УДК 519.237.8

Т. И. Чегерова, В. Г. Чегеров (Могилев)

ВОЗМОЖНОСТИ ВИЗУАЛИЗАЦИИ РЕЗУЛЬТАТОВ СТАТИСТИЧЕСКОГО АНАЛИЗА С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА R.

Аннотация. *В статье приводятся графические возможности языка R при проведении кластерного анализа, направленные на обоснованное принятие решения о количестве и составе выделяемых кластеров.*

Summary. *The article presents the graphical capabilities of the R language during cluster analysis, aimed at making informed decisions about the number and composition of allocated clusters.*

Ключевые елова: *кластерный анализ, методы кластеризации, язык программирования R.*

Keywords: *cluster analysis, clustering methods, programming language R.*

В настоящее время большинство научных данных в области психологии, медицины, социологии, фактически всех гуманитарных наук представлены в факторном виде, то есть имеют лингвистическую оценку, иногда оцифрованную по какой-либо шкале, чаще используются разные шкалы для разных данных, а какие-то данные остаются в лингвистическом виде: «Да», «Нет», «Может быть» и т.п. Часто при анализе таких данных стоит задача группировки наблюдений в типичные группы, либо объединение факторов в группы по мере их сходства. Одним из методов, позволяющих решить такие вопросы, является кластерный анализ.

Важным условием применения кластерного анализа является приведение результатов измерений к единой шкале. Общепринятым методом шкалирования является нормализация (стандартизация) данных и далее вычисление Евклидового расстояния. Другой метод – расстояние Говера – шкалирует все переменные, приводя к диапазону 0-1. Метод часто используется со смешанными числовыми и категориальными данными [1, с.290]. В кластерном анализе существует множество методов для установки расстояния между элементами и расстояния между кластерами. Ра-

нее самым популярным был метод Центроида и К-средних, поскольку можно было при некотором навыке осуществить расчет вручную. Сейчас в наиболее популярных статистических пакетах, таких как SPSS, Statistica используется в основном Евклидово расстояние между элементами и метод Варда с использованием квадрата Евклидова расстояния между кластерами, либо метод К-средних. Иллюстрация результатов кластеризации представлена в них, как правило, дендрограммой и графиком средних.

Основной вопрос вызывает не применение определенных алгоритмов кластерного анализа, а графическая интерпретация результатов, так как решение о количестве выделяемых групп (кластеров) остается за исследователем. Сейчас подавляющее большинство ученых в Европе, Америке и Азии при публикации результатов, и при собственно анализе данных используют широкие возможности графической интерпретации языка R, который является продуктом бесплатным и не имеет аналогов по количеству готовых библиотек, создаваемых самими авторами передовых идей в области не только прикладной статистики, но и вообще в области любого анализа данных. Возможности визуализации в R ограничены только знаниями исследователей. Мы представляем ниже свой вариант визуализации этапов кластерного анализа в среде R. Кроме представления данных, при кластерном анализе визуализация необходима, чтобы понять, какие данные попали в каждую группу и в чем отличие групп, сгенерированных компьютером в ходе кластеризации, между собой. В первую очередь это необходимо для интерпретации результатов эксперимента.

В психологии и социальных науках обычно используется агломеративная иерархическая кластеризация, поскольку в настоящий момент этот алгоритм кластеризации считается предпочтительным для работы с факторными данными, хотя единого мнения нет, все зависит от самих данных и характера их распределения.

В качестве иллюстрации использованы данные исследования по формированию познавательных умений детей старшего дошкольного возраста. Всего исследовались 16 познавательных умений, которые характеризовали различные познавательные процессы, такие как восприятие, мышление, внимание, память и воображение. В нашем распоряжении были результаты оценки выполнения диагностических заданий детьми старшего дошкольного возраста, основанных на классических дидактических играх. Каждое диагностическое задание проводилось индивидуально и оценивалось по 4-балльной шкале, что соответствует четырем уровням сформированности познавательных умений детей старшего дошкольного возраста [2]. Проведение иерархического кластерного анализа было обусловлено необходимостью провести эмпирическую классификацию познавательных умений, которая позволила бы упорядочить выделенные умения и разбить их на относительно однородные группы.

Для кластеризации данных использовался агломеративный вариант, как наиболее подходящий для выявления небольших кластеров (в нашем случае всего 16 параметров для 200 детей). Матрица расстояний (или матрица несходства) строилась на основании метода Говера. Далее к матрице расстояний применялись различные алгоритмы кластеризации: метод Варда, комплексный метод, Мак Килти, метод К-средних. Диаграммы «локтя» или «оценка силуэтов» показывают целесообразность разбиения на 4 или 5 групп (рисунок 1).

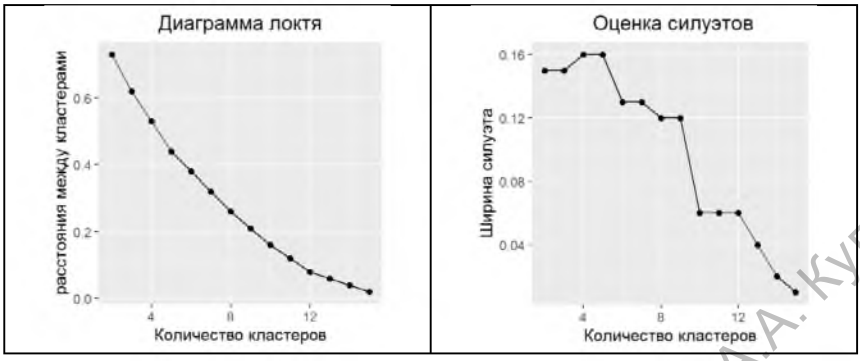

Рис. 1. Диаграммы локтя и оценка силуэтов
для определения оптимального количества выделяемых кластеров

По диаграмме локтя предпочтительно выделить 5 кластеров (расстояние между кластерами перестает заметно увеличиваться), а по диаграмме оценки силуэтов 4 либо 5 кластеров (максимальные значения ширины силуэта) являются равноправным выбором. Наиболее подходящим для данной задачи оказался метод Варда. Джо Вард, работая в области педагогической психологии, впервые опубликовал данный подход к кластеризации данных в 1963 году. В результате своих исследований, им был сделан вывод, что классические методы кластеризации данных (с Евклидовым расстоянием) слабо выявляют кластеры при факторных данных. Он предложил свой алгоритм кластеризации, подходящий именно для факторных данных. Этот метод рассматривает данные с точки зрения дисперсионного анализа. Метод Варда стремится создать небольшие кластеры, в нашем случае это как раз оправдано. Можно уменьшить число кластеров до 4, это позволяет сделать приведенная выше диаграмма оценки силуэтов (равноправно выделить 4 или 5 кластеров).

Применяя метод Варда, строим дендрограмму с выделением 4 и 5 кластеров (рисунок 2).

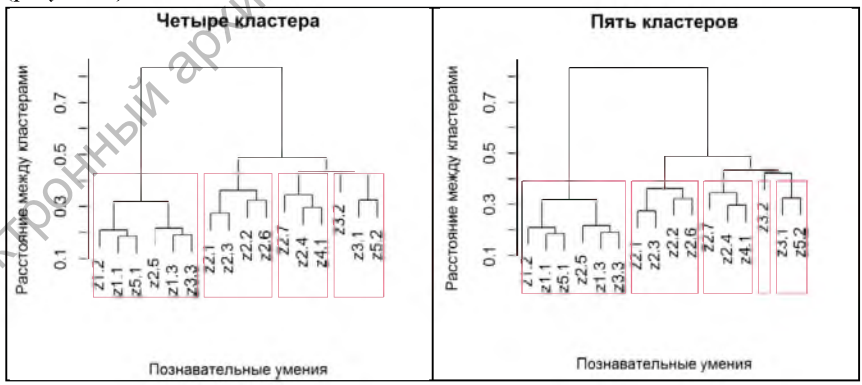

Рис. 2. Агломеративная кластеризация с применением метода Варда

Выбор между приведенными вариантами кластеризации определяется исследователем, с учетом характера переменных, объединенных в один кластер.

Аналогичный подход используется при кластеризации наблюдений эксперимента с целью выявить естественным образом формирующиеся группы в данных. Иллюстрацией является исследование, направленное на выявление структурных и типологических особенностей психологического благополучия детей (n=409) старшего дошкольного возраста [3]. В качестве структурных характеристик психологического благополучия ребёнка рассматривались: 1) система отношений, заданная социальной ситуацией развития (позитивные межличностные отношения со взрослыми, сверстниками, отношение к себе); 2) положительные и нейтральные эмоциональные состояния; 3) субъектная позиция в ведущей деятельности; 4) возрастные психологические новообразования (саморегуляция в познавательной деятельности, децентрация); всего 16 показателей.

Первоначальный этап кластеризации заключался в нормировании переменных относительно их максимального значения. Необходимость осуществления данной процедуры обусловлена существенными различиями диапазона величин изучаемых параметров и недопустимостью объединения разноразмерных показателей. Далее была осуществлена иерархическая кластеризация с использованием метода Варда. В качестве метрики использовалось Евклидово расстояние, что позволило получить наиболее компактные (круглые) кластеры и оптимизировать минимальную дисперсию внутри них. Согласно полученным данным объединённая выборка может быть разделена на три кластера. На рисунке 3 приведена дендрограмма, иллюстрирующая процесс объединения респондентов в группы.

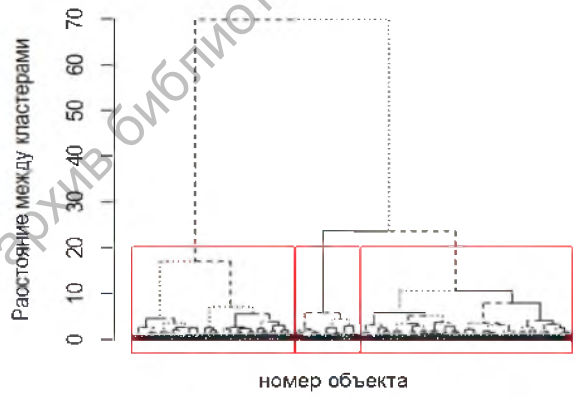

Рис. 3. Дендрограмма объединения кластеров

Визуальная оценка дерева кластеризации показывает, что отчетливо выделяются две крайние группы (n_1=65, n_2=205) и небольшая группа посредине (n_3=139). Для подтверждения обоснованности трёхкластерной структуры был построен график «каменистой осыпи» (рисунок 4).

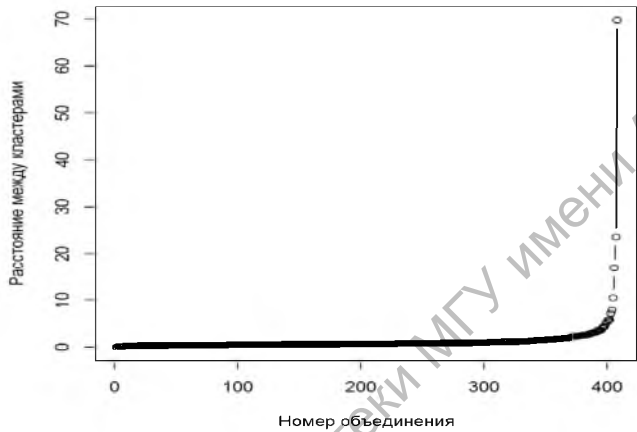

Рис. 4. График «каменистой осыпи»

Данный график позволяет определить количество кластеров, так как показывает расстояние между ними на каждом шаге объединения. Как видно на рисунке, самые большие расстояния приходятся на последние три шага, то есть объединение последних трёх кластеров.

Таким образом, использование графических возможностей языка программирования R дает исследователю дополнительные способы визуализации результатов анализа данных для принятия обоснованного решения определения количества и состава кластеров в проводимом научном исследовании.

Литература

1. Брюс, П. Практическая статистика для специалистов Data Science: пер. с англ. / П. Брюс, Э. Брюс. – СПб.: БХВ-Петербург. 2018. – 304 с.
2. Мукосей, О. М. Теоретические подходы к формированию познавательных умений у детей старшего дошкольного возраста / О. М. Мукосей, Е. И. Комкова // Пралеска. – 2020. – № 8. – С. 3–7.
3. Елупахина, А. В. Современные подходы к исследованию психологического благополучия ребёнка / А. В. Елупахина // Веснік адукацыі. – 2019. – № 8. – С. 50–55.