К вопросу практического применения алгоритмов кластерного анализа

On the issue of practical application of cluster analysis algorithms

УДК 004
Получено: 10.10.2023 Одобрено: 03.11.2023 Опубликовано: 25.12.2023

Силин А.В.
Канд. техн. наук, доцент, заведующий кафедрой ВТИТ, Новомосковский институт (филиал) Российского химико-технологического университета, Новомосковск, Россия
e-mail: silin.a.v@muctr.ru

Silin A.V.
PhD in Technical Sciences, Associate Professor, Head of the Department of VTIT, Novomoskovsk Institute (branch) of the Russian University of Chemical Technology, Novomoskovsk, Russia
e-mail: silin.a.v@muctr.ru

Силина И.В.
старший преподаватель кафедры ВТИТ, Новомосковский институт (филиал) Российского химико-технологического университета, Новомосковск, Россия
e-mail: silina.i.v@muctr.ru

Silina I.V.
senior lecturer of the department of VTIT, Novomoskovsk Institute (branch) of the Russian University of Chemical Technology, Novomoskovsk, Russia
e-mail: silina.i.v@muctr.ru

Аннотация
Данные в широком смысле означают фактический материал, поставляющий информацию для изучаемой проблемы и являющийся основой для обсуждения, анализа и принятия решений. Кластерный анализ представляет собой процедуру, на основе заданного правила объединяющую объекты или переменные в группы. В работе обеспечивается проведение группировки многомерных данных с помощью таких мер близости, как выборочный коэффициент корреляции и его модуль, косинус угла между векторами, евклидово расстояние. Группировка проводится по центрам, по ближайшему соседу и по выбранным эталонам. Программа написана в среде VS на языке C++.
Ключевые слова: кластерный анализ, кластер, функция расстояний между векторами, дендрограмма, матрица, мера расхождения, метод группировки множества объектов.

Abstract
Data in a broad sense refers to the factual material that provides information for the problem being studied and provides the basis for discussion, analysis and decision making. Cluster analysis is a procedure that, based on a given rule, combines objects or variables into groups. The work provides grouping of multidimensional data using such proximity measures as the sample correlation coefficient and its modulus, the cosine of the angle between vectors, and the Euclidean distance. Grouping is carried out by centers, by nearest neighbor and by selected standards. The program is written in the VS environment in C++.

Keywords: cluster analysis, cluster, function of distances between vectors, dendrogram, matrix, measure of divergence, method of grouping multiple objects.

Кластерный анализ традиционно относится к одной из развязностей классификационного анализа. Позволяет выполнить разбиение некого множества рассматриваемых объектов и их описаний с последующей группировкой [1] с образованием кластеров. Такой подход относится к многомерным статистическим методам, которые хорошо работают для больших объёмов исходных данных [2]. Долгое время, несмотря на разработанный математический аппарат, многомерные статистические методы не находили широкого применения из-за недостатка производительности компьютерных систем, позволяющих обработать большие массивы данных за приемлемое время.

Области применения техники кластеризации весьма разнообразны, но наиболее часто используются в экономике, а также в биологии, психологии, медицине. Причём, вне зависимости от особенностей областей применения, используется стандартный набор инструментов кластерного анализа [3].

Кластеры представляют собой группы однородности и задача кластерного анализа состоит в разбиении множества объектов на некоторое число кластеров на основании признаков исследуемых объектов так, чтобы каждый объект принадлежал только одной группе разбиения. Любой объект кластеризации представляется в виде точки в n-мерном пространстве признаков, а минимум расстояния между точками указывает на сходство между объектами. В предлагаемой авторами работе рассмотрена разработка пакета программ, реализующего как классические методы кластерного анализа [1, 3], так и позволяющего реализовывать различные специфические подходы обработки. Наблюдения и переменные (объекты) можно обрабатывать, используя различные меры расстояния и различные правила объединения кластеров.

Кластер-анализ можно охарактеризовать как метод группировки произвольного множества объектов X с помощью выбранной меры близости $\rho(x, y)$ объектов, в качестве которых рассматриваются точки x и y [2]. Группируемое множество точек может быть задано в виде матрицы

$$X = \begin{bmatrix} x_n & \cdots & x_{ln} \\ \vdots & & \vdots \\ x_{ml} & \cdots & x_{mn} \end{bmatrix},$$

столбцы которой определяют какие-то признаки, а строки - наблюдения, в которых фиксируются указанные признаки. Матрица X может рассматриваться как множество m - мерных векторов-столбцов (точек), или как множество n - мерных векторов-строк (точек).

В качестве функций расстояний $\rho(x, y)$ между векторами x и y в работе используются следующие соотношения.

1. Выборочный коэффициент корреляции между векторами x и y

$$\rho(x, y) = I_{x,y} = \frac{1}{m-1}\sum_{i=1}^{m}\frac{x_i-\bar{x}}{S_x} * \frac{y_i-\bar{y}}{S_Y}, \quad (1)$$

где $x=(x_1, \dots, x_m)$ и $y=(y_1, \dots, y_m)$ - векторы размерности m;

$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$, $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ - средние значения компонент векторов x и y;

$S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$, $S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$ - выборочные дисперсии компонент векторов x и y.

2. Модуль выборочного коэффициента корреляции между векторами x и y

$$\rho(x, y) = |I_{x,y}|, \quad (2)$$

3. Косинус угла между векторами x и y в m-мерном пространстве R^m

$$\rho(x,y) = \cos(\hat{x,y}) = \frac{(x,y)}{|x|*|y|}, \quad (3)$$

где $(x,y) = \sum_{i=1}^{m} x_i * y_i$ – скалярное произведение векторов *x* и *y*;

$|x| = \sqrt{\sum_{i=1}^{m} x_i^2}$, $|y| = \sqrt{\sum_{i=1}^{m} y_i^2}$ – длины векторов *x* и *y*.

4. Модуль косинуса угла между векторами *x* и *y* в пространстве R^m

$$\rho(x,y) = |\cos(\hat{x,y})|, \quad (4)$$

5. Евклидово расстояние между точками *x* и *y* в пространстве

$$\rho(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}, \quad (5)$$

6. Информационная мера расхождения, для которой все компоненты векторов *x* и *y* должны быть строго больше нуля.

$$\rho(x,y) = \frac{\chi * \Upsilon}{\chi + \Upsilon} * \sum_{i=1}^{m}\left(\frac{x_i}{\chi} - \frac{y_i}{\Upsilon}\right) * \ln\left(\frac{x_i * \Upsilon}{y_i * \chi}\right), \quad (6)$$

где $x = \sum_{i=1}^{m} x_i$, $y = \sum_{i=1}^{m} y_i$.

Группировка точек с помощью выбранной функции ρ в работе проводится тремя способами.

1. По центрам групп, способ при котором группировка точек множества *X* в естественные группы проводится по степени связи между точками и средними свойствами групп. Этот метод описан Парксом и состоит в следующем. Для *n* столбцов матрицы *X* с помощью функции $\rho(x_i, x_j)$ вычисляется матрица взаимных расстояний $R_n = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{bmatrix}$,

где $x_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{mi} \end{bmatrix}$, $x_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{mj} \end{bmatrix}$, *i*-ый и *j*-ый столбцы матрицы *X*;

$\rho(x_i, x_j)$ – расстояние между столбцами x_i и x_j.

В матрице *X* выбираются такие два столбца i_1 и j_1, которые имеют максимальную схожесть $\rho_{i1,j1}$ между собой. Если таких пар несколько, то выбирается первая из них. Далее, два выбранных столбца i_1 и j_1 заменяются в
матрице *X* одним, компоненты которого являются центром тяжести соответствующих компонент векторов x_{i1} и x_{j1}. Этому столбцу присваивается номер i_1, а столбец j_1 из матрицы *X* исключается. В результате будем иметь *n*-1 столбцов. Номера столбцов, объединяемых на первом шаге, и мера их сходства заносятся в первую строку матрицы $q = \begin{bmatrix} i_1 & j_1 & \rho_1 \\ \cdots & \cdots & \cdots \\ i_t & j_t & \rho_t \\ \cdots & \cdots & \cdots \\ i_{n-1} & j_{n-1} & \rho_{n-1} \end{bmatrix}$.

Для полученных *n*-1 столбцов вычисляется матрица R_{n-1} взаимных расстояний. Далее выбираются следующие наиболее схожие два столбца i_2 и j_2 и заменяются одним, соответствующим их центру тяжести. Новому столбцу присваивается номер i_2, а столбец j_2 из матрицы *X* исключается. Матрица *X* имеет на втором шаге группировки *n*-2 столбца. Номера столбцов i_2, j_2 и их уровень связи заносятся во вторую строку матрицы *q*. При объединении столбцов, которые были получены объединением других столбцов на предыдущих шагах группировки, их центр тяжести определяется с учетом количества векторов, которые они заменяют. После *n*-1 объединения получим всего один вектор-столбец, компоненты которого

являются центрами тяжести соответствующих компонент всех столбцов матрицы X. При этом будут заполнены n-1 строк матрицы q.

Этот метод дает хорошие результаты в том случае, если множество группируемых точек образует отдельные компактные группы по выбранной метрике ρ, а расстояния между этими группами достаточно велики по сравнению с размерами самих групп.

Если же точки не удовлетворяют указанному свойству, то при группировке, когда усредняются свойства отдельных групп, могут появиться точки со свойствами, отличными от множества свойств группируемых точек, что может привести к неверной интерпретации дендрограмм уровней связей [4]. Этот недостаток группировки по центрам групп устраняется в способе группировке по ближайшему соседу.

2. По ближайшему соседу, способ при котором группировка точек множества X в естественные группы проводится по расстоянию между границами групп точек. Этот вид группировки не изменяет набор свойств, объединяемых в группы точек. При объединении групп определяются меры сходства каждого элемента из одной группы с каждым элементом из другой группы. И те группы, которые содержат точки самые похожие, объединяются. В случае метрики (5) схожесть групп превращается в близость граничных точек. При этом объединяются те две группы, между границами которых расстояние минимально.

При определенных условиях этот способ группировки позволяет собрать в группы точки, подчиняющиеся одной и той же функциональной зависимости, в то время как группировка по центрам групп разбивает то же множество точек на группы с разными свойствами точек внутри каждой из них.

3. По эталонам, способ разбиения точек множества X на группы по силе их связи с эталонными точками. Эта группировка заключается в следующем. Задается эталонное множество в виде матрицы $E = \begin{bmatrix} e_{11} & \cdots & e_{1k} \\ \cdots & \cdots & \cdots \\ e_{m1} & \cdots & e_{mk} \end{bmatrix}$.

Рассмотрим группировку столбцов. Для каждого столбца x_i матрицы X определяется такой столбец эталонной матрицы E, который более всего схож со столбцом x_i. Номера s и i, а также мера сходства ρ_{si} заносятся в матрицу q.

После разнесения столбцов матрицы X по эталонам матрицы E получаем k групп. При этом некоторые группы могут быть пустыми, т.е. содержать только сам эталон и не содержать ни одного столбца из матрицы X. Столбцы матрицы X, попавшие в группу, образованную эталонным столбцом e_s, упорядочиваются по мере сходства с ним.

Матрицы X и E можно объединить в одну, указав, какие столбцы при этом будут считаться эталонными (7).

$$X' = X + E = \begin{bmatrix} x_{11} & \cdots & e_{11} & e_{12} & x_{12} & \cdots & e_{1k} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m1} & \cdots & e_{m1} & e_{m2} & x_{m2} & \cdots & e_{mk} & \cdots & x_{mn} \end{bmatrix}, \qquad (7)$$

Номера эталонных столбцов в объединенной матрице X' задаются массивом $l_э$ целых чисел, а номера эталонных строк массивом $d_э$.

Итак, результаты каждого из трех видов группировки задаются матрицей q, по которой в дальнейшем строится дендрограмма зависимости точек и их групп.

В пакете предусмотрена возможность проводить вычисления как для столбцов, так и для строк матрицы X, независимо от вида группировки и выбираемых мер близости. Работа со столбцами или строками задается признаком транспонирования матрицы X. Если признак равен нулю, то все вычисления, кроме Q - анализа, проводятся для столбцов матрицы X. При равенстве признака единице все вычисления, кроме указанного, проводятся для строк матрицы X. Если признак равен нулю, то Q - анализ проводится для строк матрицы X, а при равенстве признака единице - для ее столбцов.

Предоставляется возможность проводить группировку строк или столбцов, не только всей матрицы X, но также и любой ее подматрицы. Способ задания выборки подматрицы их X приведен ниже.

Перед группировкой выбираемая матрица, а если проводятся расчеты для всей матрицы, то и сама матрица X подвергается преобразованию. В пакете реализованы следующие преобразования.

1. Преобразование состоит в том, что каждая компонента вектора-столбца, делится на абсолютное значение максимальной по модулю его компоненты. При этом получается система векторов, концы которых лежат на гранях куба с длиной ребра равной 2 и центром в начале координат.

Например,

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 4 & 0.01 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.75 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

Эта нормировка «уравнивает» столбцы между собой, делает их более равноправными. Применяется в том случае, если столбцы соответствуют величинам, представленным в различных единицах измерения.

2. Преобразование состоит в том, что каждый элемент матрицы делится на абсолютное значение максимального по модулю элемента этой матрицы.

Например,

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 4 & 0.01 \end{pmatrix} \rightarrow \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & -0.75 & 0 \\ 0 & 1 & 0.0025 \end{pmatrix}.$$

Это преобразование сохраняет соотношение расстояний между точками и между длинами самих векторов, что особенно важно в некоторых случаях для проведения Q - анализа. Может применяться также для случая величин, имеющих одинаковые единицы измерения. При этом преобразовании концы векторов лежат на гранях прямоугольного параллелепипеда, помещенного в куб.

3. Преобразование заключается в том, что каждая компонента вектора - столбца матрицы делится на длину этого вектора.

Например, $X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 4 & 0.01 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.6 & 0 \\ 0 & 0.8 & 1 \end{pmatrix}.$

При этом преобразовании сохраняются угли между векторами. Сами вектора делаются более равноправными, чем при преобразовании 1, т.е. имеют одну и ту же длину. Их концы лежат на сфере с радиусом, равным единице и центром в начале координат.

Это преобразование может применяться для группировки величин, имеющих различные единицы измерения. Такая нормировка удобна также для функций (3) и (4).

Все приведенные выше три вида преобразования не влияют на вид группировки с использованием функций (1), (2), (3), (4). Однако, группировки с помощью функций (5) и (6) могут быть различными.

Выбираемая подматрица из X задается массивами целых чисел d и l. Массив d указывает номера выбираемых строк, а массив l - номера выбираемых столбцов. Рассмотрим примеры выбора строк.

Массив d имеет вид $d = (m, n)\ (d_i, \ldots\ldots, d_m)$ и состоит из $m+2$ целых чисел, где:

- m - число чисел, задающих номера строк;
- n - признак выбора. При n=1 указанные в массиве d строки выбираются из матрицы X и группировка проводится для выбранных строк. При n=0 указанные в массиве d строки выбрасываются из матрицы X и группировка проводится для оставшихся строк;
- $d_i, \ldots, d_m$ - числа, задающие номера строк.

Число элементов в массиве d не обязательно должно быть равно числу заданных строк, поскольку строки могут задаваться целыми интервалами. Рассмотрим возможные варианты задания массива d.

Пусть матрица X имеет размеры m = 200, n = 60. Нужно привести группировку с использованием строк (5,6,7,8,9,10,145,171,172,173). Эти строки можно указать также в виде (5-10,145,171-173).

Варианты задания выбора указанных строк будут следующие

1. d = (10,1) (5,6,7,8,9,10,145,171,172,173);
2. d = (5,1) (5, -10, 145, 171,-173);
3. d = (8,0) (1, -4,11,-144,146,-170,174,-200).

Таким образом, чтобы задать диапазон строк достаточно использовать два числа: первое из них – положительное, определяющее начало диапазона, второе – отрицательное, определяющее конец диапазона. Чтобы задать все выбираемые строки матрицы X, можно воспользоваться следующими массивами:

4. d = (2,1) (1,-200);
5. d = (0).

Выбранная матрица может быть распечатана с сохранением номеров строк и столбцов исходной матрицы.

Задание массива эталонных строк и массива эталонных столбцов аналогично заданию массива d. Разница состоит в том, что задание вида $d_э$=(0) и $l_э$=(0) означает отказ от эталонов соответственно строк или столбцов.

Пример. Пусть для матрицы X(200,60) заданы массивы.

d = (2,1) (5,-100);

l = (3,0) (20,-40,55);

$d_э$ = (3,1) (1,7,13);

$l_э$ = (0).

Это означает, что из матрицы X выбираются строки с 5 по 100 и из них выбрасываются элементы, принадлежащие столбцам с 20 по 40 и 55 столбцу. Далее в качестве эталонных строк выбираются строки 7 и 13, а первая эталонная строка будет отброшена, так как она не вошла в выбранную матрицу. Для столбцов в выбранной матрице эталонов не задано, поскольку $d_э$=(0). Таким образом, в этом примере группировка по эталонам возможна только для строк выбранной матрицы.

Отсутствие информации в матрице X задается числом $1xE^{10}$, означающим прочерк. Если строки матрицы, задаваемой массивами d и l, имеют хотя бы один прочерк, то такие строки исключаются из нее. Это намного облегчает работу с матрицей, так как не нужно менять массив d, исключать из него строки с прочерками и сохранять массив l для различных выбираемых групп строк матрицы X.

Анализ строк выбираемой матрицы проводится всегда до её преобразования по признаку транспонирования. В выбранной матрице могут оказаться строки или столбцы с одинаковыми элементами. В этом случае для некоторых видов анализа такие столбцы или строки исключаются из дальнейшего рассмотрения и выдается соответствующее сообщение, если число столбцов или строк стало меньшим двух.

Строки или столбцы матрицы X могут иметь названия. Использование названий облегчает анализ дендрограмм. Присваивать названия можно отдельным строкам или столбцам. Название можно занумеровать, указав номер той строки (столбца), которой оно принадлежит.

Ввод исходной информации в матрицу X может быть осуществлен или непосредственно, или из ранее подготовленного файла, набитого по столбцам.

В пакете программ могут быть заданы: способ группировки по центрам групп, по ближайшему соседу и по эталонам; вид нормировки выбранной матрицы, её транспонирование, количество дублей дендрограмм, а также выбранные группировки функции расстояний.

При вводе информации проверяются количественные ограничения, которым должны удовлетворять исходная и выбираемая матрицы. О любом невыполненном ограничении выдается соответствующее сообщение.

Предлагаемый в работе подход к анализу данных определяется как быстрым расширением парка компьютеров, так и развитием их математического обеспечения, а также тем, что выполнение многих реальных процедур без компьютеров просто невозможно. Следует добавить, что современные методы и процедуры анализа данных осуществляют проверку набора данных при помощи пакетов статистических программ, обработку отсутствующих наблюдений в многомерном случае.

Методы предлагаемого пакета программ были использованы для решения тех же задач, что реализованы и в пакетах Statictica, SPSS и показали сопоставимые результаты. В отличие указанных пакетов разработанный пакет программ является бесплатным, а его возможности могут быть расширены за счёт включения дополненных специфических методов для решения конкретных задач пользователя при разумном увеличении рабочих массивов для хранения анализируемых данных.

Планируется включить в разработанный пакет программ примеры для демонстрации наилучших способов использования предлагаемого программного обеспечения, таких как выбор подходящей функции определения расстояния, использования простых программ для сложного анализа, интерпретации вывода результатов работы пакета программ. На перспективу может ставиться задача расширения спектра алгоритмов, что позволит использовать разработанные программные средства для решения большего разнообразия задач и получение более адекватных результатов.

Список использованных источников

1 Е.В. Гублер Применение непараметрических критериев статистики в медико-биологических исследованиях / Е. В. Гублер, А. А. Генкин - Л.: Медицина, 1973. - 144 с.

2 С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; под ред. С.А. Айвазяна. Прикладная статистика. Классификация и снижение размерности - М.: Финансы и статистика, 2018. – 607 с., ил.

3 Alan Agresti. An Introduction to Categorical Data Analysis. Description: Third edition. Hoboken, NJ: John Wiley & Sons, 2019. – 390 p.

4 Олендерфер М. С., Блэшфилд Р. К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. - М.: Финансы и статистика, 2017. – 215 с.