DOI 10.22394/1726-1139-2020-5-116-127

Иерархические методы кластеризации в задаче поиска аномальных наблюдений на основе групп с нарушенной симметрией[1]

*Кисляков А. Н., Поляков С. В.**

Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации (Владимирский филиал), г. Владимир, Российская Федерация; *svp2206@yandex.ru

РЕФЕРАТ

Работа направлена на решение актуальной проблемы идентификации и интерпретации аномальных наблюдений при исследовании социально-экономических процессов. Предлагаемый в работе метод основан на использовании кластерного подхода к выявлению аномальных наблюдений. Кластеризация выполняется иерархическими методами, которые представляют собой совокупность алгоритмов упорядочивания данных, направленных на создание дендрограмм, состоящих из групп наблюдаемых точек. В качестве метрики расстояний между элементами в случае смешанных данных, состоящих из числовых и категориальных переменных предлагается использовать расстояние Гауэра. Оценка качества кластеризации выполняется на основе показателя суммы квадратов метрических расстояний между объектами внутри кластера и средней ширины силуэта. Эти показатели позволяют выбрать оптимальное количество кластеров и оценить качество результатов разбиения. Дендрограмма может быть использована для исследования групп симметрии кластерных систем и причин нарушения симметрии. Выявление аномалий осуществляется путем анализа результатов иерархической кластеризации и выявления ветвей дендрограммы, располагающихся на начальных уровнях построения дерева и не имеющих ветвлений. Реализованная методика позволяет более точно интерпретировать результаты кластеризации относительно определения ошибок первого и второго рода виде аномальных наблюдений в наборе данных. С помощью описанной методики возможно эффективно исследовать социально-экономические системы и управлять их развитием.

Ключевые слова: кластерный анализ, сетевые графы, нарушение симметрии, аномальные наблюдения, деревья решений.

Для цитирования: *Кисляков А. Н., Поляков С. В.* Иерархические методы кластеризации в задаче поиска аномальных наблюдений на основе групп с нарушенной симметрией // Управленческое консультирование. 2020. № 5. С. 116–127.

Hierarchical clustering methods in a task to find abnormal observations based on groups with broken symmetry

*Aleksey N. Kislyakov, Sergey V. Polyakov**

Russian Presidential Academy of National Economy and Public Administration (Vladimir Branch), Vladimir, Russian Federation; *svp2206@yandex.ru

ABSTRACT

The work is aimed at solving the actual problem of identification and interpretation of anomalous observations in the study of socio-economic processes. The proposed method is based on the use of a cluster approach to detecting anomalous observations. Clustering is performed using hierarchical methods, which are a set of data ordering algorithms aimed at creating dendrograms consisting of groups of observed points. In the case of mixed data consisting of numeric and categorical variables, it is proposed to use the Gower

[1] Исследование выполнено в рамках работ по гранту РФФИ 18-07-00170 А «Создание прогностических моделей эволюции природных, живых и социально-экономических систем на основе конечных групп нарушенной симметрии».

distance as a metric for distances between elements. Clustering quality is evaluated based on the sum of squares of metric distances between objects within the cluster and the average width of the silhouette. These indicators allow you to select the optimal number of clusters and evaluate the quality of the split results. The dendrogram can be used to study the symmetry groups of cluster systems and the causes of symmetry breaking. Anomaly detection is performed by analyzing the results of hierarchical clustering and identifying branches of the dendrogram that are located at the initial levels of tree construction and do not have branches. The implemented method makes it possible to more accurately interpret the results of clustering with respect to determining errors of the first and second kind in the data set. Using the described method, it is possible to effectively investigate socio-economic systems and manage their development.

Keywords: cluster analysis, network graphs, symmetry breaking, anomalous observations, decision trees.

For citing: Kislyakov A. N., Polyakov S. V. Hierarchical clustering methods in a task to find abnormal observations based on groups with broken symmetry // Administrative consulting. 2020. No. 5. P. 116–127.

Введение

Круг задач, решаемых на основе перспективного анализа данных и прогнозного моделирования постоянно расширяется. На основе математических моделей возможна выработка рекомендаций и составление прогнозов показателей социально-экономического развития. Сам подход к использованию данных для принятия решений становится ключевым фактором экономического роста [6; 7]. Данные являются основой для построения прогнозных моделей и от их качества зависит результат принятия решений.

При этом исходный набор данных может содержать нехарактерные показатели или аномальные наблюдения [1]. Существует достаточно способов [7; 16] выявления аномальных наблюдений от самых простых — визуальных, а также методов, основанных на критериях проверки статистических гипотез, до более сложных использующих алгоритмы линейной регрессии и заканчивая рекуррентными нейронными сетями. Обычно выделяют следующие возможные источники аномальных наблюдений.

1. Неумышленная, ошибочная запись значения наблюдаемого показателя — ошибка первого рода, подлежащая устранению.
2. Резкое изменение внешних и внутренних условий, связанных с формированием результата наблюдения (ошибка второго рода, которую следует учитывать в модели).
3. Резкие отличия показаний объектов исследования, неочевидно находящихся в различных условиях функционирования (действие скрытых факторов).
4. Умышленное искажение с определенной целью результатов наблюдений.

В этой связи на первый план выходит задача не только поиска, идентификации, но и интерпретации аномальных наблюдений. Так, например, визуальная оценка наличия таких наблюдений на диаграмме рассеяния не всегда дает однозначный ответ на вопрос является ли аномальным наблюдением данная точка. Еще сложнее обстоят дела, когда изучается поведение человека как составляющей социально-экономической системы на основе вектора признаков (факторов) на основе групп с нарушением симметрии [8; 10].

Подходом к поиску аномалий, заслуживающим особого внимания, является использование методов кластерного анализа. Суть кластерных методов заключается

в оценке метрических расстояний между объектами (точками), характеризующимися векторами признаков. Если значение этого расстояния удалено от центров кластеров более чем на определенную величину, то наблюдение можно считать аномальным.

В этом понимании кластерные методы схожи с метрическими и статистическими методами [1; 15] поиска аномальных наблюдений, однако при более детальном рассмотрении кластерные методы позволяют решить более широкий круг задач, связанных с интерпретацией выявленных аномалий.

Целью работы является разработка метода идентификации аномальных наблюдений на основе иерархической кластеризации. При этом основной задачей является выбор метода иерархической кластеризации, а также метрики расстояний в пространстве признаков. Данный подход позволяет добиться более сбалансированных результатов кластеризации при возможности определения количества кластеров.

При исследовании реальных систем, объектов и их структур, изменяющихся в процессе развития, чаще, где это возможно, обращаются к математическим моделям, относящимся к алгебраическим структурам. Классификацию алгебраических структур можно построить на основе визуализаторов. Визуализатор — тип программного обеспечения, предназначенный для преобразования различной информации в зрительные образы. В качестве одного из результатов работы визуализатора является дендрограмма [17], которая используется для представления результатов иерархической кластеризации. Дендрограмма может быть использована для исследования групп симметрии кластерных систем и причин нарушения симметрии [8; 9] (наличии, появлении аномальных наблюдений).

Методы исследования

Методы иерархической кластеризации представляют собой совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева, графа) состоящего из групп наблюдаемых точек.

Исходными данными для проведения кластерного анализа служит матрица расстояний между объектами, сформированная с использованием той или иной метрики. Распространенная мера удаленности объектов друг от друга, используемая чаще всего — евклидово расстояние [11; 12].

Основным преимуществом иерархических методов кластеризации является отсутствие необходимости начального предопределения количества групп разбиения — кластеров.

Результат иерархической кластеризации представляется в виде дендрограммы [13; 14]. Для разбиения на кластеры также необходимо определить метод объединения, т. е. метод, который позволит выявить наиболее сильные связи между группами объектов и метрики разбиения.

Иерархические методы разбиения на кластеры позволяют выбрать из двух вариантов объединения.

1. Агломеративная кластеризация начинается с n кластеров, где n — число наблюдений: предполагается, что каждое из них представляет собой отдельный кластер. Затем алгоритм пытается найти и сгруппировать наиболее схожие между собой точки данных — так начинается формирование групп.
2. Дивизионная кластеризация выполняется противоположным образом — изначально полагается, что все n точек данных, представляют собой одну большую группу, а далее наименее схожие из них разделяются на отдельные подгруппы.

В этом смысле методы иерархической кластеризации весьма схожи с методом Isolation Forest («изолированный лес») [1], но не являются полностью идентичными. Суть алгоритма Isolation Forest состоит в том, что аномалию можно изолировать с помощью меньшего количества случайных разделений по сравнению с образцом обычного класса, поскольку отклонения встречаются реже и не укладываются в общую статистику имеющегося набора данных.

Так, при «случайном» способе построения деревьев выбросы будут попадать в «листья» (вершины ветвей дерева), т. е. выбросы проще «изолировать». Выделение аномальных значений происходит на первых итерациях работы алгоритма на небольшой глубине построения дерева — когда относительное расстояние между кластерами больше 0,5 [1; 16], потому как именно при этих значениях наблюдается выделение аномалий в отдельную ветвь. Вторым признаком аномалии является отсутствие дальнейшего разветвления этой ветви дерева.

Дендрограмма также имеет древовидную структуру, однако принципы построения отличаются от принципов построения случайного и изолированного леса. При этом иерархическая кластеризация с использованием дендрограмм позволяет интерпретировать найденные аномалии: если точку можно выделить в отдельный кластер, и признаки ее отделимы от всех кластеров, то данное аномальное наблюдение следует учесть в модели, если же аномалия наблюдается только по одному или двум-трем признакам, то существует вероятность, что данное наблюдение является ошибочным и его следует исключить из рассмотрения для получения более адекватных характеристик модели.

Таким образом, предлагается выстроить следующую методику автоматической идентификации аномальных наблюдений на основе методов иерархической кластеризации с использованием показателей отделимости элементов друг от друга внутри кластера и показателя отделимости кластеров между собой:

1) загрузка исходных данных с указанием упорядоченных значений факторов;
2) вычисление матрицы отличий на основе расстояний Гауэра;
3) выполнение иерархического разбиения данных агломеративным и дивизионным методом;
4) построение дендрограммы для каждого вида кластеризации, первичная идентификация аномальных уровней.
5) выбор количества кластеров на основе показателей разделимости признаков точек внутри кластера и разделимости самих кластеров.
6) выбор метода объединения в кластеры.
7) построение дендрограммы с разбиением на кластеры, интерпретация идентифицированных аномалий.

Результаты

Рассмотрим в качестве примера возможность кластеризации тестовой базы данных, содержащей признаки поведенческой активности клиентов сети ресторанов быстрого питания. Тестовая выборка состояла из 100 клиентов ($n = 100$) и нескольких признаков. Пример выборки показан на рис. 1. Для программной реализации кластерных методов использовались пакетные функции: daisy(), diana(), clusplot() из библиотеки кластеризации на языке R.

По сути данная выборка состоит из идентификатора сделки и нескольких показателей по каждому из признаков, характеризующих эти сделки [2; 3]. Каждая из сделок является вершиной сетевого графа («листья» дерева), а связи между этими вершинами характеризуются мерами схожести каждой пары сделок.

Идентификатор	Доход с клиента	Город	Способ заказа	День недели	Тип продукта
1	большой	Кострома	через сайт	ср	десерты
2	большой	Владимир	через сайт	ср	напитки
3	большой	Владимир	через сайт	вс	пицца
4	средний	Кострома	через сайт	ср	напитки
5	средний	Кострома	лично	чт	напитки
6	большой	Кострома	через сайт	сб	блюдо дня
7	большой	Кострома	через сайт	пт	десерты
8	средний	Кострома	лично	сб	напитки
9	большой	Владимир	через сайт	вт	напитки
10	малый	Владимир	лично	вс	пицца
11	малый	Владимир	через сайт	чт	десерты
12	большой	Кострома	через сайт	чт	картофель
13	средний	Кострома	лично	вт	напитки
14	большой	Кострома	через приложение	пт	десерты
15	малый	Кострома	через сайт	чт	пицца
16	средний	Иваново	лично	ср	блюдо дня
17	большой	Ярославль	лично	пт	десерты
18	малый	Владимир	через приложение	ср	напитки
19	малый	Кострома	через сайт	сб	картофель

Рис. 1. Пример тестовой выборки
Fig. 1. Example of test selection

В качестве метрики расстояний между элементами в случае смешанных данных, состоящих из числовых и категориальных переменных предлагается использовать расстояние Гауэра [11], которое при оценке сходства (различия) допускает, одновременное использование смешанных переменных (качественных и количественных), измеренных по различным шкалам

$$S_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}},$$

где W_{ijk} — весовой коэффициент, принимающий значение 1, если сравнение объектов по признаку k следует учитывать, и 0 — в противном случае; S_{ijk} — вклад в сходство объектов, зависящий от того, учитывается ли признак k при сравнении объектов i и j. В случае бинарных признаков $W_{ijk} = 0$, если признак k отсутствует у одного или обоих сопоставляемых объектов. В итоге чем меньше расстояние Гауэра, тем лучше классификация отражает структуру данных. Эта метрика позволяет выстроить структуру дендрограммы.

На основе расстояния Гауэра вычисляется матрица отличий [4; 5], — таблица, заполненная в идеальном случае 0 и 1. Значение «0» означает отсутствие связи между признаками, значение «1» говорит об однозначной связи. Использование такой матрицы отличий позволяет реализовать основной принцип кластеризации: объединить в группы вершины, которые находятся ближе всего друг к другу, или разделить наиболее удаленные друг от друга. На рис. 2 приведены результаты такого разбиения и построены дендрограммы на основе агломеративного и дивизионного подхода.

На рис. 2 можно в обоих случаях наблюдаются «тупиковые» ветви дендрограммы, которые заканчиваются на уровнях построения дерева больше или равных 0,5. Это и есть те самые аномалии, причину появления которых и предстоит выяснить. Для этого необходимо в первую очередь оценить качество кластеризации. Таким образом, идентификация аномалий осуществляется путем выявления ветвей дендрограммы, располагающихся на начальных уровнях построения дерева и не имеющих ветвлений.

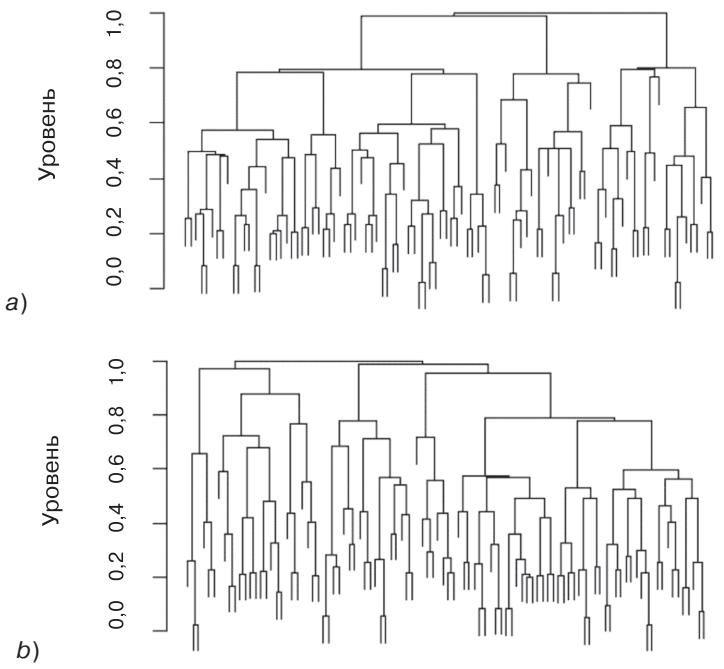


Рис. 2. Результат дивизионной *(a)* и агломеративной *(b)* кластеризации
Fig. 2. Result of divisional (a) and agglomerative (b) clustering

Для реализации высокого качества кластеризации необходимо, чтобы расстояние между точками внутри кластера (или компактность) было минимальным, а расстояние между группами (отделимость) — максимально возможным. Для этого необходимо выполнить расчет основных характеристик качества кластеризации: суммы квадратов метрических расстояний между объектами внутри кластера и среднюю ширину силуэта [12]. Особого пояснения требует показатель ширины силуэта, который для i-го объекта (s_i) определяется соотношением

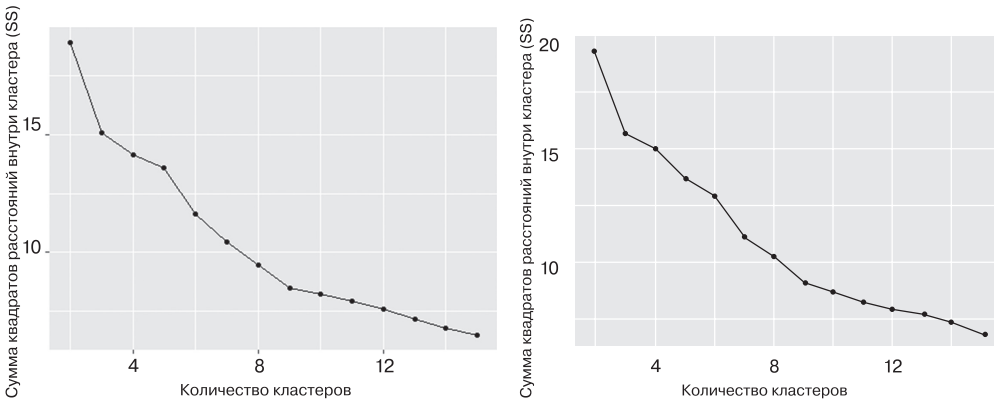
$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)},$$

где a_i — среднее расстояние между i-м объектом и всеми членами кластера, которому он принадлежит (если объект в кластере один, то $a_i = 0$), b_i — среднее расстояние между i-м объектом и членами другого ближайшего кластера (на практике рассчитывается среднее расстояние до членов всех остальных кластеров и выбирается минимум).

Ширина силуэта характеризует степень разброса принадлежности объектов к кластеру и может варьироваться от –1 до 1. Если она близка к единице, объект находится фактически в центре кластера и его принадлежность к нему не вызывает сомнений. Если s_i близко к нулю, объект лежит между двумя кластерами. Отрицательные значения s_i свидетельствуют о, вероятно, неверной классификации объекта.

Резкое изменение показателя суммы квадратов метрических расстояний между объектами внутри кластера (резкий изгиб на графике — рис. 3) при определенном количестве кластеров позволяет сделать вывод о том, что дальнейшее разбиение на кластеры теряет смысл.

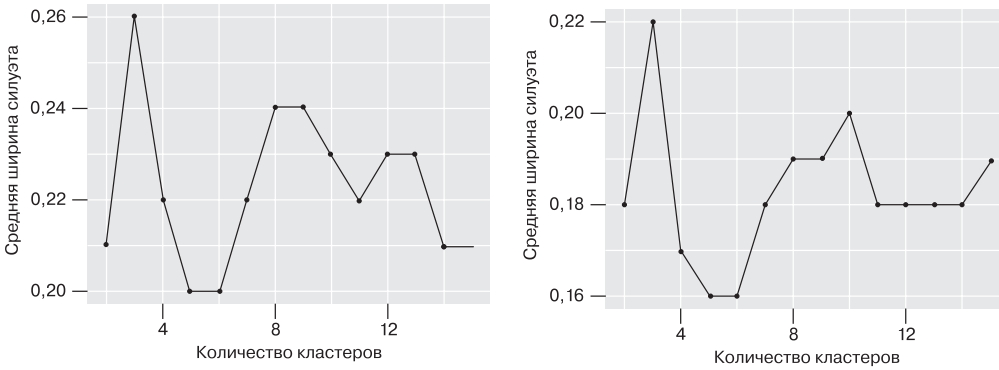
В данном случае (рис. 3) сумма квадратов расстояний между объектами внутри кластера не дает однозначных выводов о выборе количества кластеров. И в том и в другом случае логично предположить, что в данном наборе присутствует 8–10 кластеров.



a *b*

Рис. 3. Сумма квадратов метрических расстояний между объектами внутри кластера для дивизионной (*a*) и агломеративной (*b*) кластеризации
Fig. 3. Sum of squares of metric distances between objects within the cluster for divisional (a) and agglomerative (b) clustering

При использовании метода оценки силуэтов, следует выбирать такое количество кластеров, которое дает максимальную ширину силуэта, потому что для интерпретации результатов нужны кластеры, которые достаточно далеко отстоят от друга, чтобы считаться отдельными. Если набор данных будет разбиваться на более мелкие группы, тем лучше кластеры отделимы друг от друга, однако так дело может дойти до отдельных точек и процесс потеряет смысл. С другой стороны, деление множества на малое количество кластеров усложняет дальнейшую интерпретацию, поэтому целесообразнее оценивать локальные экстремумы. Как видно из рис. 4, ширина силуэтов в обоих случаях имеет локальный максимум при 8–10 кластерах.



a) *b*)

Рис. 4. Средняя ширина силуэта для дивизионной (*a*) и агломеративной (*b*) кластеризации
Fig. 4. Average silhouette width for divisional (a) and agglomerative (b) clustering

Именно это количество и будет использоваться в дальнейшем при построении дендрограмм с разбиением на кластеры (рис. 5).

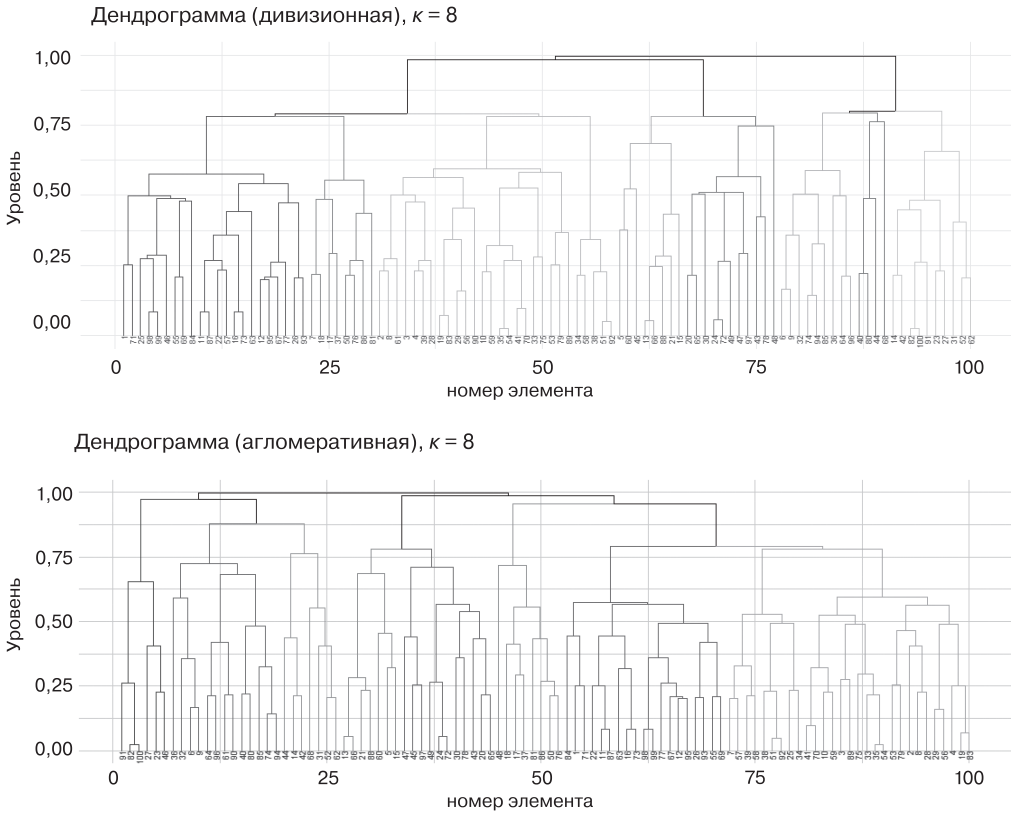


Рис. 5. Результаты иерархической кластеризации
Fig. 5. Results of a hierarchical clustering

Обсуждение

Как видно из рис. 5, наиболее отличными от остальных ветвей дендрограммы являются ветви с наблюдением № 48 и № 68, которые начинаются на уровне большем и равном 0,5 и далее не разветвляются, тем не менее алгоритмы их определили к одному из кластеров. Также можно наблюдать, что оба метода кластеризации показывают различные результаты объединения, однако данные аномалии выделяются бесспорно в каждом их этих методов. Кроме того, при дивизионной кластеризации мы можем наблюдать ветвь, состоящую всего из четырех точек. Эти точки также можно считать аномальными второго рода, однако три из них имеют больше общих признаков.

Теперь для того чтобы оценить, как будет работать алгоритм, умышленно попытаемся внести изменения двух видов в исходный набор данных: сначала необоснованно увеличим показатель дохода от одной сделки при покупке одного из видов продукции — пусть, к примеру, это будет наблюдение № 15, затем заметено увеличим доходы с продаж напитков по субботам (таких наблюдений было немного, их номера: № 52, 53, 62, 63, 79, 89 и 91). Построим теперь дендрограммы и сравним результаты.

Рисунок 6 показывает, что при дивизионной кластеризации наблюдение № 68 уже не является аномальным, зато наблюдение № 48 выделено даже в отдельный кластер в виде одной точки. Кроме того, ожидаемо отнесено к аномалиям наблюдение и № 15. Следует также обратить внимание на наблюдение № 44, которое тоже отнесено к аномалиям. В свою очередь аномалии продаж напитков по субботам даже не выделены в отдельный кластер.

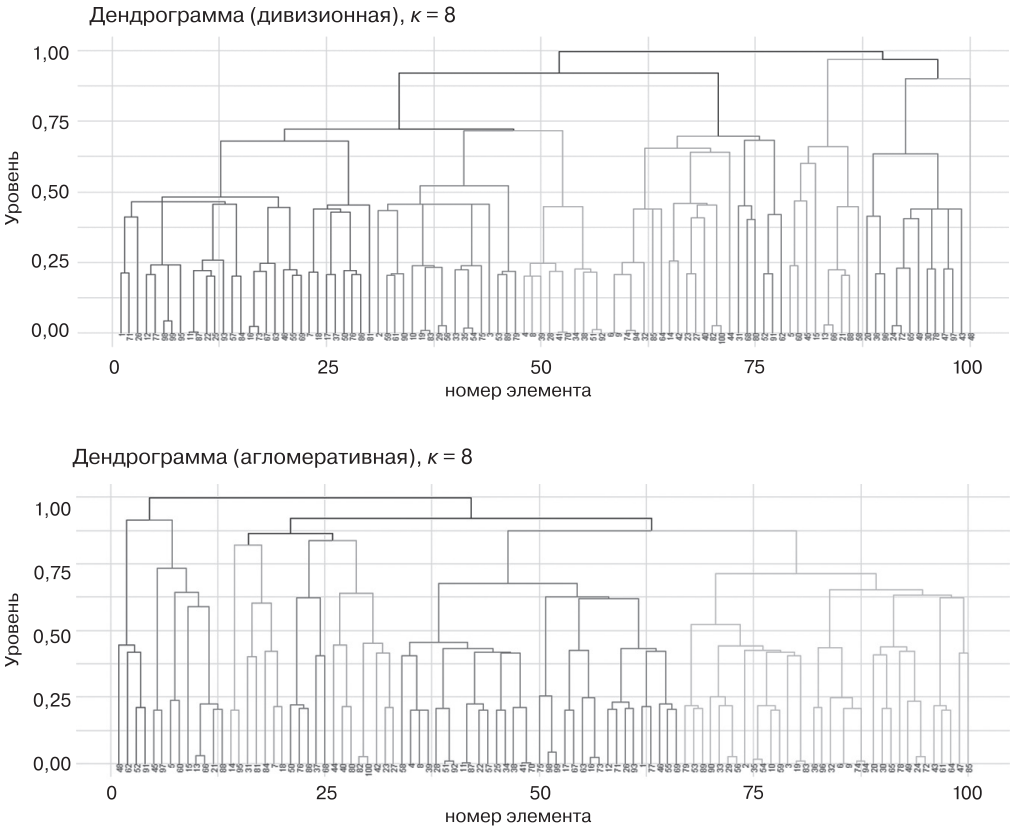


Рис. 6. Результаты кластеризации с аномальными наблюдениями
Fig. 6. Results of clustering with abnormal observations

Если рассматривать результат агломеративной кластеризации, то легко заметить, что здесь даже наблюдение № 48 уже не является аномальным, как, впрочем, и наблюдение № 68. Однако наблюдение № 15 по-прежнему относится к аномальным, кроме этого наблюдения больше не существует ветвей без дальнейшего деления на уровне построения дерева более или равном 0,5. Часть аномальных наблюдений (продажи напитков по субботам — ошибки второго рода) с номерами 52, 62, 91 отнесены к одному кластеру, но не определены как аномалии.

Таким образом, решив подобную обратную задачу, можно сделать вывод о том, что данная методика позволяет выполнить не только поиск аномальных наблюдений, но и объяснить их происхождение для дальнейшего учета в модели. Кроме того, подобный анализ дендрограммы позволяет определить наиболее значимые группы сделок и степень полезности каждого кластера с возможностью построения карты предпочтений потребителей.

Выводы

Результаты анализ дендрограмм позволяет сделать выводы.

1. Для получения более осмысленных результатов интерпретации поведенческой активности объектов в кластерах необходимо выявление аномальных наблюдений в наборе данных.
2. Использование расстояния Гауэра в качестве метрики кластеризации позволяет выполнить расчет расстояний между числовыми и категориальными переменными.
3. Методы иерархической кластеризации дают возможность автоматизированной оценки количества кластеров на основе характеристик качества кластеризации: суммы квадратов метрических расстояний между объектами внутри кластера и средней ширины силуэта.
4. Анализ дендрограмм позволяет более точно интерпретировать результаты кластеризации относительно определения ошибок первого и второго рода виде аномальных наблюдений в наборе данных.
5. Описанная методика дает возможность управлять развитием социально-экономических систем, исследуя свойства симметрии групп элементов и их признаков.

Литература

1. *Барский М. Е., Шиков А. Н.* Исследование алгоритма поиска аномалий isolation forest // Фундаментальные и прикладные научные исследования: актуальные вопросы, достижения и инновации. Сб. статей XXIII Международной научно-практической конференции 5 мая 2019 г. Пенза : Наука и Просвещение (ИП Гуляев Г. Ю.), 2019. С. 113–117.
2. *Белоцерковская М. Г.* Кластеризация клиентской базы участников программы лояльности // Московский экономический журнал. 2017. № 2. С. 112–119.
3. *Галямова А. Ф., Тархов С. В.* Управление взаимодействием с клиентами коммерческой организации на основе методов сегментации и кластеризации клиентской базы // Вестник УГАТУ. 2014. Т. 18, № 4 (65). С. 149–156.
4. *Кисляков А. Н.* Интеллектуальный анализ потребительского спроса в условиях информационной асимметрии // Современная экономика: проблемы и решения. 2019. № 10 (118). С. 8–17.
5. *Кисляков А. Н. Тихонюк Н. Е.* Модель ценообразования однородного рынка с учетом асимметричности информации // Инновационное развитие экономики. 2019. № 1. С. 93–100.
6. *Кисляков А. Н.* Методы и инструменты анализа данных в экономике и управлении: учебно-методическое пособие. Владимир : Владимирский филиал РАНХиГС, 2019. 161 с.
7. *Поляков С. В., Кисляков А. Н.* Основы математического моделирования социально-экономических процессов : учебно-методическое пособие. Владимир: Владимирский филиал РАНХиГС, 2017. 269 с.
8. *Рау В. Г., Кисляков А. Н., Тихонюк Н. Е., Рау Т. Ф.* Принцип нарушения асимметрии в моделях развития экономических систем опыт и проблемы // Региональная экономика: опыт и проблемы. Материалы XI международной научно-практической конференции (Гутманские чтения) 15 мая 2018 г. / под общ. ред. А. И. Новикова, А. Е. Илларионова. Владимир : Владимирский филиал РАНХиГС, 2018. С. 201–211.
9. *Рау В. Г., Поляков С. В., Рау Т. Ф., Фирсов И. В. и др.* Некоторые особенности применения групп нарушенной симметрии для «визуализации» процессов в природных, «живых» и социально-экономических системах // Региональная экономика: опыт и проблемы. Материалы XII международной научно-практической конференции (Гутманские чтения) 15 мая 2019 г. / под общ. ред. А. И. Новикова, А. Е. Илларионова. Владимир : Владимирский филиал РАНХиГС, 2019. С. 111–119.
10. *Тихонюк Н. Е., Кисляков А. Н.* Экономические модели работы с асимметрией информации: эволюция подходов // Региональная экономика: опыт и проблемы. Материалы XI международной научно-практической конференции (Гутманские чтения) 15 мая 2018 г. / под общ. ред. А. И. Новикова, А. Е. Илларионова. Владимир : Владимирский филиал РАНХиГС, 2018. С. 236–244.
11. *Якимов В. Н., Шурганова Г. В., Черепенников В. В., Кудрин И. А. и др.* Методы сравнительной оценки результатов кластерного анализа структуры гидробиоценозов (на примере

зоопланктона реки Линда Нижегородской области) // Биология внутренних вод. 2016. № 2. С. 94–103.
12. *Alboukadel K.* Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning (Multivariate Analysis). Vol. 1. 1st ed. / Publisher: CreateSpace Independent Publishing Platform, 2017.
13. *Murtagh F., Contreras P.* Methods of Hierarchical Clustering // Computing Research Repository — CORR, 2011.
14. *Nielsen F.* Introduction to HPC with MPI for Data Science // Springer International Publishing, 2016.
15. *Gareth J., Witten D., Hastie T., Tibshirani R*. An Introduction to Statistical Learning with Applications in R / Publisher: Springer, 2013.
16. *Hastie T., Tibshirani R., Friedman J*. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer, 2017.
17. *Tripathi Sh., Bhardwaj A., Poovammal E.* Approaches to Clustering in Customer Segmentation // International Journal of Engineering &Technology, 2018, N 7 (3.12). P. 802–807.

Об авторах:

Кисляков Алексей Николаевич, доцент кафедры информационных технологий Владимирского филиала РАНХиГС (г. Владимир, Российская Федерация), кандидат технических наук; ankislyakov@mail.ru

Поляков Сергей Владимирович, доцент кафедры информационных технологий Владимирского филиала РАНХиГС (г. Владимир, Российская Федерация), кандидат технических наук; svp2206@yandex.ru

References

1. Barsky M. E., Shikov A. N. Research of the algorithm for finding anomalies of isolation forest // Fundamental and applied scientific research: topical issues, achievements and innovations. collection of articles of the XXIII International Scientific and Practical Conference on May 5, 2019 Penza: Science and Enlightenment (IP Gulyaev G. Y.), 2019. P. 113–117. (In rus)
2. Belotserkovskaya M. G. Clustering of client base of loyalty program participants // Moscow Economic Journal [Moskovskii ekonomicheskii zhurnal]. 2017. № 2. P. 112–119. (In rus)
3. Galyamova A. F., Tarhov S. V. Management of interaction with clients of commercial organization on the basis of methods of segmentation and clustering of client base // Journal of USATU [Vestnik UGATU]. 2014. V. 18, No. 4 (65). P. 149–156. (In rus)
4. Kislyakov A. N. Intelligent analysis of consumer demand in conditions of information asymmetry // Modern economy: problems and solutions [Sovremennaya ekonomika: problemy i resheniya]. 2019. № 10 (118). P. 8–17. (In rus)
5. Kislyakov A. N. Tikhonyuk N. E. Model of homogeneous market pricing taking into account asymmetric information // Innovative development of the economy [Innovatsionnoe razvitie ekonomiki]. 2019. № 1. P. 93–100. (In rus)
6. Kislyakov A. N. Methods and Tools of Data Analysis in Economy and Management: Educational and Methodological Manual. Vladimir : RANEPA Vladimir branch, 2019. 161 p. (In rus)
7. Polyakov S. V., Kislyakov A. N. Basics of mathematical modeling of socio-economic processes: educational and methodological manual. Vladimir : RANEPA Vladimir branch, 2017, 269 p. (In rus)
8. Rau V. G., Kislyakov A. N., Tikhonyuk N. E., Rau T. F. Principle of violation of asymmetry in models of economic systems development experience and problems // Regional economy: experience and problems. Materials of the XI International Scientific and Practical Conference (Gutman Readings) on May 15, 2018 / under general ed. of A. I. Novikov and A. E. Illarionov. Vladimir: RANEPA Vladimir branch, 2018. P. 201–211. (In rus)
9. Rau V. G., Polyakov S. V., Rau T. F., Firtsov I. V., Togunov I. A. Some features of using groups of broken symmetry to “visualize” processes in natural, “living” and socio-economic systems // Regional economy: experience and problems. Materials of the XII International Scientific and Practical Conference (Gutman Readings) May 15, 2019 / under the general ed. of A. I. Novikov and A. E. Illarionov. Vladimir: RANEPA Vladimir branch, 2019. P. 111–119. (In rus)
10. Tikhonyuk N. E., Kislyakov A. N. Economic models of work with asymmetry of information: evolution of approaches // Regional economy: experience and problems. Materials of the XI International Scientific and Practical Conference (Gutman Readings) on May 15, 2018 /

under general ed. of A. I. Novikov and A. E. Illarionov. Vladimir : RANEPA Vladimir branch, 2018. P. 236–244. (In rus)
11. Yakimov V. N., Shurganova G. V., Cherepennikov V. V., Kudrin I. A., Ilin M. Y. Methods of comparative assessment of results of cluster analysis of hydrobiocenosis structure (on the example of zooplankton of Linda River of Nizhny Novgorod region) // Biology of internal waters [Biologiya vnutrennikh vod]. 2016. № 2. P. 94–103. (In rus)
12. Alboukadel K. Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning (Multivariate Analysis). Vol. 1. 1st ed. / Publisher: CreateSpace Independent Publishing Platform, 2017.
13. Murtagh F., Contreras P. Methods of Hierarchical Clustering // Computing Research Repository — CORR, 2011.
14. Nielsen F. Introduction to HPC with MPI for Data Science // Springer International Publishing, 2016.
15. Gareth J., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R / Publisher: Springer, 2013.
16. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Springer, 2017.
17. Tripathi Sh., Bhardwaj A., Poovammal E. Approaches to Clustering in Customer Segmentation // International Journal of Engineering &Technology, 2018, N 7 (3.12). P. 802–807.

About the authors:

Aleksey N. Kislyakov, Associate Professor of the Chair of Information Technology of Vladimir Branch of RANEPA (Vladimir, Russian Federation), PhD in Technical Science; ankislyakov@mail.ru
Sergey V. Polyakov, Associate Professor of the Chair of Information Technology of Vladimir Branch of RANEPA (Vladimir, Russian Federation), PhD in Technical Science; svp2206@yandex.ru