Сизых Дмитрий Сергеевич

Сизых Наталья Васильевна

DOI 10.31483/r-103748

ФОРМИРОВАНИЕ ОДНОРОДНЫХ ГРУПП ОБУЧАЮЩИХСЯ МЕТОДОМ КЛАСТЕРНОГО АНАЛИЗА С ЦЕЛЬЮ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ПРОЦЕССА ОБУЧЕНИЯ

Аннотация*: актуальной является проблема формирования достаточно ровных групп обучающихся с целью повышения эффективности преподавания для качественного освоения учебного материала. В настоящее время имеется большой выбор инструментария для многомерного анализа данных, разнообразный и достаточно доступный для использования пользователями-педагогами. Однако описание методик для понимания процесса многомерной классификации (группировки), имеющееся в наличии в настоящее время, относительно сложное для преподавателей без математической подготовки. Поэтому цель данной работы состоит в том, чтобы представить достаточно понятное с методической точки зрения описание процесса многомерной классификации, в частности, – методики кластерного анализа.*

Ключевые слова*: группа обучающихся, процесс обучения, эффективность преподавания.*

Abstract*: an urgent problem is the formation of sufficiently equal groups of students in order to increase the effectiveness of teaching for the qualitative development of educational material. Currently, there is a large selection of tools for multidimensional data analysis, diverse and sufficiently accessible for use by teaching users. However, the description of techniques for understanding the process of multidimensional classification (grouping), currently available, is relatively difficult for teachers without mathematical training. Therefore, the purpose of this work is to present a description of the multidimensional classification process that is sufficiently understandable from a methodological point of view, in particular, the methods of cluster analysis.*

Keywords*: group of students, learning process, teaching effectiveness.*

Введение

В настоящее время актуальным является проблема формирования достаточно ровных групп обучающихся, с целью повышения эффективности преподавания для качественного освоения учебного материала. Данная проблема обострилась в настоящее время по причине того, что разница в уровне подготовки выпускников школ становится достаточно существенной по разным причинам сформированного социального общества. Кроме того, в вузы поступают студенты с различным уровнем знаний поскольку прием идет одновременно и по жесткому конкурсу на бюджет и по достаточно лояльному конкурсу на условиях оплаты обучения. Таким образом, если не формировать более однородные группы студентов, то процесс их обучения будет менее эффективным. К сожалению, уже не работает правило, что недостаточно подготовленные студенты будут стремиться дотянуться к хорошо подготовленным. И без формирования однородных групп страдает процесс обучения и хорошо подготовленным студентам не интересно учиться в «слабом темпе».

Существует множество методов и инструментов для формирования таких однородных групп. Но следует отметить, что наиболее эффективно формировать группы по набору нескольких показателей. Например, оценок по профильным предметам, учета интересов, бэкграунда и прочее. Инструментарий для многомерного анализа данных разнообразный и достаточно доступный для использования пользователями-педагогами имеется в настоящее время. Однако описание методик для понимания процесса многомерной классификации (группировки), имеющееся в наличии в настоящее время, относительно сложное для преподавателей без математической подготовки. Поэтому цель данной работы состоит в том, чтобы предоставить достаточно понятное, с методической точки зрения, описание процесса многомерной классификации, то есть методики кластерного анализа.

Кластерный анализ – многомерная статистическая процедура классификации и разбивки объектов (синонимы: единица, событие, образ, паттерн, предмет) в сравнительно однородные группы, которые называются кластерами (англ.

cluster – гроздь, скопление). Разбиение объектов проводится согласно значениям их признаков (синонимы: показатель, параметр, свойство, переменная, характеристика), которые представляет собой конкретное свойство объекта. При этом должно соблюдаться условие: формируемые кластеры должны быть однородными (гомогенными) внутри и разнородными (гетерогенными) по отношению друг к другу по заданным характеристикам. Таким образом, цель кластерного анализа, как метода изучения однородности сложных и неочевидно взаимосвязанных объектов, состоит в выделение кластеров из исследуемой совокупности объектов.

Методы кластеризации применяются в самых разнообразных областях: в экономике, маркетинге, рыночных исследованиях, социологии, медицине и пр. Достаточно часто данные методы применяются в экономике для принятия решений по управлению, планированию, прогнозированию, для анализа различных производственно-хозяйственных ситуаций. Кроме того, с помощью методов кластерного анализа можно анализировать временные ряды: сопоставлять периоды с близкими показателями, выделять группы показателей со схожей динамикой и пр.

Кластерный анализ выполняет следующие основные задачи:

– классификация данных, выявление кластерной структуры, схем группирования объектов, разработка типологии,

– формирование иерархической структуры выборки, *таксономический анализ, построение* древообразной иерархической структуры, возможность быстрой навигации и поиска данных*;*

– исследование и анализ данных: определение однородности, исследование и анализ отдельных кластеров данных, получение данных для принятия решений;

– сжатие данных, сокращение выборки данных, получение новой выборки из эталонных элементов;

– выявление новизны, выделение нетипичных объектов, которые не входят ни в один из кластеров, выявление кластеров с новыми свойствами;

– использование в качестве предварительной описательной стадии исследования, на которой определяется возможное наилучшее решение;

– использование в качестве предварительного этапа для дальнейших исследований и анализа данных, в частности для анализа больших данных Data Mining и пр.

Поставленная цель и задачи применения кластерного анализа непосредственно определяют используемый метод и алгоритм кластеризации. При этом учитывается количество признаков и их взаимосвязи, а также заданный критерий качества. В общем случае методы кластеризации базируются на трех общих принципах:

– *эвристический (нет формальной модели* изучаемого явления и критерия для сравнения различных решений, алгоритм базируется на интуитивных соображениях);

– *экстремальный (*не формулируется исходная модель, задается критерий качества разбиения на кластеры, наиболее эффективен для задач с четко определенной целью);

– *статистический (*имеется вероятностная модель исследуемого процесса, используется для теоретического исследования и воспроизводимости результатов).

Данные принципы лежат в основе множества различных моделей кластеризации. Выбор возможных вариантов формирования кластерных моделей, их размерности, смысловой интерпретации полученных результатов требуют творческого подхода исследователя, его профессионализма, интуиции.

Достоинства кластерного анализа:

– проводится разбиение объектов одновременно по нескольким различным признакам;

– нет никаких ограничений по виду кластеризуемых объектов;

– множество исходных данных может иметь практически произвольную природу (признаки имеют разнообразный вид);

– имеется возможность рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы информации, делая их компактными и наглядными.

Недостатки и ограничения кластерного анализа:

– позволяет обрабатывать лишь двухмерные массивы данных (объекты и их признаки), поэтому при необходимости проводится преобразование структуры исходного массива данных;

– состав и количество кластеров зависит от выбираемых критериев разбиения;

– при малом количестве кластеров имеется риск искажений и потери индивидуальных особенностей отдельных объектов за счет замены их обобщенными значениями;

– часто игнорируется возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров;

– на этапе кластерного анализа нет априорных гипотез относительно кластеров и проверка статистической значимости неприменима.

Особенности применения методов кластерного анализа:

– эвристический характер, требующий профессионализма и осторожности в выборе моделей кластерного анализа и их параметров для правильной трактовки полученных результатов;

– наиболее сложным является определение меры однородности объектов;

– для одних и тех же данных при применении разных кластерных методов можно получить различные результаты кластеризации;

– модели кластерного анализа самостоятельно формируют (привносят, навязывают) структуру в исходных данных, которая в некоторых случаях может не совпадать с реальной.

Часто кластерный анализ применяют совместно с факторным. Следует различать, что кластерный анализ направлен на разбиение объектов по совокупности их признаков на однородные группы, а факторный – на изучение связи между признаками, характеризующими исследуемые объекты. Таким образом, в кластерном анализе уменьшается количество объектов путем их группировки, при этом признаки (переменные) не группируются, а выступают в качестве критериев.

В общем случае, как факторный, так и кластерный анализ формируют однородные группы. В факторном анализе данные группы формируются относительно признаков на основании степени тесноты корреляционной связи, а в кластерном, как правило, формируются группы объектов на основании различных статистических мер, среди которых наиболее часто используется расстояние между признаками и кластерами. В кластерном анализе корреляция используется как мера сходства, а не различия, кроме того, нет потери данных. Кластерный анализ может с одинаковым успехом проводиться как для классификации объектов, так и классификации признаков (аналог факторного анализа). Когда проводится кластеризация признаков, термины «объект» и «признак» меняются местами.

Если объекты оцениваются большим количеством признаков, непосредственное проведение кластерного анализа с большим массивом данных представляется затруднительным или практически невозможным. При этом кластерный анализ используются в паре с факторным: на первом этапе с помощью факторного анализа снижается количество признаков, а затем с помощью кластерного анализа проводится группировка объектов. Разработано много направлений сочетания факторного и кластерного анализов, среди которых наиболее современный – лингвистический подход к обработке данных.

Наиболее общие классификации понятий и определений в кластерном анализе:

– различают параллельные, иерархические и функциональные процедуры;

– выделяют как внутренние методы кластерного анализа (признаки классификации равнозначны), так и внешние (существует один главный признак, остальные определяют его);

– внутренние методы можно разделить на иерархические (процедура классификация имеет древовидную структуру) и неиерархические;

– иерархические методы подразделяются на агломеративные (объединяющие) и дивизимные (разъединяющие).

В качестве общих характеристик кластеров рассматриваются следующие показатели:

– диаметр кластера, показывающий максимальное расстояние между любыми двумя точками кластеров;

– радиус кластера, показывающий максимальное расстояние от центроида до любой из точек кластера;

– плотность, определяемая как количество объектов в кластере поделенное на «объем»;

– межкластерное расстояние – расстояние между центрами, между ближайшими точками, среднее расстояние между парами объектов.

Многие характеристики кластеров используют понятие центра кластера (центроида). В евклидовом пространстве центроидом является среднее арифметическое точек кластера. В неевклидовом пространстве центроида нет, но при этом центром (кластроидом) выбирается одна из точек кластера, минимизирующая или максимальное расстояние до остальных точек, или сумму расстояний или сумму квадратов расстояний.

Теоретическая справка и модель

Процесс кластеризации состоит в разбиении множества объектов исходной выборки на подмножества непересекающихся кластеров таким образом, чтобы каждый объект принадлежал только одному кластеру. При этом объекты, принадлежащие одному и тому же кластеру должны быть сходными, а объекты, принадлежащие разным кластерам, должны быть разнородными. Таким образом, объекты разбиваются в кластеры на основе выбранной меры их схожести (различия), а сам процесс кластерного анализа заключается в *выявлении уровня схожести* всех исследуемых объектов и их последовательном *объединении в порядке возрастания уровня различия* между ними. Кроме меры схожести процесс кластеризации зависит и от ряда других факторов: используемой метрики *для оценки сходства* объектов; указанного априорно или оцениваемого количества кластеров. Поэтому выбор параметров кластеризации является неоднозначным и

достаточно неопределенным, во многом зависящим от целей кластеризации, вида и характеристик признаков объектов.

Исходные данные при проведении кластерного анализа задаются прямоугольной таблицей или матрицей признаков, в которой каждому объекту соответствует определенный набор признаков

$$X = \left\{ x_{ij} \right\}, (i = 1,2,...,\mathrm{n}), (\mathrm{j} = 1,2,...,\mathrm{m})$$

$$X = \begin{pmatrix} x_{11} & \dots & x_{ij} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & xij & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{pmatrix},$$

где x_{ij} – j-ый признак i-того объекта;

m – количество признаков, характеризующих исследуемые объекты;

n – количество исследуемых объектов для кластеризации.

При этом каждый из исходных n объектов рассматривается как точка в p-мерном пространстве признаков:

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{im} \end{bmatrix}, i = 1,2,...,n$$

Признаки объектов могут быть заданы как числовыми значениями, так и нечисловыми (порядковыми или номинальными, в частности, дихотомическими). Как правило, к исходным данным предъявляются требования их однородности и полноты. Все объекты должны быть одной природы и иметь сходный набор признаков. Если признаки заданы разными типами переменных, желательно их преобразовать и свести к какому-то одному типу. Вопрос полноты признаков решается в зависимости от цели кластеризации, выбранного метода и ряда иных подходов.

Поскольку признаки могут иметь разную размерность, разные порядки величин, различные приоритеты, то перед проведением процесса кластеризации

желательно нормализовать и стандартизировать значения признаков, что позволит привести все признаки к единому масштабу. Результаты кластерного анализа чувствительны к процессу стандартизации данных (может снижаться качество разбиения на кластеры) и к наличию «выбросов» (объектов со значениями признаков, значимо отличающихся от аналогичных признаков остальных объектов). При существенной разнице в размерностях признаков процесс стандартизации может снизить точность кластеризации. Иногда после нормировки могут появляться и возрастать «шумовые» эффекты каких-то признаков, что будет снижать дискриминирующие возможности других признаков. Следует отметить, что если кластерному анализу предшествует факторный анализ, то выборка признаков не нуждается в корректировке, поскольку стандартизация данных проводится на этапе факторного анализа. Кроме того факторный анализ формирует некоррелированные между собой факторы. Если имеется высокая корреляция между какими-то признаками, характеризующими объекты кластеризации, то возрастает влияние данных признаков на результаты распределения объектов по кластерам. Поэтому при проведении кластеризации необходимо проанализировать показатель взаимосвязи признаков и постараться избавиться от признаков с высоким показателем взаимной корреляции.

Существует много подходов для нормализации и стандартизации признаков. При кластеризации чаще всего используют z-стандартизацию, в которой стандартизованные переменные рассчитываются как отношение отклонений индивидуальных значений показателей от средней по показателю к среднеквадратическому отклонению показателя:

$$x_{il} \to \frac{x_{ij} - \tilde{x}_j}{\tilde{\sigma}_j}$$

где x_{il} – нормированное значение j-того признака для i-того объекта;

$\tilde{x}_j$ – среднее выборочное значение *j-того признака;*

$\tilde{\sigma}_j$ – выборочное среднеквадратичное отклонение *j-того признака*.

$$\tilde{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\tilde{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \tilde{x}_j)^2}$$

Таким образом, переходим от исходной матрицы признаков *X* к нормированной матрице. Данная нормировка данных проводится к нулевому среднему значению признака и к единичной дисперсии.

Иногда для стандартизации используют минимальные и максимальные значения признаков, например:

$$x_{il} \to \frac{x_{ij} - \min\limits_{j} x_{ij}}{\max\limits_{j} x_{ij} - \min\limits_{j} x_{ij}}$$

Рассмотрим вопросы оценки результатов и качества кластеризации. Наиболее простыми способами оценки качества кластеризации являются следующие показатели, характеризующие свойства кластеров:

Плотность распределения объектов внутри кластера (скопление точек в пространстве данных), которая может быть оценена *дисперсией расстояния от центра кластера до каждого из его объектов. Плотность кластера больше, если* меньше значение дисперсии, а, значит, ближе к центру кластера находятся его объекты.

Размер кластера, который может оцениваться радиусом, если кластер имеет круглую форму или является гиперсферой, в противном случае используется более сложная оценка.

– *отделимость кластеров друг от друга, т.е. их локальность, которая оценивается* взаимной удаленностью кластеров друг от друга и степенью их возможного перекрытия в многомерном пространстве признаков.

В кластерном анализе используют *функционал качества разбиения Q(S), с помощью которого можно сравнить качество различных способов разбиения.* При этом качество разбиения оценивается экстремумом данного функционала

(минимум или максимум). Наиболее часто используются следующие функционалы качества разбиения:

– *сумма попарных расстояний* между объектами внутри кластера, или среднее внутрикластерное расстояние (должно быть минимально):

$$Q1 = \sum_i \sum_{x,y \in X_i} d(x,y) \to \min$$

Можно использовать и внутрикластерные дисперсии расстояний.

среднее межкластерное расстояние

$$Q2 = \sum_{i<j} \sum_{x \in X_i, y \in X_j} d(x,y) \to \max$$

суммарная выборочная дисперсия разброса элементов относительно центров кластеров (кластерных центроидов) используется для метрического линейного пространства:

$$Q3 = \sum_i \frac{1}{|X_i|} \sum_{x \in X_i} d^2(x, c_i) \to \min$$

где $c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$ – центр кластера X_i.

Иногда используется несколько критериев одновременно или соотношения критериев качества кластеризации, например соотношение внутрикластерной дисперсии к межкластерной и пр. Часто в качестве оценки связанности кластеров используют отношение среднего внутрикластерного расстояния к межкластерному:

$$Q_{i,j} = \frac{Q1_i + Q1_j}{2Q2_{i,j}}$$

где $Q1_i$, и $Q1_j$ – средние внутрикластерные расстояния классов i и j; $Q2_{i,j}$ – среднее межкластерное расстояние между этими же кластерами. Функция принятия решения (оценочная функция) имеет вид:

$$S = \frac{1}{k} \sum_{i=1}^{k} \max_j Q_{i,j},$$

где k – количество кластеров.

По значению *S* часто определяется оптимальное количество кластеров в иерархическом анализе. Наилучшему разбиению объектов на кластеры соответствует минимальное значение функции S. Оценочная функция S равна 1, если все различия между объектами равны между собой или, если все объекты соединены в один кластер. Локальные минимумы функции S позволяют выявить разные уровни объединения (под- и над- структуры).

При выборе показателя качества разбиения объектов на кластеры используются и эмпирические соображения. Например, хорошим считается разбиение на кластеры, в котором имеется значительное отличие средних внутрикластерных показателей от общего среднего значения (для оценки значимости различий применяется t-критерий Стьюдента). В некоторых случаях, когда невозможно формализовать цель кластеризации, в качестве критерия качества используется возможность содержательной интерпретации полученных кластеров.

Наиболее сложным в кластеризации, как и в любой классификации объектов, является выбор меры их сходства. Существует и может быть предложено множество различных способов оценки меры сходства объектов: по расстоянию, по корреляции, ассоциативности (для бинарных и номинальных признаков), по различным статистическим и вероятностным оценкам и пр. (Снит и Сокэлом, 1973). При этом могут учитываться как технические, так и физические показатели, априорно задаваться различные весовые показатели и приоритеты, учитываться статистические особенности распределений признаков, шкалы их измерений и пр. Количественное оценивание однородности сводится к введению понятия метрики. От выбора данной метрики зависит окончательный вариант разбиения объектов на кластеры с учетом принятого алгоритма. Поэтому выбор меры близости объектов проводится эмпирически в зависимости от целей кластеризации, особенностей объектов и характеризующих их признаков, статистических особенностей распределения *признаков и пр.* Наиболее часто однородность объектов кластеризации определяется либо вычислением расстояний, либо заданием некоторой решающей функции, характеризующей степень близости объектов.

В задачах классификации и, в частности, кластерного анализа, существует необходимость оценивать меру близости как отдельных объектов между собой, так и групп объектов (кластеров) между собой или объекта и кластера. Поэтому рассмотрим раздельно две группы методов, используемых для оценки мер близости.

Как правило, в практике кластерного анализа в качестве меры оценки однородности объектов используется расстояние между объектами (метрика расстояния, метрика меры близости объектов, меры сходства), которое определяется расстоянием между векторами признаков. *Мера близости объектов, т.е. расстояние d*(*x*, *y*) *между объектами x и y* в пространстве признаков, характеризующих объекты кластеризации с введенной метрикой (в метрическом пространстве), должно удовлетворять следующим условиям (аксиомам):

– неотрицательность $d\,(x, y) \geq 0$;

– симметрия $d\,(x, y) = d\,(y, x)$;

– *неразличимость идентичных объектов: расстояние для двух идентичных объектов равно нулю,* при $x=y$ имеем $d\,(x, y) = 0$

– неравенство треугольника (метрическое неравенство)

– $d\,(x, y) \leq d\,(x, z) + d\,(z, y)$;

– различимость нетождественных объектов: при $x \neq y$ имеем $d\,(x, y) \neq 0$

Рассмотрим наиболее известные метрики кластерного анализа:

Евклидово расстояние – отражает геометрическое расстояние в многомерном пространстве и показывает среднее различие между объектами:

$$d(x,\mathrm{y}) = \sqrt{\sum_{j=1}^{m}(x_j - y_j)^2}$$

где m – количество признаков $j=1,2,\ldots,m$.

Используется для количественных признаков однородных с точки зрения физического смысла, с одинаковым весом и распределением, близким к нормальному. Показатель евклидова расстояния зависит от различий между координатными осями. Геометрически оно лучше всего объединяет объекты в

шарообразных скоплениях. Если признаки распределены криволинейно, то применяется матрица взвешенного евклидова расстояния. На практике используют различные варианты евклидова расстояния.

Квадрат евклидова расстояния применяется с целью придать больше веса более отдаленным друг от друга объектам;

$$d(x,y)=\sum_{j=1}^{m}(x_j-y_j)^2$$

– *взвешенное евклидово расстояние, в котором задаются весовые коэффициенты, позволяющие повысить* степень важности отдельных признаков при кластеризации*:*

$$d(x.y)=\sqrt{\sum_{j=1}^{m}\omega_j(x_j-y_j)^2}$$

где ω_j значение весов определяется дополнительным исследованием и находится в диапазоне $0< \omega_j <1$.

Нормализованное (нормированное) евклидово расстояние используется для признаков, имеющих различные единицы измерения и значительное различие по величине. При вычислении евклидова расстояния для выравнивания влияния признаков в случае различающихся дисперсий признаков, при различном масштабе данных можно использовать нормализованные данные.

Хэммингово расстояние (манхэттенское расстояние, расстояние городских кварталов) используется при наличии выбросов и определяется как среднее значение разностей по координатам, снижая влияние выбросов:

$$d(x,\mathrm{y})=\sqrt{\sum_{j=1}^{m}\left|x_j-y_j\right|}$$

Может использоваться как мера различия объектов, имеющих признаки, заданные в дихотомической шкале. В данном случае расстояние определяется как число несовпадений значений соответствующих признаков рассматриваемых двух объектов.

Расстояние Чебышева используется в случае признания объектов разными, если они различаются по какой-либо одной координате:

$$d(x,y)=\max_{1\le j\le m}\left|x_j-y_j\right|$$

Обобщенное расстояние Колмогорова – Минковского является степенным расстоянием с использованием увеличения или уменьшения веса признаков и применяется в случае, когда соответствующие объекты сильно различаются:

$$d(\mathrm{x},\mathrm{y})=\left(\sum_{j=1}^{m}\left|x_j-y_j\right|^{p}\right)^{\frac{1}{r}}$$

где p – задается априорно как постепенное взвешивание разностей по отдельным координатам $p > 1$,

r – задается априорно как прогрессивное взвешивание больших расстояний между объектами.

Таким образом, можно прогрессивно увеличить или уменьшить вес соответствующего признака.

Если оба значения r и p равны двум, то полученное расстояние совпадает с расстоянием Евклида.

При p=1 имеем Хеммингово расстояние, которое часто используется и для дихотомических признаков.

Если признаки имеют разные порядки величин, то может использоваться весовая метрика Минковского.

Расстояние Канберра учитывает процесс нормирования измеряемых величин и является неинвариантной величиной относительно сдвига векторов:

если $x \neq 0$ или $y \neq 0$, то

$$d(x,y)=\sum_{j=1}^{m}\frac{\left|x_j-y_j\right|}{\left|x_j\right|+\left|y_j\right|}$$

Расстояние Махаланобиса используется в случае зависимых признаков и их различной значимости и связано с корреляциями признаков, которые могут влиять на адекватное разбиение объектов. В случае наличия корреляции признаков,

координатные оси, по которым определяется расстояние, могут рассматриваться как неортогональные, т.е. не направлены под прямыми углами друг к другу. В физическом смысле расстояние Махаланобиса показывает расстояние между координатами рассматриваемого объекта (точка в пространстве) и центром масс делённое на ширину эллипсоида в направлении рассматриваемой точки. Центр тяжести (центр масс) – это центроид кластера, определенный положением точки, представляющей средние значения для всех признаков в многомерном пространстве, признаков рассматриваемой модели. Объект принадлежит тому кластеру, до которого расстояние Махаланобиса минимально. Таким образом, расстояния Махаланобиса вычисляется как:

$$d(x) = \sqrt{(x-\mu)S^{-1}(x-\mu)^T}$$

или как мера различия между двумя случайными векторами с равными распределениями

$$d(x,y) = \sqrt{(x-y)S^{-1}(x-y)^T},$$

где $x = (x_1, x_2, ..., x_n)$ является вектором признаков объекта, $\mu = (\mu_1, \mu_2, ..., \mu_n)$ является вектором средних значений признаков, а S – матрица ковариаций.

Если матрица ковариации является единичной, то расстояние Махаланобиса становится равным расстоянию Евклида. Если матрица ковариации диагональная (но необязательно единичная), то расстояние Махаланобиса становится равным нормализованному расстоянию Евклида.

Расстояние с использованием коэффициента корреляции:

$$d(x,y) = 1 - \left| r_{xy} \right|$$

Расстояние Брея-Картиса применяется для признаков, представленных в номинальных и ранговых шкалах (с предварительной стандартизацией):

$$d(x,y) = \sum_{j=1}^{n} \frac{\left| x_j - y_j \right|}{\left| x_j + y_j \right|}$$

Расстояние по проценту несогласия применяется для признаков, представленных в категориальных данных:

$$d(x,y) = (количествоx_j \neq y_j) / j$$

Расстояние по косинус-методу базируется на определении косинусов значений векторов признаков.

Расстояние по корреляции Пирсона базируется на оценке корреляции значений векторов признаков.

Кроме перечисленных выше мер расстояния между объектами кластеризации разработано и применяется множество специфических мер, например для двух дихотомических переменных применяется метод Лямбда (Lambda), применяются методы Shape, Hamann или Anderbergs's D., для переменных с номинальным типом шкал можно использовать методы χ^2 (Chi-square measure) и φ^2 (Phi-square measure).

Кроме представленных выше технических мер расстояния и подходов к их модификации для оценки меры близости объектов можно использовать некоторые физические показателя, например при кластеризации отраслей народного хозяйства с целью агрегирования можно использовать матрицы межотраслевого баланса.

В ряде программных систем для проведения кластерного анализа имеется возможность для пользователя задавать свою меру расстояния (настроенное расстояние).

По выбранной метрике расстояний строится матрица расстояний

$$X = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$$

Таким образом, исходные данные для проведения кластерного анализа могут описываться матрицей признаков, матрицей расстояний или матрицей сходства.

Для определения расстояния между объектом и кластером, между двумя кластерами применяются различные меры близости. Количество и разнообразие методов и подходов для определения расстояний между кластерами намного превышает возможности вычисления расстояния между двумя объектами в

многомерном пространстве признаков. При этом учитывается, что кластеры имеют определенный объем в многомерном пространстве, имеют протяженность и состоят из многих точек в отличие от точек, характеризующих объекты. Методы оценки меры близости кластеров в основном базируются на следующих трех способах определения расстояний [Лепский А.Е., Броневич А.Г.]:

– определение расстояния до центра кластера используется для хорошо определяемых кластеров (компактных), т.е. кластеров с хорошими показателями качества описания и не высокой ошибкой классификации. Пример классификации по двум и трем классам приведен на рисунке 1.

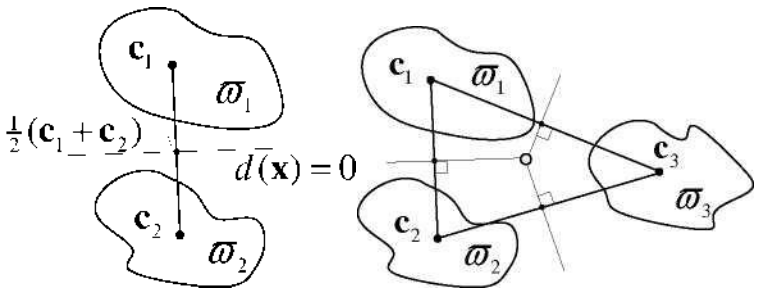

Рис. 1. Кластеризация по двум и трем кластерам методом определения расстояния до центра кластера

При кластеризации по двум классам решающая функция линейна, а разделяющая поверхность – прямая линия с уравнением $d(x) = 0$ проходит через середину отрезка, соединяющего центры классов и перпендикулярна к нему. При кластеризации по трем классам границами классов выступают серединные перпендикуляры между центрами классов, а точка пересечения этих перпендикуляров является центром окружности, описанной вокруг центров классов. При кластеризации по трем и более кластерам все пространство признаков с заданными в нем центрами классов разбивается на отдельные области, которые называются *клетками Вороного,* а множество всех клеток Вороного – *диаграммой Вороного.*

– способ постепенного присоединения объектов (способ ближайшего соседа) применяют в том случае, когда цена ошибки неправильной кластеризации велика, но ошибки в исходных данных должны быть невелики, поскольку

данный способ чувствительный к значениям отдельных данных, которые могут оказаться ошибочными;

– определение расстояния до эталонного объекта применяется при большом значении дисперсии признаков объектов кластера. В данном случае кластер описывается несколькими эталонными объектами, вокруг которых хорошо группируются объекты кластера. Вычисляется расстояние между анализируемым объектом и ближайшим к нему эталонным объектом кластера.

Все применяемые методы объединения объектов в кластеры строятся на перечисленных выше способах оценки расстояний между кластерами. Наибольшее применение нашли следующие методы объединения объектов в кластеры:

– метод одиночной связи (ближайшего соседа, single linkage) состоит в том, что расстояние определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах, при этом сужается пространство признаков и кластеры объединяются по ближайшей границе. Поскольку кластеры формируются постепенным присоединением объектов, то в итоге получается длинная цепочка объектов, то есть «волокнистые» кластеры. При этом кластеры объединяются вместе только отдельными элементами, случайно оказавшимися ближе остальных друг к другу. В данном методе расстояние определяется по формуле:

$$d_{i,s\oplus t} = 1/2 * d_{is} + 1/2 * d_{it} - 1/2 * \left| d_{is} - d_{it} \right|,$$

где $d_{i,s\oplus t}$ – расстояние между i-тым кластером и объединенным $s \oplus t$ кластером;

d_{is} – расстояние между кластерами i и s;

d_{it} – расстояние между кластерами i и t.

– метод полной связи (наиболее удаленных соседей, complete linkage) состоит в том, что расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах. При этом пространство признаков растягивается и кластеры объединяются по дальней границе. Данный метод непригоден для «цепочных» протяженных кластеров. Расстояние определяется по формуле:

$$d_{i,s\oplus t} = 1/2 * d_{is} + 1/2 * d_{it} + 1/2 * \left| d_{is} - d_{it} \right|.$$

Решение принимается, как и в методе ближнего соседа, по минимальному значению расстояния.

– метод невзвешенного попарного среднего расстояния (UPGMA unweighted pair-group method using arithmetic averages) состоит в том, что расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод не изменяет пространство признаков (объекты объединяются в соответствии с расстоянием до центра класса) и является эффективным и для «цепочного» типа кластеров, и для удаленных кластеров. Расстояние оценивается по формуле:

$$d_{i,s\oplus t} = \frac{n_s}{n_s + n_t} * d_{is} + \frac{n_t}{n_s + n_t} * d_{it},$$

где n_i, n_s, n_t – число объектов соответственно в кластерах i, s, t.

– метод взвешенного попарного среднего (WPGMA weighted pair-group method using arithmetic averages) идентичен предыдущему методу, но при вычислениях расстояний используется весовой коэффициент, равный размеру соответствующих кластеров (числу объектов, содержащихся в них). Поэтому данный метод более эффективный в том случае, когда формируются неравные размеры кластеров.

– невзвешенный центроидный метод (UPGMC unweighted pair-group method using the centroid average) состоит в том, что расстояние между двумя кластерами определяется как расстояние между их центрами тяжести:

$$d_{i,s\oplus t} = \frac{n_s}{n_s + n_t} * d_{is} + \frac{n_t}{n_s + n_t} * d_{it} - \frac{n_s}{n_s + n_t} * \frac{n_t}{n_s + n_t}$$

– взвешенный центроидный метод (медианный, WPGMC weighted pair-group method using the centroid average) идентичен предыдущему, за исключением того, что при вычислениях расстояний используются веса для учёта разницы между размерами кластеров (числом объектов в них). Метод эффективен при значительных различиях в размерах кластеров.

– метод Варда использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод состоит из множества этапов, на которых минимизируется сумма квадратов для любых двух кластеров, которые могут быть сформированы на каждом этапе. Таким образом, постепенно объединяются кластеры, которые в наименьшей степени повышают гетерогенность внутри формируемых кластеров. Данный метод является достаточно эффективным, но способствует формированию кластеров малого размера.

В общем случае результат кластерного анализа представляется набором кластеров, состоящих из множества исследуемых объектов, но в некоторых случаях, когда объект может быть отнесен к нескольким кластерам, указывается вероятность его принадлежности к кластеру.

Алгоритмы и практический пример

В общем случае методы кластерного анализа предполагают выполнение следующих основных этапов:

Постановка проблемы кластеризации: определение целей

Выбор данных для кластеризации согласно поставленной цели и имеющегося массива данных. Данные выбираются как по объектам, так и по признакам. Оценка и анализ статистических показателей по выбранным данным, выяснение коррелированности признаков и пр. При необходимости выполнение стандартизации и нормализации признаков (выбор подходящего метода). Принятие решений по результатам проведенного анализа и формирование признакового пространства для кластеризации объектов.

Выбор метрики для оценки сходства объектов, внутрикластерного сходства и межкластерного различия наблюдаемых объектов и кластеров. Выбор критерия качества для оценки результатов кластеризации.

Выбор метода проведения кластерного анализа.

Проведение процесса кластеризации. Оценка критерия качества. Если заранее не было задано количество кластеров, то принимается решение о количестве кластеров и процесс кластеризации повторяется.

Проверка достоверности результатов кластерного решения.

Оценка, анализ и интерпретация результатов кластеризации.

Отдельной проблемой кластерного анализа является установление оптимального количества кластеров. При этом возможны следующие три ситуации:

– число классов априори известно;

– число классов заранее неизвестно и подлежит оценке;

– число классов заранее неизвестно, но его определение и не входит в условие задачи; требуется только построить дендрограмму.

Метод и алгоритм кластеризации, необходимый для качественного решения поставленной цели выбирается в зависимости от количества признаков, их взаимосвязи, подобранного критерия качества. Методы кластеризации строятся на следующих подходах: вероятностный, искусственного интеллекта, логический, графовый, статистический и пр. В общем случае используемые кластерные методы можно разделить на несколько групп:

– иерархические;

– итеративные;

– поиска модальных значений плотности;

– по факторным данным;

– методы сгущений;

– с использованием графов.

Наиболее часто для экономических исследований и бизнес-анализа используются иерархические и итеративные методы (как пример, к-средних), поэтому в данном пособии наиболее подробно рассмотрим эти методы.

Иерархические методы кластеризации (древовидная кластеризация, иерархическое дерево) используют алгоритмы последовательной пошаговой группировки объектов по показателю близости друг к другу на основе матрицы расстояний. Таким образом, при помощи многошагового процесса формируются кластеры по данным меры сходства или расстояния между объектами. Чем меньше расстояние, тем формируется большее число классов. Поскольку на каждом шаге иерархической кластеризации требуется вычисление матрицы расстояний, то для

практической реализации данного процесса необходима емкая машинная память и большое количество времени, которое растёт пропорционально третьей степени количества наблюдений. Это основной недостаток иерархических процедур и поэтому его реализация для кластеризации большого количества объектов с большим числом признаков является нецелесообразной и может быть даже не реализуемой. Однако следует отметить достаточно высокую точность иерархических методов кластеризации.

Многошаговая иерархическая кластеризация в большинстве случаев проводится в два этапа: на первом этапе определяется оптимальное число кластеров для разбиения объектов, а на втором – проводится кластеризация по ранее установленному количеству кластеров.

Разделяют алгоритмы иерархической кластеризации на агломеративные (чаще применяемые на практике) и дивизимные. *Агломеративный* алгоритм *(объединяющий, метод слияния, формирование кластеров снизу вверх)*, рассматривает на первом шаге все объекты в качестве кластеров, которые на последующих шагах увеличиваются путем объединения ДО тех пор, пока не будет сформирован единственный кластер, содержащий все объекты. Дивизимные методы являются противоположными агломеративным по процедурам кластеризации. *Дивизимный* алгоритм (*разделяющий, метод дробления, построение кластеров сверху вниз)* рассматривает на первом шаге один кластер, в котором объединены все исследуемые объекты, на последующих шагах данный кластер поэтапно делится на более мелкие кластеры ДО тех пор, пока будет сформировано необходимое количество кластеров. *Таким образом,* на каждом шаге количество кластеров возрастает, а мера расстояния между ними уменьшается. В отличие от агломеративных методов дивизимные не требуют пересчета матрицы расстояний на каждом шаге классификации.

Практическая реализация любого иерархического алгоритма, не взирая на различное количество, последовательность и метрику выделяемых кластеров, предусматривает выполнение следующих пяти действий:

1. Постановка цели, выбор и предварительный анализ данных. Задание критерия качества.

2. Выбор метрики для оценки сходства объектов и построение матрицы расстояний.

3. Выбор метрики для внутрикластерного сходства и межкластерного различия наблюдаемых объектов и кластеров.

4. Выбор метода проведения иерархического кластерного анализа и выполнение процедуры кластеризации. Если априорно не задано оптимальное количество кластеров, то процедура кластеризации проводится в два этапа.

5. Проверка достоверности результатов кластерного решения. Оценка, анализ и интерпретация результатов кластеризации.

Проанализируем применение методов иерархического кластерного анализа на реальном примере.

Используя данные базы Евростат, изучим особенности развития различных стран Евросоюза по показателю инновационности предприятий []. Одним из этапов для достижения данной цели является выполнение кластеризации стран Евросоюза по показателю Инновационности предприятий (исходные данные см. табл. Приложение 1). В данном примере признаками инновационности предприятий стран Евросоюза являются показатели, приведенные в базе Евростат:

– число предприятий, использующих какой-либо вид инноваций (% от общего числа предприятий; рассматриваются малые от 10 сотрудников, средние и большие предприятия);

– продукция инновационных предприятий из общего объема, %;

– организационные инновации из общего числа организационных мероприятий, %;

– инновационные процессы от общего объема производственных процессов, %;

– инновационный маркетинг из общего числа маркетинговых мероприятий, %.

Кластеризацию стран Евросоюза по показателю инновационности предприятий будем проводить различными методами с целью изучения их особенностей и применимости:

– иерархический агломеративный метод в два этапа с определением оптимального количества кластеров по исходным показателям базы Евростат;

– иерархический агломеративный метод по признакам, полученным в результате проведения факторного анализа;

– иерархический агломеративный метод кластеризации показателей базы Евростат;

– итеративный метод k-means по исходным показателям базы Евростат;

– итеративный метод k-means по признакам, полученным в результате проведения факторного анализа.

Общий сравнительный анализ приведен в конце данного раздела.

Проиллюстрируем алгоритм иерархического кластерного анализа на примере кластеризации стран ЕС по инновационности предприятий.

Пример 1.

Для расчетного примера выберем 6 стран, а в качестве признаков возьмем факторы, полученные в результате факторного анализа. В качестве объектов возьмем страны: Дания, Франция, Финляндия, Швеция, Австрия, Португалия, а признаки – укрупненные факторы Инновационное производство и Инновационное управление и маркетинг. Исходная таблица данных для рассматриваемого примера имеет вид:

Таблица 1

Исходные данные

Объекты		Признаки	
Обозначение	Страна	Инновационное производство	Инновационное управление и маркетинг
1	Дания	0,19805	0,45875
2	Франция	0,27982	0,47377
3	Финлян-дия	0,9568	0,04895
4	Швеция	1,02114	0,07407

5	Австрия	0,39664	0,72057
6	Португа-лия	0,36248	0,83139

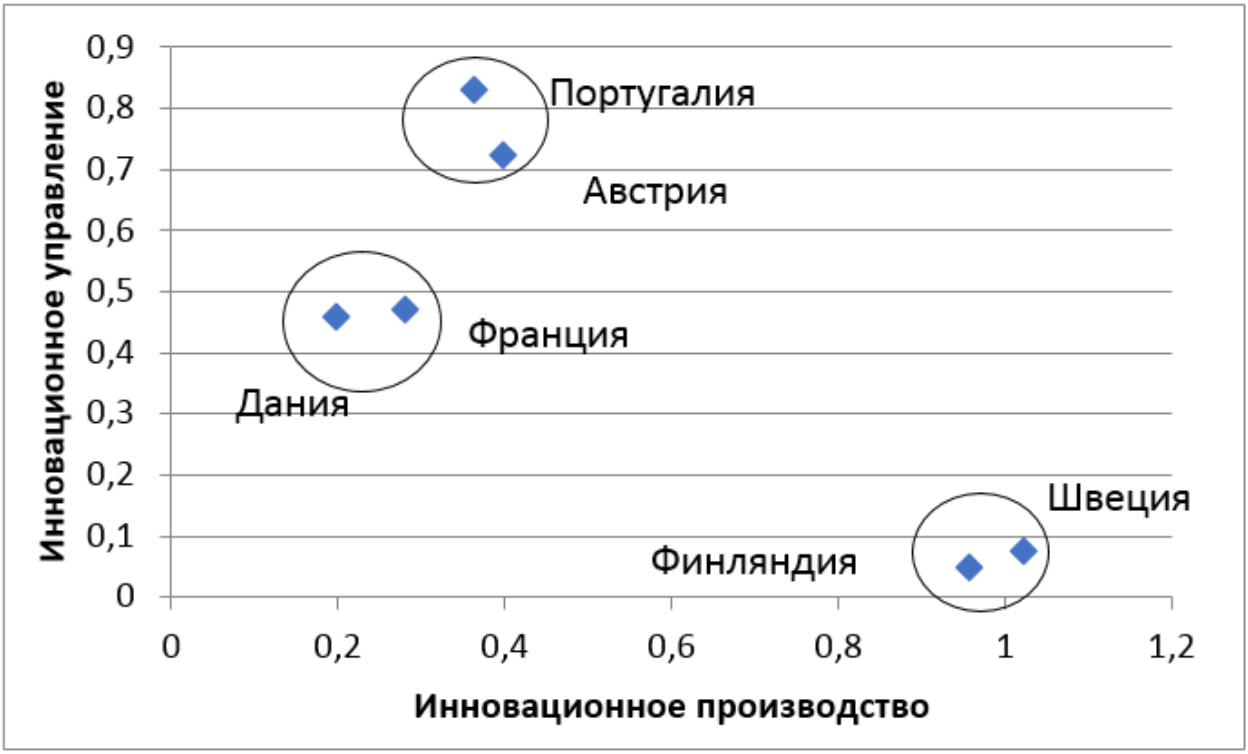

Рис. 2. Расположения стран в двумерном пространстве признаков примера 1

Анализ исходных данных наглядно показывает, что можно выделить три кластера, в которые входит по два объекта:

– Финляндия и Швеция (объекты 3 и 4) имеют высокий показатель инновационного производства, но очень низкий уровень по инновационному управлению;

– Австрия и Португалия (объекты 5 и 6) имеют высокий уровень инновационного управления, но средний уровень инновационного производства;

– Дания и Франция (объекты 1 и 2) имеют низкий уровень инновационного производства и средний уровень инновационного управления.

Поскольку признаки являются результатом факторного анализа, то не требуется их стандартизация, кроме того, данные показатели некоррелированные. Значения признаков значимо не различаются между собой. Исходя из этого, выберем в качестве меры близости признаков обычное евклидово расстояние. В

качестве метода кластеризации используем иерархический агломеративный (объединительный) метод. В качестве меры близости кластеров – метод ближайшего соседа, хотя, оптимальным для данного примера может являться метод дальнего соседа, поскольку в данном случае нет «цепочных» протяженных кластеров.

Проанализируем алгоритм данного метода на приведенном примере.

1. Определим расстояние между объектами и построим матрицу близости объектов:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

$$d_{11} = \sqrt{(0{,}1981 - 0{,}1981)^2 + (0{,}459 - 0{,}459)^2} = 0$$

$$d_{12} = \sqrt{(0{,}1981 - 0{,}2798)^2 + (0{,}459 - 0{,}474)^2} = 0{,}083138$$

$$d_{13} = \sqrt{(0{,}1981 - 0{,}9568)^2 + (0{,}459 - 0{,}049)^2} = 0{,}86234$$

$$d_{14} = \sqrt{(0{,}1981 - 1{,}0211)^2 + (0{,}459 - 0{,}074)^2} = 0{,}908546$$

$$d_{15} = \sqrt{(0{,}1981 - 0{,}3966)^2 + (0{,}459 - 0{,}721)^2} = 0{,}3286$$

$$d_{16} = \sqrt{(0{,}1981 - 0{,}3625)^2 + (0{,}459 - 0{,}831)^2} = 0{,}407306$$

Аналогично находим расстояния между всеми шестью объектами и строим матрицу расстояний (будем записывать в виде таблиц).

Таблица 2

Начальная таблица расстояний между объектами
(соответствует расстоянию между кластерами)

кластеры	(*1*)	(*2*)	(*3*)	(*4*)	(*5*)	(*6*)
(*1*)	0	0,083	0,862	0,9085	0,3286	0,407
(*2*)	0,083	0	0,799	0,842	0,273	0,367
(*3*)	0,862	0,799	0	*0,069*	0,875	0,9826
(*4*)	0,9085	0,842	0,069	0	0,899	1,0037
(*5*)	0,3286	0,273	0,875	0,899	0	0,1159
(*6*)	0,4073	0,367	0,9826	1,0037	0,1159	0

Поскольку выбран агломеративный вариант иерархического метода, то на начальном первом этапе считаем объекты одиночными кластерами, а расстояния между объектами будут являться расстояниями между кластерами. По методу ближайшего соседа объединение кластеров проводим последовательно, начиная от кластеров с наименьшим расстоянием. Поскольку наименьшее расстояние составляет d_{34}=0,069, то будем на первом шаге объединять кластеры 3 и 4.

Расстояние между кластерами по методу ближайшего соседа рассчитывается как:

$$d_{i,s\oplus t} = 1/2 * d_{is} + 1/2 * d_{it} - 1/2 * \left| d_{is} - d_{it} \right|$$

Таким образом, объединяем кластер s=3 и кластер t=4 и пересчитаем расстояния:

$$d_{1,3\oplus 4} = 1/2 * d_{13} + 1/2 * d_{14} - 1/2 * \left| d_{13} - d_{14} \right| =$$
$$= 0,5 * 0,862 + 0,5 * 0,9085 - 0,5 * \left| 0,862 - 0,9085 \right| = 0,862$$

$$d_{2,3\oplus 4} = 1/2 * d_{23} + 1/2 * d_{24} - 1/2 * \left| d_{23} - d_{24} \right| =$$
$$= 0,5 * 0,799 + 0,5 * 0,842 - 0,5 * \left| 0,799 - 0,842 \right| = 0,799$$

Подобным образом пересчитываем все расстояния и получаем матрицу на один порядок меньше предыдущей:

Таблица 3

Таблица расстояний между кластерами после первого объединения

	(*1*)	(*2*)	(*3,4*)	(*5*)	(*6*)
(*1*)	0	*0,083138*	0,862344	0,328615	0,407306
(*2*)	0,083138	0	0,799233	0,273052	0,367049
(*3,4*)	0,862344	0,799233	0	0,874559	0,982562
(*5*)	0,328615	0,273052	0,874559	0	0,115965
(*6*)	0,407306	0,367049	0,982562	0,115965	0

Поскольку в новой матрице минимальным расстоянием между кластерами является d_{12}=0,083, то будем на втором шаге объединять кластеры 1 и 2 и проведем пересчет расстояний аналогично расчету на первом шаге. В результате получим матрицу 4*4:

Таблица 4

Таблица расстояний между кластерами после второго объединения

	(1,2)	(3,4)	(*5*)	(*6*)
(1,2)	0	0,799233	0,273052	0,367049
(3,4)	0,799233	0	0,874559	0,982562
(*5*)	0,273052	0,874559	0	*0,115965*
(*6*)	0,367049	0,982562	0,115965	0

d_{56}=0,115965

Далее объединяем 5 и 6 кластеры и получим матрицу 3*3:

Таблица 5

Таблица расстояний между кластерами после третьего объединения

	(1, 2)	(3, 4)	(5, 6)
(1,2)	0	0,799233	*0,273052*
(3,4)	0,799233	0	0,874559
(5,6)	0,273052	0,874559	0

$d_{(1,2),(5,6)}$=0,273052

На 4-м шаге объединяем кластеры (1, 2) и (5, 6), получим матрицу 2*2:

Таблица 6

Таблица расстояний между кластерами после четвертого объединения

	(1, 2), (5, 6)	(3,4)
(1, 2), (5, 6)	0	*0,799233*
(3, 4)	0,799233	0

$d_{(1,2),(5,6),(3,4)}$ = 0,799233.

Таким образом, на 5-м шаге получаем единый кластер, в который вошли все объекты. На рисунке показана диаграмма этапов объединения объектов с указанием расстояния между кластерами.

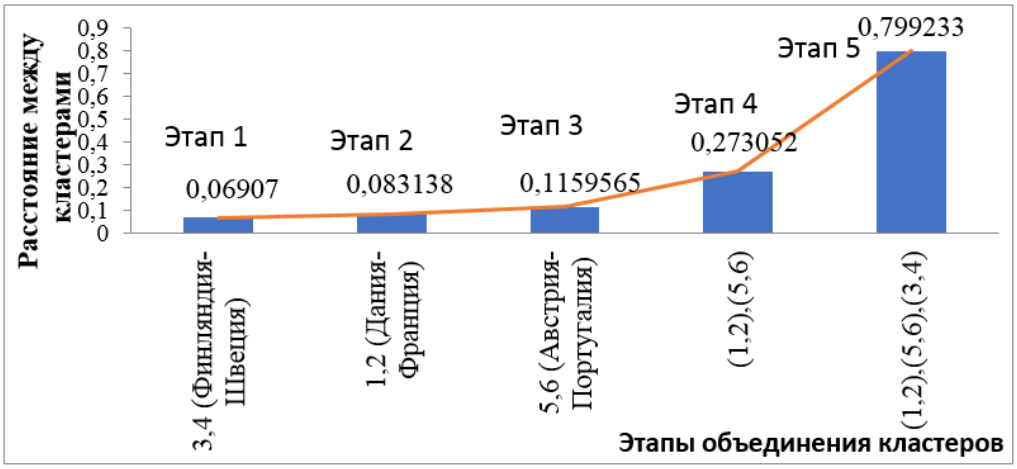

Рис. 3. Диаграмма этапов объединения объектов
и расстояние между кластерами

Сопоставим полученные результаты рассмотренного примера с аналогичными данными в SPSS. В качестве результатов иерархической агломерации в SPSS строится таблица с указанием порядка агломерации кластеров. Данные приведенной таблицы полностью соответствуют этапам рассмотренного примера.

Таблица 7

Порядок объединения кластеров по результатам анализа в SPSS

Порядок агломерации (*кластеров*)						
Этап	Объединенный кластер		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	3	4	,069	0	0	5
2	1	2	,083	0	0	4
3	5	6	,116	0	0	4
4	1	5	,273	2	3	5
5	1	3	,799	4	1	0

В данной таблице приводится следующая информация (по столбцам):

1 – этапы присоединения кластеров;

2 – первый объединяемый кластер;

3 – второй объединяемый кластер;

4 – коэффициенты, которые показывают расстояние между объединяемыми кластерами;

5 – этап объединения, на котором появляется впервые Кластер 1 (из столбца 2);

6 – этап объединения, на котором появляется впервые Кластер 2 (из столбца 3);

7 – следующий этап объединения, на котором появиться вновь один из объединяемых кластеров (из столбцов 2 или 3).

Как правило, по результатам иерархического кластерного анализа строится дендограмма процесса кластеризации в виде дерева иерархической структуры, каждый из уровней которой соответствует одному из шагов процесса последовательного объединения кластеров. Она показывает объединенные кластеры и расстояния между ними. Дендограммы строятся как вертикальные, так и горизонтальные. На одной оси дендограммы расположены названия или номера кластеров, а на другой оси – расстояние между кластерами при их соединении. По дендограмме можно узнать количество образованных кластеров на любом этапе, проведя прямую линию, перпендикулярную ветвям дендограммы: количество кластеров равно количеству точек пересечения прямой линии и ветвей дендограммы, а состав кластеров соответствует номерам объектов у корней. На рис. 4 показана дендограмма рассматриваемого примера, построенная с помощью SPSS. Расстояние между кластерами перешкалировано для наглядности в диапазоне чисел от 0 до 25.

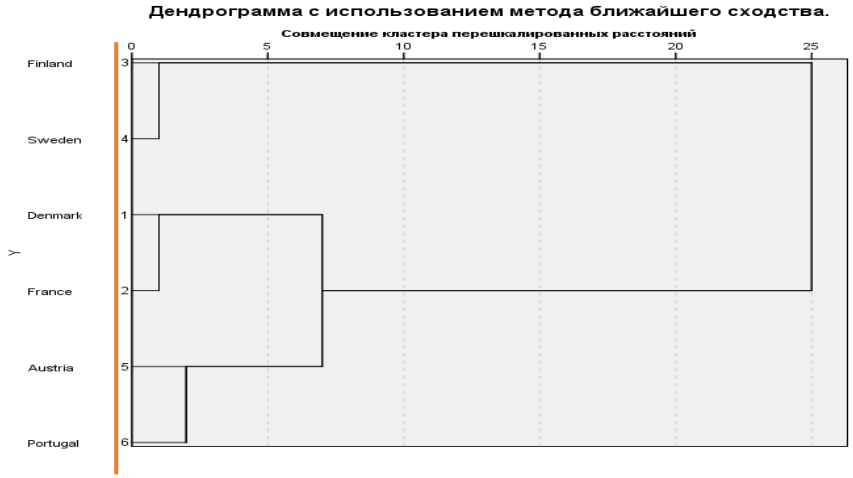

Рис. 4. Дендограмма шагов объединения объектов
(красной линией указано пересечение дендограммы по первому существенному скачку расстояния между объединяемыми кластерами $d_{(1,2),(5,6)}$)

Достаточно сложным вопросом при кластеризации является определение оптимального числа кластеров разбиения объектов. Методы кластерного анализа не имеют автоматической (самостоятельной) процедуры установления оптимального числа кластеров, но предоставляют аналитическую информацию для этого процесса. При объединении объектов в кластеры необходимо учитывать: чем меньше количество кластеров, тем больше объектов в них входит, а с увеличением числа объектов в кластере растет их гетерогенность. Поэтому оптимальным считается такое количество кластеров, при котором в них включается как можно больше объектов при наименьшей гетерогенности внутри кластера.

В иерархической кластеризации, как правило, число кластеров для разбиения определяют по динамике значения расстояния последовательно объединяемых объектов, что указано в таблице порядка агломерации (значение коэффициента объединения) или визуально на дендограмме. По таблице порядка агломерации анализируем динамику изменений значений коэффициента (расстояние между объединяемыми кластерами) и определяем, на каком шаге процесса объединения кластеров происходит первый относительно большой скачок его показателя. Данное значение коэффициента можно считать пороговым значением, до которого объединялись кластеры, находящиеся на достаточно малых расстояниях друг от друга, а, начиная с данного шага, происходит объединение более далеких кластеров. Оптимальное количество кластеров определяется как разница между общим числом объектов и номером шага объединения, предшествующего шагу с пороговым значением коэффициента. В приведенном выше примере на четвертом шаге объединения значение коэффициента 0,273, а на предыдущем шаге было значение коэффициента 0,116, рост составил 0,157 единиц, поэтому значение 0,273 можно считать пороговым значением, поскольку рост коэффициента на предыдущих шагах был значительно ниже 0,1. Таким образом, в

приведенном примере оптимальным можно считать следующее количество кластеров: 6 – 3 = 3. Можно считать количество кластеров и следующим образом: если скачок расстояния на 4-м шаге, то количество кластеров равно количеству шагов после шага со скачком расстояния плюс 2. В нашем примере после шага со скачком расстояния остался один шаг (5-й), поэтому оптимальное количество кластеров составит: 1 + 2 = 3.

Если используем дендограмму для определения оптимального числа кластеров, то визуально можем определить первый существенный рост расстояния объединений кластеров $d_{(1,2),(5,6)}$, что является пороговым значением. При этом, если условно «разрежем» дендограмму прямой линией через данное пороговое значение, то количество пересечений прямой линии с ветвями дендограммы и будет определять оптимальное число кластеров. В нашем примере имеем 3 пересечения прямой с ветвями дендограммы, значит, выделяем 3 кластера.

В некоторых случаях для определения оптимального количества кластеров используется критерий «*Ellbow*», который оценивается по показателю гетерогенности кластеров. Данный метод аналогичен использованию метода «каменистой осыпи» в факторном анализе. Строится график зависимости количества кластеров и значений их гетерогенности. График аналогичен графику на рис.3. Анализ данных графика показывает, что скачек гетерогенности кластеров происходит при переходе от шага 3 к 4, т.е. при сокращении числа кластеров с 3 до 2. Поэтому оптимальное число, при котором будут получены сравнительно однородные кластеры, составляет 3.

Необходимо отметить, что нет единого универсального метода определения оптимального числа кластеров, в некоторых случаях оно может определяться исследователем априорно по значениям некоторых показателей кластеров, их однородности, степени удаленности друг от друга, по данным внутригрупповой дисперсии или вариации и пр. При этом должно соблюдаться условие статистической значимости размера кластеров и их практической приемлемости. В приведенном выше примере, по анализу точечной диаграммы (см. рис. 2) можно априорно указать в качестве оптимального числа кластеров 3, что было

подтверждено в процессе кластеризации. Однако такие простые случаи кластеризации не встречаются, чаще приходится иметь дело с большим количеством объектов, классифицируемых по большому количеству признаков. В данном случае возможно применение следующего процесса определения оптимального количества кластеров:

1. Устанавливаем первичное значение числа кластеров стандартно по таблице агломераций в соответствии с существенным скачком коэффициента объединения кластеров.

2. Проводим кластеризацию по полученному количеству кластеров. Информацию по распределению объектов по кластерам можно сохранить как новую переменную. Создадим в исходном файле данных новую переменную, распределяющую все объекты по кластерам, и построим ее линейное распределение. Проанализируем полученные кластеры и определим их значимость по количеству объектов входящих в них. Если все кластеры значимы, то переходим к анализу кластеров.

3. Если есть кластеры не значимые, то их можно сократить. Учитывая общее число исследуемых объектов и необходимость статистической значимости кластеров можно априорно установить критическое значение по кластерам для данного процесса кластеризации.

4. Определяем, сколько кластеров состоят из значимого количества объектов и проводим кластеризацию по данному числу кластеров. Проводим кластеризацию с новым числом кластеров и создаем новую переменную. Построим линейное распределение новой переменной, определяем число кластеров со значимым числом объектов. Процедура данного пункта повторяется до тех пор, пока все кластеры будут состоять из значимого числа объектов. Полученное число кластеров можно считать оптимальным.

После определения оптимального количества кластеров, проводим повторную кластеризацию и получаем окончательные результаты разбиения объектов на кластеры. Анализ полученных результатов проводим по таблице кластерных профилей, которая показывает средние значения признаков для каждого

кластера. Процесс анализа состоит в сравнение средних значений признаков по кластерам, в установлении свойств объектов, характерных для полученных кластеров. Таким образом, проводим идентификацию и составляем характеристики полученных кластеров.

Общие выводы и рекомендации по методам кластеризации:

– методы кластеризации имеют большую эвристическую составляющую и во многом зависят от профессионализма пользователя или эксперта;

– результаты кластеризации существенно зависят от выбранной метрики, поэтому необходимо тщательно проводить предварительный анализ исходных данных и учитывать рекомендации по выбору различных мер сходства;

– многие алгоритмы требуют задания начальных условий, что также связано с субъективными решениями и учетом рекомендаций;

– поскольку в методах кластеризации формирование кластеров проводится за счёт формализованного подхода на основе мер сходства, то обоснование установленного числа кластеров в соответствии с некоторым субъективным критерием не всегда логично. Выбор количества кластеров является субъективной процедурой;

– поскольку нет однозначно наилучшего критерия качества кластеризации, то очень сложно логически оценить полученную кластеризацию. Многие критерии качества кластеризации могут показывать разные результаты. Поэтому для оценки качества и анализа полученных результатов рекомендуется привлекать экспертов.

Список литературы

1. Айвазян С.А. Классификация многомерных наблюдений / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М.: Статистика, 1974.

2. Айвазян С.А. О структуре и содержании пакета программ по прикладному статистическому анализу / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин // Алгоритмическое и программное обеспечение прикладного статистического анализа. – М., 1980.

3. Беккер В.А. Об анализе структуры матрицы коэффициентов связи / В.А. Беккер, М.Л. Лукацкая // Вопросы экономико-статистического моделирования и прогнозирования в промышленности. – Новосибирск, 1970.

4. Браверман Э.М. Структурные методы обработки данных / Э.М. Браверман, И.Б. Мучник. – М.: Наука, 1983.

5. Воронин Ю.А. Теория классифицирования и ее приложения / Ю.А. Воронин. – Новосибирск: Наука, 1987.

6. Дубровский С.А. Прикладной многомерный статистический анализ / С.А. Дубровский. – М.: Финансы и статистика, 1982.

7. Дюран Н. Кластерный анализ / Н. Дюран, П. Оделл. – М.: Статистика, 1977.

8. Дюран Б. Кластерный анализ / Б. Дюран. – М.: Книга по Требованию, 2012. – 128 с.

9. Елисеева И.И. Группировка, корреляция, распознавание образов / И.И. Елисеева, В.С. Рукавишников. – М.: Статистика, 1977.

10. Романовский В.И. Избранные труды. – Т. 2. Теория вероятностей, статистика и анализ / В.И. Романовский. – М., 1980. – 973 с.

11. Судаков С.А. Кластерный анализ в психиатрии и клинической психологии (+ CD-ROM) / С.А. Судаков. – М.: Медицинское информационное агентство, 2010. – 164 с.

12. Фостер Дж. Автоматический синтаксический анализ / Дж. Фостер. – М., 1975. – 445 с.

13. Фурман Я.А. Введение в контурный анализ и его приложения к обработке изображений и сигналов / Я.А. Фурман. – М., 2003. – 353 с.

Сизых Дмитрий Сергеевич – канд. техн. наук, доцент департамента финансового менеджмента ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики», Россия, Москва.

Сизых Наталья Васильевна – канд. техн. наук, доцент департамента математики ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики», Россия, Москва.