УДК 378.147
DOI: 10.35750/2071-8284-2020-1-208-217

А.Н. Литвиненко
доктор экономических наук, профессор
Санкт-Петербургский университет МВД России
Российская Федерация, 198206, Санкт-Петербург, ул. Летчика Пилютова, д. 1
ORCID: 0000-0003-0599-4027. E-mail: Lanfk@mail.ru

Л.В. Большакова
кандидат физико-математических наук, доцент
Санкт-Петербургский университет МВД России
Российская Федерация, 198206, Санкт-Петербург, ул. Летчика Пилютова, д. 1
ORCID: 0000-0002-5176-8832. E-mail: blv5505@mail.ru

Методика применения кластерного анализа при выполнении выпускных квалификационных работ слушателями Санкт-Петербургского университета МВД России

Аннотация: Актуальность темы определяется необходимостью обеспечения качества выпускных квалификационных работ слушателями университета. Данная цель достигается в том числе и с применением методов математико-статистического анализа. В статье рассматривается вопрос практического применения методики кластерного анализа слушателями университета при выполнении выпускной квалификационной работы. Авторами проанализированы достоинства и недостатки данного метода и определены границы его использования. С методической точки зрения представляет интерес обоснование границ применения метода кластерного анализа. Обращается внимание на постановку частных задач, решение которых предполагает достижение цели кластерного анализа. Авторы описывают этапы кластеризации, выделенные при использовании восходящего иерархического метода. В работе описаны основные правила и последовательность применения пакета «Statistica» с англоязычной версией для решения конкретной задачи кластеризации большого числа объектов. В качестве примера приведена задача кластеризации одиннадцати субъектов Северо-Западного федерального округа на предмет выявления схожих черт их инновационного развития. Авторы подчёркивают важность использования метода с точки зрения выявления общих свойств объектов в выделяемых кластерах. Содержательными являются и закономерности, которыми описываются взаимоотношения отдельных групп объектов.

Ключевые слова: группировка, кластерный анализ, метод анализа, параметры, свойств объектов, этапы кластеризации.

Для цитирования: Литвиненко А.Н., Большакова Л.В. Методика применения кластерного анализа при выполнении выпускных квалификационных работ слушателями Санкт-Петербургского университета МВД России // Вестник Санкт-Петербургского университета МВД России. – 2020. – № 1 (85). – С. 208–217. DOI: 10.35750/2071-8284-2020-1-208-217.

Alexander N. Litvinenko
Dr. Sci. (Econ.), Professor
Saint-Petersburg University of the MIA of Russia
1, Letchika Pilyutova str., Saint-Petersburg, 198206, Russian Federation
ORCID: 0000-0003-0599-4027. E-mail: Lanfk@mail.ru

Lyudmila V. Bolshakova
Cand. Sci. (Phys.-Math.), Docent
Saint-Petersburg University of the MIA of Russia
1, Letchika Pilyutova str., Saint-Petersburg, 198206, Russian Federation
ORCID: 0000-0002-5176-8832. E-mail: blv5505@mail.ru

The technique of applying cluster analysis when Performing graduate qualification work by students St. Petersburg University of the Ministry of Internal Affairs of Russia

Annotation: The relevance of the topic have determined by needs of ensuring the quality of graduation qualifications by university students. This aim have to achieved, inter alia, using methods of mathematical and statistical analysis. The article discusses the practical application of the cluster analysis methodology by university students in the Graduation qualification work. The authors have analysed advantages and disadvantages of this method and determined the boundaries of its use. From a methodological point of view, it have interested to justify the boundaries of the Cluster analysis method application. Attention is to drawn the particular problems formulation the solution of which involves achieving the aim of cluster analysis. The authors describe the stages of clustering have identified with using the ascending hierarchical method. The article describes the basic rules and the sequence of using the STATISTICA package with the English version to solve the specific problem of clustering a large number of objects. As an example, the task of clustering eleven subjects of the Northwestern Subject of Russia in order to identify similar features of their innovative development. The authors emphasize the importance of using the method in terms of identifying the general properties of objects in allocated clusters. The regularities that describe the relationship of individual groups of objects are also substantial.

Keywords: grouping, cluster analysis, analysis method, parameters, properties of objects, clustering stages.

For citation: Litvinenko A.N., Bolshakova L.V. The methodology of applying cluster analysis when performing graduate qualification work by students of the St. Petersburg University of the Ministry of Internal Affairs of Russia // Vestnik of St. Petersburg University of the Ministry of Internal Affairs of Russia. – 2020. – № 1 (85). – P. 208–217. DOI: 10.35750/2071-8284-2020-1-208-217.

Введение

Различные процессы и явления, представляющие предмет научных исследований по разным направлениям, в том числе педагогическим, психологическим, экономическим, достаточно часто зависят от большого числа параметров-факторов, их характеризующих. В связи с этим применение в таких исследованиях, в том числе в выпускных квалификационных работах, математико-статистического анализа для получения определённых выводов, а также для подтверждения или опровержения конкретных результатов, значительно увеличивает научную ценность и достоверность этих работ.

К сожалению, использование этих методов в научных работах, особенно гуманитарного направления, крайне редко и не всегда правильно. Одной из причин является достаточно частое отсутствие понятного и методически верного изложения этих методов в процессе обучения.

Целью данной статьи является рассмотрение возможностей применения многомерного математико-статистического анализа – кластерного анализа – при выполнении выпускных квалификационных работ, а также при проведении слушателями научных исследований. В Санкт-Петербургском университете кластерный анализ рассматривается при изучении нескольких учебных дисциплин, в том числе «Математические методы в психолого-педагогических исследованиях», «Эконометрика», «Математические основы обработки информации» при подготовке специалистов разного профиля. Владение методами математической статистики необходимо как для многих специалистов, так и при проведении научных исследований, в том числе при написании выпускных квалификационных работ разного уровня сложности, где приходится иметь дело с обработкой эмпирических данных.

Для правильного применения любого метода, для получения результатов, адекватных действительности, обучающийся должен чётко представлять себе следующие моменты: основные задачи, решаемые данным методом, его сфера применения, достоинства и недостатки; содержание метода, его основные этапы; возможность использования при применении метода компьютерных программ, стандартных пакетов и т.д. Всё это должно стать основой для преподавателя, который знакомит обучающихся с конкретным методом.

Кластерный анализ, его суть и основные этапы

Изучение методов кластерного анализа предполагает изложение теоретического материала на лекции и выполнение конкретных заданий в компьютерном классе. Материал лекции включает в себя рассмотрение вопросов, связанных с понятием группировки объектов или явлений, необходимостью и сферой применения кластерного анализа, рассмотрением его основных этапов. На практических занятиях в компьютерных классах должно быть изучено применение данного метода при решении конкретных задач, которые в дальнейшем могут по-

мочь обучающимся при написании выпускной работы.

Основное содержание лекции целесообразно изложить следующим образом. При исследовании различных процессов и явлений нередко возникает задача группировки рассматриваемых объектов по значениям характеризующих их признаков. Разделение на группы (группировку) можно проводить различными способами в зависимости от постановки задачи. Это может быть группировка: по значению самого важного (информативного) признака; по нескольким ранжированным признакам; по некоторому обобщённому показателю [1, с. 7–12; 137–143].

Группировка по нескольким признакам одновременно лучше всего, по мнению авторов, может быть проведена с помощью методов кластерного анализа.

Преимуществами методов кластерного анализа являются следующие: они не требуют дополнительной априорной информации; число групп-классов может определяться и изменяться в результате исследования; у них не существует ограничений, связанных с распределением исходных переменных, поэтому они могут быть применены как к количественным, так и к качественным переменным; ввиду необязательности распределения исходных переменных по нормальному закону они могут успешно применяться в социальных науках.

К недостаткам методов кластерного анализа можно отнести в первую очередь их громоздкость. Однако с появлением пакетов прикладной статистики этот недостаток становится несущественным. Кроме этого, выбор меры «схожести» или меры «различия» объектов и групп объектов в соответствии со значениями их признаков, выбор критерия качества кластеризации и оптимального количества кластеров носит часто субъективный характер.

Для применения методов кластерного анализа необходимо чтобы, во-первых, выбранные объекты в принципе (по смыслу) допускали желательное разбиение на кластеры и, во-вторых, единицы измерения, т.е. масштаб, были выбраны правильно. Второе условие является важным для получения правильных результатов, и для его выполнения с каждым признаком производят процесс нормировки.

Сфера применения кластерного анализа достаточно широка и многообразна ввиду универсальности его методов [2–6][1]. Так, в маркетинге решается задача сегментации рынка по группам товаров; в социологии – задача разбиения респондентов на однородные группы, в юриспруденции – задача распределения по видам преступлений; в экономике – при распределении банков на группы в зависимости от их кредитоспособности или предприятий в зависимости от их платежеспособности и т. д.

Главная цель кластерного анализа – это исследование структуры выборочной совокупности, которая состоит из определённого числа объектов. Каждый объект характеризуется конкретным набором признаков-факторов, значения которых считаются заданными. В зависимости от конкретной постановки проблемы достижение цели кластерного анализа может быть связано с решением следующих задач.

1. Исследование однородности совокупности представленных объектов. Если совокупность является однородной по некоторому критерию, то для дальнейшего статистического анализа могут успешно применяться методы многомерного исследования, например, методы корреляционно-регрессионного анализа. Если совокупность является неоднородной, то применение этих методов может привести к результатам, далёким от реального положения дел. Поэтому проверка однородности имеет исключительно важное значение.

2. Разделение совокупности объектов на однородные в определённом смысле группы. Эта задача непосредственно вытекает из первой тогда, когда выявлена неоднородность совокупности. В этом случае кластерный анализ позволяет разделить её на некоторое число однородных групп (кластеров) с учётом значений признаков, характеризующих объекты.

3. Сжатие информации (данных). Если совокупность состоит из очень большого числа объектов, то после объединения объектов в группы для дальнейшего исследования можно либо рассматривать наиболее типичных представителей от каждой группы, либо получить от каждой группы представителя с обобщенными показателями. И в том, и в другом случае происходит существенное сокращение информации.

4. Обнаружение аномалий. В процессе кластеризации могут быть выделены некоторые нетипичные объекты, которые явно выделяются из всей совокупности. Например, следует особое внимание обратить на кластеры, состоящие из небольшого числа объектов (1–3 объекта).

По принципу образования групп методы кластерного анализа можно разделить на иерархические, использующие алгоритмы последовательной группировки (кластеризации) объектов, и неиерархические методы, основанные на определении концентрации (сгущения) объектов, поэтому их иногда называют центрографическими (метод К-средних). И, наконец, существует метод двухвходового объединения, при котором кластеризация происходит одновременно по объектам и признакам. Этот метод применяется только в тех случаях, когда есть уверенность в том, что одновременная кластеризация даст кластеры, имеющие смысловое значение.

Иерархические методы делятся на агломеративные (восходящие) и дивизимные (нисходящие). Первые предполагают, что на первом

[1] См. также: Большакова Л. В., Примакин А. И., Яковлева Н. А. Применение кластерного и дискриминантного анализов в процессе обработки и интерпретации статистических данных при обеспечении экономической и информационной безопасности хозяйствующего субъекта // Вестник Санкт-Петербургского университета МВД России. – 2014. – № 2 (62). – С. 148–156.

шаге все объекты по отдельности образуют кластеры: сколько объектов, столько и кластеров. Затем происходит пошаговое объединение объектов в укрупнённые кластеры, начиная с объектов, наиболее близких по определённому критерию. Объединение происходит до тех пор, пока не будет составлен один кластер. Каждый шаг данного объединения описывается аналитически и графически в виде дендрограммы (дерева), по которой можно определить сколько кластеров необходимо оставить для дальнейшего исследования. Дивизимные методы являются логической противоположностью агломеративным. Для них вначале процедуры все объекты образую один кластер, а затем на каждом шаге происходит «расслоение» на более мелкие кластеры [7, с. 78–85].

Рассмотрим более подробно основные этапы кластеризации при использовании восходящего иерархического метода.

Первый этап: постановка задачи

Пусть имеется n объектов C_1, …, C_n, которые характеризуются m количественными признаками (переменными) V_1, … V_m. Таким образом, для каждого объекта C_i должно быть известно значение каждого признака V_j, где $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. Очевидно, что основная трудность, и в связи с этим необходимость применения кластерного анализа, возникает тогда, когда число признаков, характеризующих каждый объект, становится больше двух.

После постановки задачи необходимо оценить значения признаков. Если значения хотя бы одного признака существенно отличаются по величине от значений других, рекомендуется провести нормировку признаков.

Второй этап: нормировка признаков

Нормировка (стандартизация) признака предполагает переход от обычной случайной величины (признака) к стандартной, т.е. для каждого признака находят среднее μ и среднеквадратическое отклонение σ его значений. Новые значения признака определяются по следующему правилу: от старого значения вычитается μ, затем разность делится на σ.

После построения таблицы стандартизированных данных начинают процесс кластеризации с выбора функции расстояний между объектами.

Третий этап: критерий близости объектов

Нами рассматривается восходящий иерархический метод. Таким образом, имеется n кластеров, т.е. каждый объект считается отдельным кластером. На данном этапе определяются наиболее близкие, с точки зрения определённого критерия, объекты, т.е. те объекты, которые можно объединить в один кластер. Критерием близости объектов является расстояние между объектами, которое зависит от значений признаков, характеризующих объекты. В кластерном анализе используются следующие основные расстояния между объектами.

1. Евклидово расстояние между двумя объектами определяется как квадратный корень из суммы квадратов разностей соответствующих значений признаков, характеризующих эти объекты. Эта функция расстояния является наиболее простой и наиболее часто применяемой. При наличии лишь двух или трёх признаков – это длина отрезка, соединяющего точки (объекты).

2. Квадрат евклидова расстояния между двумя объектами определяется как сумма квадратов разностей соответствующих значений признаков, характеризующих эти объекты. Эта функция расстояния похожа на предыдущую, однако её применение связано с необходимостью придания большего веса объектам, наиболее удалённым друг от друга, так как в расчётах при возведении в квадрат большие разности вносят и больший вклад. Формула наиболее часто используется, если в дальнейшем (четвёртый этап) будет применяться центроидный метод или метод Варда.

3. Расстояние городских кварталов (манхэттенское расстояние) между двумя объектами определяется как сумма модулей разностей соответствующих значений признаков, характеризующих эти объекты. Эта функция расстояния в большинстве случаев дает такие же результаты, как и при применении формулы евклидова расстояния. Однако необходимо отметить, что влияние отдельных больших разностей (выбросов) для этой функции существенно уменьшается, так как эти разности в квадрат не возводятся.

4. Расстояние Чебышева между двумя объектами определяется как максимальное значение из всех модулей разности между соответствующими значениями признаков, характеризующих эти объекты. Применение данной формулы является достаточно редким. Чаще всего эту формулу используют в том случае, когда два объекта необходимо считать различными, если они различаются по какой-либо одной координате (одному значению).

Для нахождения расстояния между объектами существуют ещё несколько формул, имеющих специальное применение. Например, степенное расстояние, которое представляет в основном математический интерес как универсальная функция. Или формула расстояния, основанная на проценте несогласия, которая применяется в том случае, когда данные являются категориальными.

После выбора формулы расстояния между объектами и применения её для всех пар объектов получают матрицу расстояний между объектами. Далее выбирают из этой матрицы элемент(ы) с наименьшим значением и объединяют соответствующие объекты в кластер. В результате получают новую матрицу расстояний, размер которой будет меньше размера предыдущей матрицы. Для дальнейшего объединения объектов в кластеры необходимо выбрать критерий близости между кластер-группами.

Четвёртый этап: критерий близости кластер-групп

Этап состоит из нескольких шагов, на каждом из которых происходит объединение наиболее близких друг к другу кластеров. При этом каждый из объединяемых кластеров может состоять либо из одного, либо из нескольких объектов. Процесс продолжается до тех пор, пока все объекты не попадут в один кластер.

Критерий близости между кластерами – расстояние между этими кластер-группами. Это расстояние может быть определено по следующим основным правилам (принципам).

1. Правило ближайших соседей (метод одиночной связи) определяет расстояние между кластерами как расстояние между двумя объектами из разных кластеров, которые расположены наиболее близко друг к другу.

2. Правило дальних соседей (метод полной связи) можно рассматривать как альтернативу предыдущему правилу. По нему за расстояние между кластерами принимается расстояние между двумя объектами из разных кластеров, которые расположены наиболее удалённо друг от друга.

3. Правило средней связи (метод невзвешенного попарного среднего) – расстояние между кластерами определяет среднее расстояние между всеми парами объектов из разных кластеров.

4. Правило средневзвешенной связи (метод взвешенного попарного среднего) отличается от предыдущего тем, что в данном методе учитывается весовой коэффициент, который определяется размером соответствующих кластеров (числом объектов, содержащихся в кластере).

5. Правило центра тяжести (невзвешенный центроидный метод) определяет расстояние между кластерами как расстояние между центрами тяжести кластеров.

6. Медианное правило (взвешенный центроидный метод) идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров. Данный метод является более предпочтительным, чем предыдущий, если есть основания предполагать, что имеются существенные отличия в размерах кластеров.

7. Правило Варда (метод Варда) использует для определения расстояния между кластерами методы дисперсионного анализа, что существенно отличает его от всех вышеописанных методов. Метод достаточно эффективен и направлен на объединение близко расположенных кластеров.

В результате применения какого-либо из перечисленных правил получают новую матрицу расстояний, по которой происходит дальнейшее соединение объектов и кластеров в новые кластер-группы до образования одного кластера.

Пятый этап: представление результатов

На данном этапе происходит представление полученных результатов кластеризации в виде таблиц и графически в виде дендрограммы. Далее сам исследователь решает, какое число кластеров его устраивает. Особое значение для ответа на этот вопрос оказывает величина, характеризующая расстояние между кластерами, отмеченная на дендрограмме.

Кластеризацию большого числа объектов производят, как правило, с использованием статистических пакетов, например, «Statistica», SPSS, «Stadia», «StatGraphics». Изучение работы с этими пакетами должно происходить на практических занятиях в компьютерных классах с пошаговым решением конкретного примера.

Опишем основные моменты применения наиболее доступного пакета «Statistica» с англоязычной версией для решения следующей задачи.

Примером использования данного метода при написании выпускной квалификационной работы может служить исследование, в котором решается задача кластеризации одиннадцати субъектов СЗФО на предмет выявления схожих черт их инновационного развития. В качестве субъектов были выбраны: С1 – Карелия; С2 – Коми; С3 – Архангельская область; С4 – Ненецкий автономный округ; С5 – Вологодская область; С6 – Калининградская область; С7 – Ленинградская область; С8 – Мурманская область; С9 – Новгородская область; С10 – Псковская область; С11 – Санкт-Петербург [8].

Анализировались следующие индикаторы (признаки) инновационной активности регионов: V1 – внутренние затраты на научные исследования и разработки (млн. руб.); V2 – количество выданных патентов на изобретения в 2017 году (штук); V3 – количество разработанных передовых производственных технологий (штук); V4 – количество используемых передовых производственных технологий (штук); V5 – инновационная активность организаций в субъекте (удельный вес организаций, осуществлявших технологические, маркетинговые и организационные инновации, в общем числе обследованных организаций; в процентах); V6 – объём произведенных инновационных товаров, работ, услуг (в процентах от общего объёма отгруженных товаров, выполненных работ, услуг); V7 – количество организаций, осуществляющих инновации, обеспечивающие повышение безопасности в процессе производства товаров, работ, услуг (в процентах от общего числа организаций, осуществляющих экологические инновации соответствующего субъекта Российской Федерации); V8 – затраты на технологические инновации (в процентах от общего объёма отгруженных товаров, выполненных работ, услуг).

Значения признаков для каждого субъекта получены на основе данных Федеральной службы государственной статистики за 2018 год (табл. 1).

Введение исходных данных

В окне «Welcome to “Statistica”» («Приглашение в “Statistica”») выбирается способ введения данных. Для этого во вкладке «Home» («Главная») выбираем команду «New» («Создать»). В выпадающем меню выбираем элемент

Таблица 1

Инновационные показатели по 11 субъектам СЗФО

признаки / субъекты	V1	V2	V3	V4	V5	V6	V7	V8
C1	943,2	27	10	660	5,9	0,3	25,0	0,4
C2	2350,0	40	1	910	3,5	0,4	20,0	0,4
C3	1522,9	56	9	1417	4,4	28,4	0	0
C4	21,5	0	0	63	4,6	0	0	0,5
C5	479,5	64	8	2992	5,4	2,9	100	0,2
C6	1094,0	46	1	859	4,3	0,3	25,0	0,3
C7	6863,5	42	18	1879	9,3	2,2	44,4	4,1
C8	2276,1	35	0	1145	8,2	1,3	12,5	0,4
C9	2751,8	44	28	1983	8,8	4,0	50,0	0,8
C10	437,7	30	1	1363	7,4	2,1	0	0,5
C11	120804,0	1541	130	8933	16,1	9,1	34,5	2,8

«Spreadsheet» («Таблица») и нажимаем «ОК». На экране появляется пустая электронная таблица, которая содержит 10 объектов (строки) и 10 признаков (столбцы).

В примере – 11 объектов, следовательно, необходимо добавить еще одну строку (объект) в таблицу. Во вкладке «Data» («Данные») выбираем «Cases» («Наблюдения») и в выпадающем меню выбираем команду «Add» («Добавить»). В поле «How many» («Сколько») вводим: 1. В поле «Insert after case» («Вставить после») вводим: 10. Если число объектов меньше десяти, то выбираем команду «Delete» («Удалить»).

Для уменьшения числа признаков с 10 до 8 во вкладке «Data» («Данные») выбираем «Variables» («Переменные»), затем в выпадающем меню команду «Delete» («Удалить»). В окне «Delete Variables» в поле «From variable» («От переменной») вводим: 9, а в поле «To variable» («К переменной»): 10 и нажимаем «ОК». Таблица необходимой размерности получена. Копируем данные в эту таблицу, получаем (рис.1):

Проведение процесса стандартизации

Данные таблицы существенно отличаются по величине, следовательно, необходимо провести процесс стандартизации. Для этого выделяем таблицу с данными, на вкладке «Data» («Данные») выбираем команду «Standardize» («Стандартизация»). В двух полях, связанных с объектами и переменными, отмечаем «All» («Все»), а в поле «Weight» («Вес») – «Off» («Не задавать»). После нажатия кнопки «ОК» получаем стандартизованные данные для дальнейшей работы.

Выполнение процесса кластеризации

На вкладке «Statistics» («Анализ») выбираем вначале команду «Mult / Exploratory» («Многомерный анализ»), а затем «Cluster» («Кластерный анализ»). В появившемся окне «Clustering Method» («Методы кластеризации») представлены три метода:

– «Joining tree clustering» («Иерархическая классификация»),

Data: Spreadsheet1* (8v by 11c)

	1 Var3	2 Var4	3 Var5	4 Var6	5 Var7	6 Var8	7 Var9	8 Var10
1	943,2	27	10	660	5,9	0,3	25	0,4
2	2350	40	1	910	3,5	0,4	20	0,4
3	1522,9	56	9	1417	4,4	28,4	0	0
4	21,5	0	0	63	4,6	0	0	0,5
5	479,5	64	8	2992	5,4	2,9	100	0,2
6	1094	46	1	859	4,3	0,3	25	0,3
7	6863,5	42	18	1879	9,3	2,2	44,4	4,1
8	2276,1	35	0	1145	8,2	1,3	12,5	0,4
9	2751,8	44	28	1983	8,8	4	50	0,8
10	437,7	30	1	1363	7,4	2,1	0	0,5
11	120804	1541	130	8933	16,1	9,1	34,5	2,8

Рис. 1. Исходные данные для кластеризации

Рис.2. Иерархический кластерный анализ

– «K-means clustering» («Кластеризация методом К-средних»),

– «Two-way joining» («Метод двухвходового объединения»).

Выбираем иерархическую классификацию, нажимаем «ОК». Получаем следующую картинку (рис. 2).

Далее на вкладке «Advanced» («Дополнительно») заполняем следующим образом поля: «Variables» («Переменные») – «All» («Все»), так как все переменные участвуют в кластеризации; «Input file» («Файл данных») – «Raw data» («Исходные данные»), так как данные у нас уже введены; «Cluster» («Объекты») – «Cases (Rows)» («Наблюдения (строки)»), так как проводится кластеризация объектов (строк).

Затем заполняем поля, связанные с выбором расстояния между объектами – «Distance measure» (третий этап кластерного анализа) и расстояния между кластерами – «Amalgamation (linkage) rule» (четвертый этап кластерного анализа).

Для «Distance measure» («Мера близости») в выпадающем меню имеются следующие функции, описанные ранее: «Squared Euclidean distances» («Квадрат евклидова расстояния»); «Euclidean distances» («Евклидово расстояние»); «City-block (Manhattan) («Расстояние городских кварталов»); «Chebyshev distance metric» («Расстояние Чебышева»); «Power: SUM (ABS(x–y) xxp)xx1/r» («Степенное расстояние»); «Percent disagreement» («Процент несогласия»). Выбираем «Euclidean distances».

Для «Amalgamation (linkage) rule» («Правило объединения») в выпадающем меню имеются следующие функции, описанные ранее: «Single Linkage» («Метод одиночной связи»); «Complete Linkage» («Метод полной связи»); «Unweighted pair-group average» («Метод невзвешенного попарного среднего»); «Weighted pair-group average» («Метод взвешенного попарного среднего»); «Unweighted pair-group centroid» («Невзвешенный центроидный метод»); «Weighted pair-group centroid» («Взвешенный центроидный или медианный метод»); «Ward's method» («Метод Варда»). Выбираем «Single Linkage». После нажатие кнопки «ОК» будет выполнена кластеризация.

Получение результатов кластеризации

Результаты кластеризации можно получить в окне «Joining Results» («Результаты иерархической классификации») при переходе на вкладку «Advanced» («Дополнительно») (рис. 3):

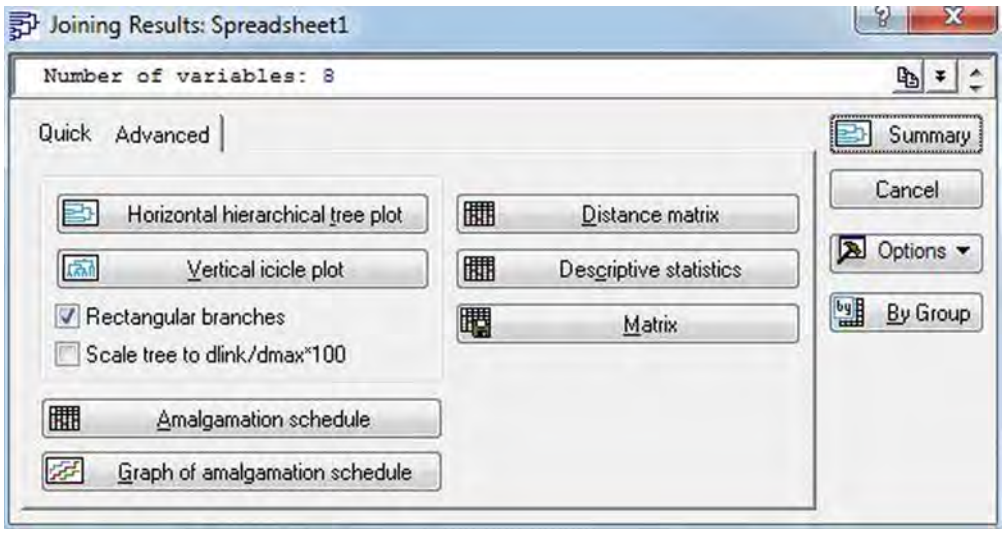

Рис. 3. Результаты иерархической кластеризации

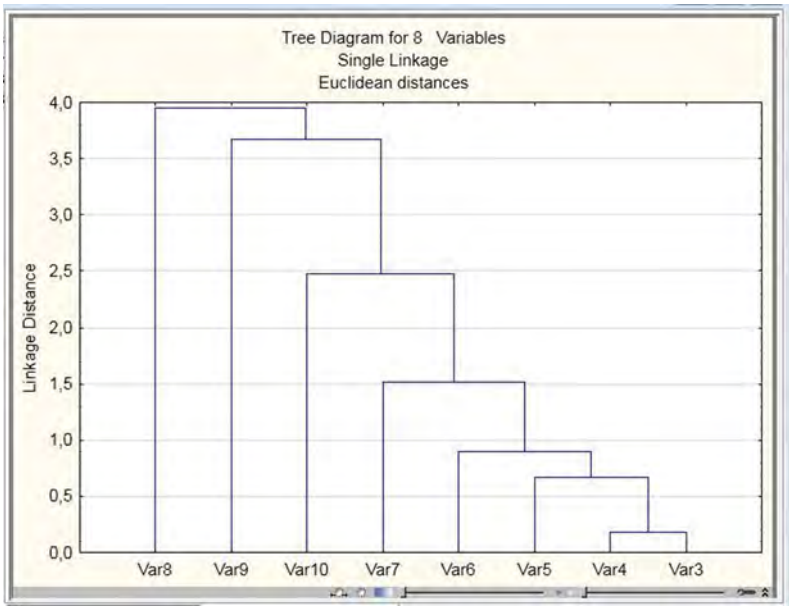

Рис. 4. Горизонтальная дендрограмма

Дендрограмму можно получить в горизонтальном или вертикальном виде при нажатии кнопки «Horizontal hierarchical tree plot» или «Vertical icicle plot» соответственно. В первом случае объекты будут располагаться на вертикальной оси, а на горизонтальной – расстояния объединения. Во втором случае – наоборот. Выберем второй случай, т.е. кнопку «Vertical icicle plot», получим следующий график (рис. 4):

По вертикальной оси отложены расстояния между кластерами. Ближе всех по сумме показателей объекты 2 и 6 на первом этапе они объединены в один кластер, остальные объекты не объединялись. На втором этапе с очень небольшим увеличением расстояния образовался кластер с двумя объектами: 8 и 10 и т.д. Дальше всех от всех объектов – объект 11.

Для получения таблицы, характеризующей схему объединения, нажимаем на кнопку «Amalgamation Schedule» («Схема объединения»), получаем (табл. 2):

Таблица 2

Схема объединения

Amalgamation Schedule Single Linkage Euclidean distances											
	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11
0,2943587	C_2	C_6									
0,5071447	C_8	C_10									
0,5193704	C_1	C_2	C_6								
0,8365631	C_1	C_2	C_6	C_4							
0,8470256	C_1	C_2	C_6	C_4	C_8	C_10					
1,482898	C_1	C_2	C_6	C_4	C_8	C_10	C_9				
2,116611	C_1	C_2	C_6	C_4	C_8	C_10	C_9	C_5			
2,600068	C_1	C_2	C_6	C_4	C_8	C_10	C_9	C_5	C_7		
3,308197	C_1	C_2	C_6	C_4	C_8	C_10	C_9	C_5	C_7	C_3	
6,605019	C_1	C_2	C_6	C_4	C_8	C_10	C_9	C_5	C_7	C_3	C_11

В первом столбце таблицы приведены расстояния до соответствующих кластеров. В каждой строке показан состав кластера на соответствующем шаге кластеризации. Например, на первом шаге (первая строка) объединяются второй и шестой объекты с расстоянием между ними равным 0,2943587. На втором шаге объединились в кластер 8 и 10 объекты (вторая строка) с расстоянием между ними равным 0,5071447 и т.д.

При необходимости можно получить первоначальную матрицу евклидовых расстояний, нажав на кнопку «Distances matrix» («Матрица расстояний»). При нажатии кнопки «Graph of amalgamation schedule» («График схемы объединения») можно посмотреть результаты древо-

видной кластеризации, т.е. диаграмму, по горизонтальной оси которой отложено число шагов, а по вертикальной – расстояния. В данном случае для объединения всех кластеров в один потребовалось 10 шагов.

И, наконец, для каждого объекта можно найти среднее значение («Means») по всем признакам и среднеквадратическое отклонение («Standard Deviations»), нажав кнопку «Descriptive statistics» («Описательная статистика»).

Кластеризация может быть проведена и с применением других методов, описанных на третьем и четвёртом этапах кластерного анализа. Для наших данных [8] можно получить, например, следующие результаты (табл. 3):

Таблица 3

Результаты кластеризации 11 субъектов СЗФО

Метод	Первый кластер	Второй кластер	Третий кластер	Четвертый кластер
Метод одиночной связи	9;7; 5; 11; 8; 10; 2; 6	3	4	1
Метод полной связи	9; 7; 5; 11; 8; 10; 4	6; 2	3	1
Метод невзвешенного попарного среднего	9; 7; 5; 11; 8; 10; 2; 6	3	4	1
Невзвешенный центроидный метод	9; 7; 5; 11; 8; 10; 2; 6	3	4	1
Метод Варда	9; 7; 5; 11; 8; 10	3; 2; 6	4	1

Результаты анализа полученных данных позволяют сделать ряд выводов:

– предсказуемым стало формирование в отдельный кластер г. Санкт-Петербурга как одного из «локомотивов инновационного развития» страны;

– неожиданно выглядит практически единогласное, с точки зрения различных методов, формирование третьего кластера с Архангельской областью. Основаниями, послужившими её обособленности в единый кластер, стали количество выданных патентов и объём произведённой инновационной продукции в 2017 году;

– обособление Ленинградской, Вологодской и Новгородской областей связано с большим процентом выданных патентов в субъектах при высоких показателях использования передовых технологий;

– согласно трём из шести методов выделена в отдельный кластер Ленинградская область. В данном случае основаниями для попадания в отдельный кластер стали значительные для округа внутренние затраты на НИР и наибольшие затраты на технологические инновации.

Заключение

После кластеризации возможны несколько путей научного исследования множества объектов.

Первый предполагает дальнейшее математико-статистическое исследование в каждой кластер-группе, а затем выявление каких-то общих свойств, закономерностей и т.п.

Второй путь связан со «сжатием» информации. Определяется новая совокупность, число объектов в которой равно числу кластеров, т.е. каждый кластер, состоящий из нескольких объектов, рассматривается как новый объект с новыми общими характеристиками или обобщёнными значениями показателей признаков, входящих в кластер.

Третий путь определяется возможностью связать дальнейшее исследование с методами дискриминантного анализа, с помощью которых можно, например, выяснить к какой кластер-группе отнести новый появившийся объект, а также сделать некоторый прогноз [9, с. 255–262].

Таким образом, представленная методика применения кластерного анализа может быть использована и при написании выпускных квалификационных работ, и при выполнении исследовательских работ слушателями любого гуманитарного направления.

Список литературы

1. *Мандель И. Д.* Кластерный анализ. – Москва: Финансы и статистика, 1988. – 176 с.

2. *Баринова В. А., Дробышевский С. М., Еремкин В. А., Земцов С. П., Сорокина А. В.* Типология регионов России для целей региональной политики // Российское предпринимательство. – 2015. – Т. 16. – № 23. – С. 4199–4204.

3. *Гордячкова О. В.* Кластерный анализ привлечения иностранных инвестиций российскими регионами // Российское предпринимательство. – 2013. – Т. 14. – № 3. – С. 116–121.

4. *Маслюкова Е. В., Зайцева Ю. Ю.* Занятость в неформальном секторе: количественные методы анализа // Экономика труда. – 2017. – Т. 4. – № 4. – С. 423–430.

5. *Савченко Т. Н.* Применение методов кластерного анализа для обработки данных психологических исследований // Экспериментальная психология. – 2010. – Т. 3. – № 2. – С. 67–86.

6. *Хайдуков Д. С.* Применение кластерного анализа в государственном управлении // Философия математики: актуальные проблемы : сборник тезисов II Международной научной конференции «Философия математики: актуальные проблемы», МГУ им. М.В. Ломоносова. – Москва: МАКС Пресс, 2009. – 287 с.

7. *Жамбю М.* Иерархический кластер-анализ и соответствия. – Москва: Финансы и статистика, 1988. – 345 с.

8. *Рубцов Г. Г., Литвиненко А. Н., Большакова Л. В.* Тенденции развития отечественной инновационной политики на примере СЗФО // Научно-технические ведомости СПбГПУ. Экономические науки. – 2020. – Т. 13. – № 1. – С. 65–78.

9. *Дубров А. М., Мхитарян В. С., Трошин Л. П.* Многомерные статистические методы. – Москва: Финансы и статистика, 2003. – 352 с.

References

1. *Mandel I. D.* Cluster analysis. – Moskva: Finance and Statistics, 1988. – 176 s.

2. *Barinova V. A., Drobyshevsky S. M., Eremkin V. A., Zemtsov S. P., Sorokina A.V.* Typology of Russian regions for the purposes of regional policy // Russian Entrepreneurship. – 2015. – Volume 16. – № 23. – S. 4199–4204.

3. *Gordyachkova O. V.* Cluster analysis of attracting foreign investment by Russian regions // Russian Journal of Entrepreneurship. – 2013. – Volume 14. – № 3. – S. 116–121.

4. *Maslyukova E. V., Zaitseva Yu. Yu.* Employment in the informal sector: quantitative methods of analysis // Labor Economics. – 2017. – Volume 4. – № 4. – S. 423–430.

5. *Savchenko T. N.* The use of cluster analysis methods for processing data from psychological research // Experimental Psychology, 2010. – Volume 3. – № 2. – S. 67–86.

6. *Khaidukov D. S.* The use of cluster analysis in public administration // Philosophy of Mathematics: Actual Problems. – Moskva: MAX Press, 2009. – 287 s.

7. *Zhambyu M.* Hierarchical cluster analysis and correspondence. Moskva: Finance and statistics, 1988. – 345 s.

8. *Rubtsov G. G., Litvinenko A. N., Bolshakova L. V.* Trends in the development of domestic innovation policy on the example of the North-West Federal District // Scientific and Technical Journal of St. Petersburg State Polytechnical University. Economic sciences. Volume 13, №1, 2020. – S. 74 – 86.

9. *Dubrov A. M., Mkhitaryan B. C., Troshin L. P.* Multidimensional statistical methods. – Moskva: Finance and Statistics, 2003. – 352 s.

Статья поступила в редакцию 30.12.2019 г.