терминологию [6]). Обратимся к решениям для двух с нашей точки зрения наиболее естественных критериев "неединости".

3. Визуализация путем максимизации дисперсии

Сначала в качестве критерия рассмотрим величину разброса проекций относительно их центра. Начнем с задачи одномерной визуализации. Здесь максимизируемую величину

разброса, очевидно, можно отождествить с дисперсией полученных одномерных изображений. Без ограничения общности будем считать, что прямая, на которую будет производиться проецирование, проходит через начало координат и задается направленным вдоль нее вектором

$$\vec{a} \;=\; (a_1, ..., a_p), \quad \sum_{i=1}^{p} a_i^2 \;=\; 1.$$

Если рассмотреть на этой прямой систему координат, индуцированную существующей многомерной, то проекции изучаемых точек будут иметь координаты, равные соответственным координатам направляющего вектора. Обозначим через $\bar{a}$ их среднее арифметическое – это координата центра проекций на прямой. Тогда нужно подобрать координаты направляющего вектора так, чтобы максимизировать

$$D = \sum_{i=1}^{p} (a_i - \bar{a})^2 = \sum_{i=1}^{p} a_i^2 - p\bar{a} \;=\; 1 - p\bar{a}.$$

Таким образом, максимальное значение выбранного критерия достигается при произвольном задании координат вектора $\vec{a}$ так, чтобы их сумма была равна 0. Это условие на координаты направляющего вектора означает, что он может быть выбран произвольным перпендикулярным вектору $\vec{l} = (1, ..., 1)$, что хорошо объясняет полученный результат с качественной стороны.

Утверждение 1. *Наибольшая дисперсия проекций вершин p-мерного симплекса на прямую реализуется на прямой, направляемой произвольным единичным вектором $\vec{a}$, сумма координат которого равна нулю. Соответствующая одномерная визуализация может быть построена изображением на оси с выбранным началом координат и направлением точек с координатами $a_1, \ldots, a_p$ соответственно.*

Теперь рассмотрим случай проецирования точек $A_1, \ldots, A_p$ на (двумерную) плоскость в $\mathbb{R}^p$, проходящую через начало координат и натянутую на векторы $\vec{a} = (a_1, \ldots, a_p)$, $\vec{b} = (b_1, \ldots, b_p)$. При этом по-прежнему будем считать эти векторы имеющими единичную длину. Если дополнительно предположить их ортогональность, т.е. выполнение условия $\sum\limits_{i=1}^{p} a_i b_i = 0$, то в системе координат, в которой векторы $\vec{a}, \vec{b}$ являются базисными на плоскости, проекция B_i каждой из точек A_i будет иметь координаты (a_i, b_i).

На этот раз максимизации подлежит величина

$$D \;=\; \sum_{i=1}^{p} |B_i\overline{B}|^2,$$

где $\overline{B}$ – средняя точка проекций, координаты которой в выбранном базисе плоскости $(\bar{a}, \bar{b})$. Действуя полностью аналогично случаю проекции на прямую, получим

$$D \;=\; 2 - p((\bar{a})^2 + (\bar{b})^2).$$

Следовательно, для достижения максимальной величины избранного критерия необходимо и достаточно, чтобы сумма координат каждого из двух базисных векторов плоскости проекций равнялась бы нулю. При этом запомним, что по каждой координат минимизация получилась производящейся независимо.

Утверждение 2. *Наибольший разброс проекций вершин p-мерного симплекса на плоскость реализуется на (двумерной) плоскости в $\mathbb{R}^p$, натянутой на произвольные векторы $\vec{a} = (a_1, \ldots, a_p)$, $\vec{b} = (b_1, \ldots, b_p)$ имеющими единичные длины и нулевые суммы координат. Если эти векторы перпендикулярны, то соответствующая двумерная визуализации дается точками $B_i(a_i, b_i)$, $i = 1, \ldots, p$.*

4. Максимум минимальных расстояний

Изменим критерий оптимальности разброса точек. Теперь будем максимизировать величину

$$DM \;=\; \min_{i,j} |B_iB_j| \,. \tag{1}$$

Причины того, что производится именно максимизация минимального расстояния между точками обсуждались в [6]. При решении задачи одномерной визуализации снова будем выбирать прямую, направляющий вектор $\vec{a} \;=\; (a_1, \ldots, a_p)$ которой имеет единичную длину. Координаты проекций $B_i, i \;=\; 1, \ldots, p$ в индуцированной системе координат на этой прямой равны соответствующим координатам направляющего вектора. Без ограничения общности можно считать, что координаты пронумерованы нужным образом для того, чтобы $DM = |a_1 - a_2|$ и даже, более того, $a_1 \le a_2 \le \ldots \le a_p$. Тогда, в силу очевидного соотношения между минимумом и средним,

$$DM \le \;\frac{1}{p-1}\sum_{i=2}^{p} |a_i - a_{i-1}| \;=\; \frac{a_p - a_1}{p-1}\,,$$

и равенство достигается тогда и только тогда, когда все разности $a_i - a_{i-1}$ одинаковы по абсолютной величине, и, следовательно, проекции наших точек расположены на прямой с одинаковым шагом, т.е. найдется такое положительное a, что

$$a_i \;=\; a_1 + (i-1)a, \quad i = 2, \ldots, p,$$

При этом оказывается $DM = a$. Чтобы найти значение a, воспользуемся тем, что вектор $\vec{a}$ имеет единичную длину. Примем a_1 равным 0. Получаем

$$1 \;=\; \sum_{i=1}^{p} a_i^2 = a^2 \sum_{i=1}^{p} (i-1)^2 \;=\; a^2 \frac{(p-1)p(2p-1)}{6},$$

откуда

$$d \;=\; a \;=\; \sqrt{\frac{6}{p(p-1)(2p-1)}}\,. \tag{2}$$

Следовательно, прямая, на которую следует проецировать исходные точки, задается направляющим вектором

$$\vec{a} \;=\; (0, d, 2d, \ldots, (p-1)d). \tag{3}$$

Если необходимо указать точное положение такой прямой в $\mathbb{R}^p$ (хотя для решения нашей задачи этого не требуется), то следует потребовать, чтобы она проходила через ту из исходных точек $A_1, \ldots, A_p$, которая при проецировании на нее окажется левее всех. Построение этой прямой может быть осуществлено поочередным перемещением вектора (3) в каждую из них и анализом расположения проекций.

Все возможные направляющие векторы прямых с требуемым разбросом проекций вершин единичного симплекса на них исчерпываются векторами, получающимися перестановками координат построенного нами направляющего вектора (3), поэтому, в принципе, можно поступить проще – провести прямую через A_1 и перенумеровать точкив порядке расположения их проекций, или наоборот, переставить координаты направляющего вектора в соответствии с этим порядком.. Подводя итог, приведем точную формулировку.

Утверждение 3. *Наибольшее по величине минимальное расстояние между проекциями вершин основания p-мерного единичного симплекса на произвольную прямую реализуется на прямой, направляющий вектор которой имеет вид (3). Соответствующая одномерная визуализация представляет собой цепочку точек, расположенных в произвольном месте оси с шагом, задаваемым формулой (2).*

Перейдем к построению двумерных визуализаций с выбранным критерием (1). Отметим, что, задавая плоскость, на которую мы производим проецирование двумя перпендикулярными единичными векторами $\vec{a} = (a_1, \ldots, a_p)$, $\vec{b} = (b_1, \ldots, b_p)$ мы в качестве проекций рассматриваемых точек всегда получаем $B_i(a_i, b_i)$, $i = 1, \ldots, p$. Поэтому задача максимизации критерия DM будет осуществляться путем надлежащего выбора координат базисных векторов плоскости.

Примем за исходное предположение то, что, максимизируя минимальное расстояние между точками, мы максимизируем и минимальное расстояние между проекциями точек на каждую из осей координат. В основе этого может лежать, например, предположение, что при минимизации дисперсий соответствующих расстояний мы пришли к необходимости минимизации дисперсий каждой из координат точек-изображений. Повторно отметим, что перестановка координат вектора (3) не приведет к какому-либо изменению ситуации с минимальными расстояниями по этой координате, поскольку в каждую из точек вида $(id, 0)$ по-прежнему упадет ровно одна из координат точек-проекций. Следовательно, можно без ограничения общности считать, что первая (индуцированная) координата проекции i-й точки на искомую плоскость равна id для каждого из $i = 1, \ldots, p$, а вторая выбирается из этого же набора чисел, причем все вторые координаты точек должны быть различны. Необходимые сдвиги начала координат в полученном рисунке можно будет сделать позднее.

Таким образом, для выбора координат векторов, порождающих наилучшую плоскость проекций, мы приходим к следующей задаче.

Задача. *Найти перестановку* $(k_1, \ldots, k_p)$ *чисел* $1, \ldots, p$, *для которой*

$$Q = \min_{i,j} \left\{ (i-j)^2 + (k_i - k_j)^2 \right\} \tag{4}$$

достигает максимального значения.

После того, как эта задача будет решена, (двумерную) плоскость, на которую следует проецировать, нужно натянуть на векторы

$$\vec{a} = (d, 2d, \ldots, pd), \ \vec{b} = (k_1 d, \ldots, k_p d),$$

где d – подбираемый масштабирующий параметр, а оптимальными визуализациями следует (после еще одной очевидной смены масштаба) признать элементы набора точек с координатами (i, k_i), $i = 1, \ldots, p$.

К сожалению, задача аналитического выражения точного значения (4) оказалась слишком сложной, но для практических целей, когда число кластеров, как правило, не слишком велико, она успешно решается перебором возможных вариантов. При p=2 решение очевидно, для p от 3 до 9 оно задается таблицей 1.

5. Обсуждение и краткие выводы

Предполагая, что исходные кластеры объективно четко разделены, нами были рассмотрены два подхода к задаче визуализации и квантификации кластерной переменной при двух естественных критериях оптимальности. Качественная визуализация подразумевает максимальную «неединость» получаемых изображений, поэтому первым рассмотренным критерием качества являлась величина их рассеивания относительно соответствующего центра. Поскольку речь шла о модификации анализа соответствий, то задача решалась в предположении, что визуализации строятся путем специального проецирования, как и в модифицируемом методе. Ее полные решения в одномерном и двумерном случае содержатся в утверждениях 1 и 2.

Вторым критерием качества являлась степень достижения максимума минимального из попарных расстояний между точками. Для одномерного случае решение содержится в

Таблица 1

Вторые координаты визуализаций вершин основания p-мерного симплекса (первые координаты – всегда $1, 2, \ldots, p$)

p	Оптимальная перестановка	Q	Максимин (Qd)
3	(3,1,2)	2	0,8944
4	(3, 1, 4, 2)	5	1,3363
5	(5, 3, 1, 4, 2)	5	0,9129
6	(6, 4, 2, 5, 3, 1)	5	0,6742
7	(5, 2, 7, 4, 1, 6, 3)	8	0,8386
8	(8, 5, 2, 7, 4, 1, 6, 3)	8	0,6761
9	(7, 4, 1, 8, 5, 2, 9, 6, 3)	10	0,7001

утверждении 3. К сожалению, здесь не удалось получить полного аналитического решения при построении визуализаций на плоскости. Добавление ограничения возможности выбора визуализаций лишь из точек квадратной сетки с целыми координатами, тем не менее, позволило получить оптимальные решения при малых p. Очевидно, можно высказать весьма правдоподобную гипотезу о виде подобной визуализации и для больших p. В этом может помочь рисунок 1 ниже.

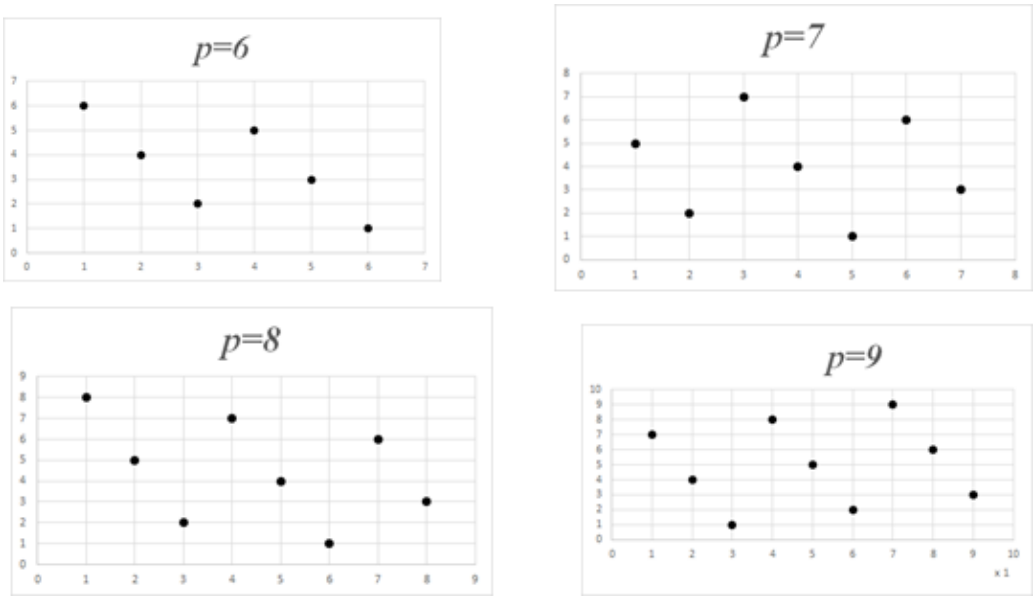

Рисунок 1. Оптимальные расположения проекций в прямоугольной сетке

Заметим также, что предложенные в таблице 1 оптимальные перестановки не являются единственно возможными – достаточно, например, симметрично отразить расположения точек на имеющемся рисунке. Тем не менее, точную формулировку гипотезы и попытки обосновать ее автор предпочитает временно отложить.

Еще одной нерешенной пока задачей является визуализация наших точек в произвольных точках плоскости, не обязательно расположенных в узлах прямоугольной сетки:

$$Q \;=\; \min_{a_i, b_j} \left\{ (a_i - a_j)^2 + (b_i - b_j)^2 \right\} \;\to\; max$$

при ограничениях

$$\sum_{i=1}^{p} a_i^2 \;=\; \sum_{i=1}^{p} b_i^2 \;=\; 1.$$

Аналитически данная задача оказалась практически не поддающейся решению, однако при $p = 3$ она все же решается сочетанием алгебраических и геометрических методов. Одним из возможных ответов будет правильный треугольник с вершинами в точках

$$A_1\left(-\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right);\ A_2\left(\frac{1}{\sqrt{5}}, -\frac{1}{\sqrt{5}}\right);\ A_3\left(\frac{\sqrt{3}}{\sqrt{5}}, \frac{\sqrt{3}}{\sqrt{5}}\right).$$

Как нетрудно вычислить, длина стороны этого треугольника (она же является максимином наших расстояний), равна $\sqrt{0.8}$. Это значение оказалось совпадающим с тем, что приведено в таблице 1 для $p = 3$, хотя при новом вычислении было снято ограничение на попадание проекций в один из узлов решетки. Видимо, из геометрических соображений можно пытаться решать и задачу при других значениях p, но тогда, скорее всего, значения последнего столбца таблицы 1 будут давать лишь оценки снизу для точных результатов.

Список литературы

1. Герасимова А.С., Дронов С.В. К проблеме оцифровки кластерной переменной // Анализ, Геометрия и топология. Труды Всероссийской молодежной школы-семинара. Ч.2. Барнаул, 2-4 октября 2013 г. — Барнаул : ИП Колмогоров И.А., 2013. — С. 54–58.

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

3. Жилин С.И. Решение задач дисперсионного и ковариационного анализа методом центра неопределенностей // Известия Алтайского государственного университета. — 2011. — № 1-2(69). — С. 54–57.

4. Системы анализа данных: IBM SPSS Statistics. [Электронный ресурс]. — URL: https://soware.ru/products/ibm-spss-statistics. Дата обращения 14.08.2024.

5. Grandars. Статистика. Общая теория статистики. Сводка и группировка статистических данных. [Электронный ресурс]. — URL: https://www.grandars.ru/student/statistika/gruppirovka-statisticheskih-dannyh.html. Дата обращения 14.08.2024.

6. Dronov S.V., Leongardt K.A. Multidimensional unfolding problem solution in the case of a single target // IOP Conf. Series: Journal of Physics: Conf. Series1210. — 2019. — 012034.