Методика кластерного анализа для проведения региональных мониторингов качества подготовки обучающихся

ИРО
ИНСТИТУТ РАЗВИТИЯ ОБРАЗОВАНИЯ
Свердловской области

Министерство образования и молодежной политики Свердловской области
Государственное автономное образовательное учреждение
дополнительного профессионального образования Свердловской области
«Институт развития образования»
Региональный центр обработки информации и оценки качества образования

Методика кластерного анализа для проведения региональных мониторингов качества подготовки обучающихся

Екатеринбург
2022

ББК 74.04(235.55)
М 54

Рецензенты:

Н. И. Сапожникова, заместитель начальника МО Управление образованием городского округа Красноуфимск;
М. Ю. Мамонтова, доцент кафедры управления в образовании ГАОУ ДПО СО «ИРО», канд. физ.-мат. наук, доцент

Авторы-составители:

С. В. Алейникова, директор РЦОИиОКО ГАОУ ДПО СО «ИРО»;
С. В. Никитин, заместитель директора РЦОИиОКО ГАОУ ДПО СО «ИРО»

М 54 Методика кластерного анализа для проведения региональных мониторингов качества подготовки обучающихся / Министерство образования и молодежной политики Свердловской области, Государственное автономное образовательное учреждение дополнительного профессионального образования Свердловской области «Институт развития образования», Региональный центр обработки информации и оценки качества образования; авт.-сост.: С. В. Алейникова, С. В. Никитин. – Екатеринбург: ГАОУ ДПО СО «ИРО», 2022. – 61 с.

Данные методические рекомендации являются описанием подходов для осуществления аналитической работы в сфере управления качеством образовательных результатов, образовательной деятельности.
Данное издание содержит региональную методику кластерного анализа, применяемую в рамках региональной системы оценки качества образования Свердловской области (далее – РСОКО), в том числе:

- описание подходов к кластерному анализу;
- описание групп данных, собираемых в рамках РСОКО;
- математический метод, необходимый для анализа данных в рамках РСОКО;
- описание возможных групп кластеров, выделяемых при помощи данной методики кластерного анализа;
- направления использования сведений о выделяемых кластерах.

Региональная методика кластерного анализа является компонентом РСОКО, применяется организацией – оператором РСОКО Свердловской области, ГАОУ ДПО СО «Институт развития образования», в рамках управленческого цикла по управлению качеством образования системы образования региона.

Утверждено Научно-методическим советом ГАОУ ДПО СО «ИРО» от 28.03.2022 № 4

ББК 74.04(235.55)

Оглавление

Введение

Данные методические рекомендации направлены в первую очередь на ознакомление с понятием и областью применения кластерного анализа, ставшего в последние годы не только популярным методом самостоятельных исследований, но и одним из широко используемых инструментов интеллектуального анализа данных (data *mining* – англ.).

Вместе с тем публикуемые исследования, в том числе описывающие различные области применения кластерного анализа, имеют различные терминологические особенности, касающиеся как самого понятия «кластер», так и видов и способов кластерного анализа и его механизмов – «меры близости/родства/схожести», «меры различия».

Меняются и способы проведения анализа, математико-статистические подходы. Безусловно, это обусловлено возможностью проведения гораздо более сложных расчетов с использованием средств компьютерной техники, а также развитием программных продуктов с одной стороны и расширением набора услуг с другой.

Понятие больших данных (big data) заставило обратить внимание на совершенствование механизмов кластерного анализа по отношению к данным различного вида, разработку, с одной стороны, «высокоточных» и «целевых» методов кластерного анализа, а с другой – наоборот, «единых подходов» и «единых механизмов».

С появлением и наращиванием образовательной статистики и пониманием необходимости построения «управления на основе данных», с развитием «интеллектуального анализа данных в сфере образования» (*educational data mining*, EDM – англ.) «развивающейся дисциплине, связанной с разработкой методов изучения уникальных типов данных, поступающих из образовательных организаций, и использованием этих методов для лучшего понимания учащихся и условий, в которых они учатся» (определение согласно международному консорциуму по интеллектуальному анализу данных в сфере образования) [56] потребовалось решение задачи разделения образовательных систем на кластеры для разработки модели улучшения образовательного процесса и институциональной эффективности. Безусловно, любые публикации в данной сфере должны быть изучены управленцами каждого уровня для понимания глубинных взаимосвязей в образовательных системах и вычленения изменений, дающих синергетический эффект при умелом управлении.

Выделение групп обучающихся с разным уровнем подготовки является довольно распространенной практической задачей «аналитической группировки».

Дополнительно стоит отметить, что довольно часто результаты объединения по одному признаку и кластеризация по ряду признаков соответствуют друг другу. В общем виде это правило демонстрирует «принцип суперпозиции», а в вырожденном случае, в котором число используемых при соединении в кластер пар признаков может быть равным единице, задача группировки по одному признаку сводится к задаче кластеризации. С другой стороны, используемая при объединении признаков мера близости сводит их к одному признаку и далее разбиение на кластеры равнозначно группировке по этому признаку. О путях формализации последней задачи мы уже писали в других наших работах – это и разбиение по «процентилям», и по квадратичным отклонениям, и по максимуму «локального расстояния», и по относительному расстоянию, по внутрикластерному коэффициенту вариации и т. д.

Еще одной популярной темой при обсуждении кластерного анализа являются «используемые инструменты». Популярность data mining и мода на использование «инструментов» такого анализа пришлись на развитие технологии блокчейна. Сегодня на практике исследователь чаще всего использует заранее запрограммированные инструменты (например, SPSS или его аналоги), в редком случае создавая их при помощи каких-то нейросетей или инструментов программирования типа библиотек для Python; основная же задача исследователя – в обоснованном выборе данных инструментов и правильной предварительной чистке данных. Из этого следует, что исследователю необходимо хорошо изучить теорию и практику применения различных методов кластерного анализа, а также осмыслить технику кластерного анализа с точки зрения общих методологических принципов статистической науки.

Предлагаемые методические материалы направлены на достижение первой из указанных целей.

Глава 1. Место кластерного анализа в управлении системами образования

1.1. Понятие кластера и кластерного анализа

В переводе слово «кластер» (англ. *cluster* – гроздь, группа, скопление) обозначает единицу, которая характеризуется самостоятельностью (отдельностью и отличиями) от иных групп и объектов. При этом большое количество современных авторов отмечают, что «каждое современное исследование уточняет характерное для него рабочее определение кластера» (Данилов [81, с. 149], Демидов [82, с. 40] и др.).

Общеупотребительное определение кластера появилось в конце 80-х годов XX века. Его предложил профессор Гарвардской школы бизнеса Майкл Юджин Портер. ***Кластер*** – это «группа географически соседствующих взаимосвязанных компаний и связанных с ними организаций, действующих в определенной сфере и характеризующихся общностью деятельности и взаимодополняющих друг друга» [91, с. 258].

Кластерный анализ – это «неконтролируемый метод, используемый для группировки объектов, близко расположенных в многомерном пространстве признаков, используемый обычно для выявления некоторой внутренней структуры, которой обладают данные» (Дж. Брок, С. Пиюр, С. Датта) [9, с. 1].

В книге Дюрана и Оделла приводится «**задача кластерного анализа**», позволяющая лучше понять данный метод: «Пусть *m* – целое число, меньшее, чем *n*. Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся в множестве *X*, разбить множество объектов *I* на *m* кластеров (подмножеств) <…> так, чтобы <…> объекты, принадлежащие одному и тому же кластеру, были *сходными*, в то время как объекты, принадлежащие разным кластерам, были неоднородными (*несходными*)» [84, с. 15]. Решением задачи кластерного анализа является сегментация, удовлетворяющая некоторому *критерию оптимальности*.

Выделим в определении слова о «неконтролируемом методе», или, как стало популярно называть его с появлением понятия «интеллектуального анализа данных» (устоявшийся на текущий момент перевод с английского языка словосочетания data mining), «методе без обучения», «методе без учителя», который означает, что для его использования нет необходимости производить предварительную экспертную классификацию (обучение – специализированный

набор методов в машинном обучении[1], предполагающий предварительную «настройку» математической модели на некоторой «выборке») или подтверждение данных в процессе работы (еще одна группа методов, подразумевающая «обучение» методом подтверждения в процессе реальной работы алгоритма).

Метод можно применять для вычленения однородных групп объектов, называемых кластерами, для последующего наблюдения за ними. Безусловно, наблюдения за объектами одного кластера должны иметь много схожих характеристик, но отличаться от наблюдений в других кластерах. (У Моуи и Сарштейда [48] под наблюдением подразумевается измерение некоторых параметров в динамике, в том числе дискретно.)

1.2. Области применения кластерного анализа

Подходы кластерного анализа широко используются в современной информатике, или науке о данных (от английского *data science*), для моделирования в области искусственного интеллекта и машинного обучения (публикации Иллинойского университета [36] [31] и многие другие).

Накопление же данных в других науках и практических областях рано или поздно приводит к возможности исследователей попробовать универсальные по сути методы информатики, применяя их к определенным наборам данных в других областях.

На сегодняшний день в практических задачах широко используются кластеры в маркетинге (например, Чакрапани, 2004 [13]), биоинформатике и генетике (например, Селински и Икстадт, 2008 [64]), археологии (Саттон и Рейнхард, 1995 [71]). В клинических исследованиях по психиатрии (Клатворди и др., 2005 [16]), включая классификацию пациентов (например, Чигнел и Стейси, 1981 [14]; Сампогна, Сера и Абени, 2004 [62]) и медсестер (Хиллхаус и Адлер, 1997 [29]). В социальных науках предлагаются исследования по популяции населения (Фагхани Сейдех Захра, И. Явшид, Никпор Аболфазл, 2016 [24]), исследования по индивидуальному подходу при интервьюировании как метода сбора данных в различных социальных науках (Ф. Радмер, Х. Аламолходел [23]).

С 2019 года развиваются проекты для обучения нейронных сетей и искусственного интеллекта для работы с данными по новой коронавирусной инфекции. Например, в Сети можно найти проект «COVID-19: наборы данных, программные интерфейсы и списки проектов», опубликованный в открытом доступе на GitHub (https://github.com/sfu-db/covid19-datasets.git), или готовые приложения для анализа наличия коронавирусной инфекции по рентгеновским снимкам методом кластерного анализа: https://cloud.yandex.ru/cases/radlogics.

[1] Machine learning в современной науке об информации.

Использование кластерного анализа в практических исследованиях реализует идею «управления системами на основе данных» (от английского data-driven management), описанную в [98], [83], для выявления групп, требующих внешнего вмешательства в управление процессами.

Например, в уже упоминавшемся нами исследовании по практической психологии Клатворти и др. кластеризация использовалась для управления «внедрением компании по укреплению здоровья, сервисов по снижению риска развития заболеваний и с риском неблагоприятных результатов исследования» (Клатворти и др., 2005 [16]). В системе образования Российской Федерации Федеральный институт оценки качества образования использует тот же подход управления на основе данных «с целью группирования школ по контекстным факторам, обуславливающим низкие результаты» (методика ФИОКО [90, с. 3]), или, наоборот, для выявления лучших практик управления, например, исследование по выявлению резильентных школ (методика ФИОКО [78], являющаяся адаптацией практики исследования PISA [63]).

Метод кластеризации давно и обширно используется в исследованиях в области образования – от стилей обучения (Беннет, 1975) [7] и образовательных концепций (Шавельсон, 1979) [65] до группировки учебных заведений (например, Боронико и Чокси, 2012) [8] и классификации учащихся (например, Ексельтурк и Топ, 2013) [77]. В педагогической психологии, например, проводятся подробные кластерные исследования «успеваемости учащихся (например, успеваемости по математике) в рамках разных кластеров, созданных на основе нескольких переменных (например, математической тревожности, объема рабочей памяти и т. д.)» (цитата об исследовании [45] по [23], перевод авторский).

За последние годы были предложены, применены и протестированы различные методики интеллектуального анализа данных в применении к сфере образования, в результате которого исследователи сходятся во мнении, что общие методы и алгоритмы исследований не подходят для применения в области управления образованием. Основной же задачей извлечения данных остается «преобразование необработанных данных, поступающих из образовательных систем, в полезную информацию, которая потенциально может оказать большее влияние на образовательные исследования и практику» [11].

Считается, что методы интеллектуального анализа образовательных данных должны отличаться от стандартных методов из-за «многоуровневой иерархии и независимости образовательных данных» [5, с. 112]. При этом достижения учащихся все чаще связывают с «атмосферой деятельности» образовательных организаций [6].

Проведены исследования в области сохранения и отчисления студентов [34] [60] колледжей, в результате которых появился метод прогностического моделирования уровня удержания студентов.

Протестированы рекомендательные системы на основе логики «продающих» сайтов, и сделаны попытки применения схожей логики в образовательном контексте (не увенчались успехом, поскольку такие рекомендации сильно зависят от области знаний [50]).

Ромео и Вентура [12] сделали обзор статей, в котором представлены публикации в период с 1995 по 2005 год по интеллектуальному анализу образовательных данных, кластеризации, классификации, анализу ассоциативных правил, анализу текста.

Зайан и Люо предложили [52] применить методы интеллектуального анализа данных к изучению онлайн-курсов, предложили [51] правила ассоциации и кластеризации для поддержки совместной фильтрации для разработки более чувствительных и эффективных систем электронного обучения.

Бейкер и др. [55] провели тематическое исследование методов прогнозирования использования интерактивных сред обучения.

Мейсрон и Язиф [47] предоставили инструменты, которые можно использовать для поддержки образовательного интеллектуального анализа данных. А Бек и Вульф [33] исследовали разработку моделей обучающихся с использованием методов прогнозирования. Следует отметить, что моделирование учащихся является новой исследовательской дисциплиной в области интеллектуального анализа данных в образовании [56].

Другая группа исследователей (Гарсия, Манеро, Вентура и др. [21]) разработала инструментарий, который работает в рамках систем управления курсом и может предоставлять извлеченную добытую информацию пользователям, не являющимся экспертами.

Методы интеллектуального анализа данных использовались для создания динамических учебных упражнений, основанных на прогрессе учащихся в курсе обучения английскому языку (например, Янг и Ляо [76]). Хотя большинство систем электронного обучения, используемых образовательными учреждениями, применяются для публикации или доступа к материалам курса, они не предоставляют преподавателям необходимых инструментов, которые могли бы тщательно отслеживать и оценивать все действия, выполняемые учащимися, чтобы оценить эффективность курса и процесса обучения [43].

Параллельно с этим разрабатываются различные программные решения для анализа данных, комбинирующие алгоритмы таким образом, чтобы помочь исследователям найти ответы на конкретные вопросы, но проблема в том, что такое программное обеспечение требует изучения[2].

[2] Можно отметить SPSS от IBM и библиотеки Python, R; альтернативой им могут быть решения с открытым исходным кодом Weka, Rapid Mineret и др.

Кластерный анализ в публикациях последних лет об исследованиях фронтального образования (образования непосредственно в классе) возникает не так часто, поскольку в последние годы (и особенно в годы пандемии COVID-19) именно электронное образование и его эффект вызывали наибольший интерес: фактически мир столкнулся с 100%-ным онлайн-образованием, чего ранее никогда не происходило.

1.3. Место кластерного анализа в управлении системами образования

При этом в России продолжаются исследования в рамках проектов Рособрнадзора «управление на основе данных» (в понимании Тима Филлипса [98]). Методика кластерного анализа является основой при выявлении действительно значимых факторов управления «черным ящиком» с использованием усовершенствованного цикла Деминга [100, с. 13], показанного на рисунке ниже.

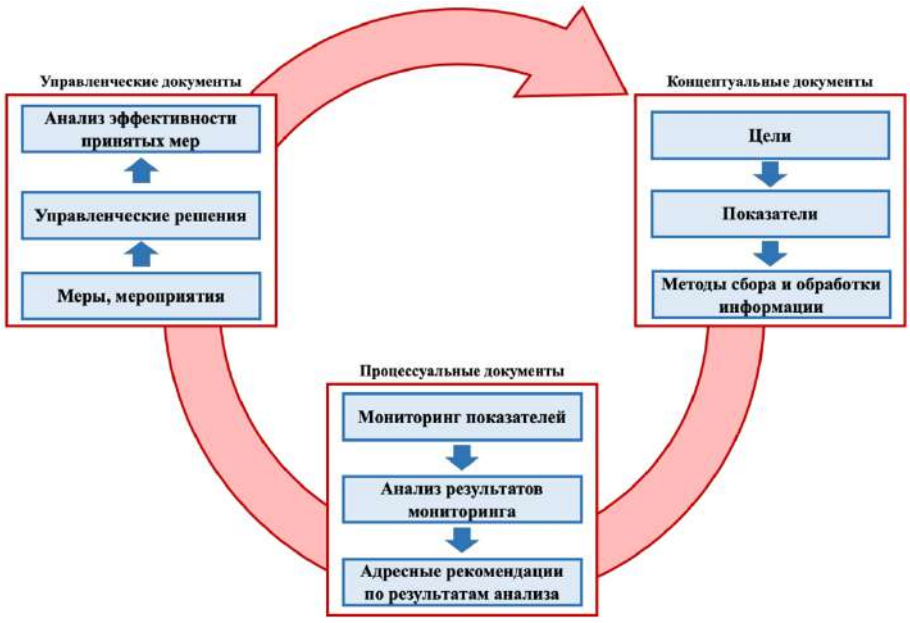

Рис. 1. Структура управленческого цикла

Напомним, что в «классической литературе» по управлению качеством выделяются четыре составляющие: «планируй», «выполняй», «проверяй», «управляй».

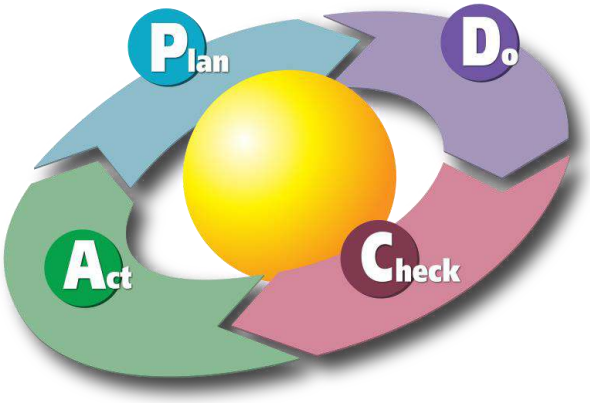

Рис. 2. Классический цикла Деминга

Читателю может показаться, что место кластерного анализа и его результатов – в разделе «проверяй». В действительности это не так. Место кластерного анализа – в определении параметров объекта управления.

Рассматривая образовательную организацию в качестве математической модели, обычно игнорируют процессы внутри системы, предполагая при этом, что значения на выходе функционально зависят от состояний на входе. «Вход» и «выход» – это внешние переменные, а у самой системы имеются какие-то параметры; управленческие воздействия приходятся на изменение «входа» таким образом, чтобы получить необходимый «выход». В этой модели задача кластерного анализа – выявить значимые переменные и параметры.

В классической теории управления одним из значимых параметров может быть «время реакции системы», или (в определении теории автоматизированного управления) «постоянная времени» системы [96], – время, через которое характеристика процесса изменится в *e* раз (*e* = 2,718) при мгновенном единичном воздействии. Например, когда результаты повысятся втрое, если увеличить финансирование системы на 1 единицу. Или при изменении компьютеров на 1 единицу. Или при покупке 1 единицы какого-то необходимого программного обеспечения. Заметим, что масштаб такой единицы измерения не указан и его следует подобрать.

Задача выбора значимых для управления «черным ящиком» параметров – это классическая задача идентификации объектов управления, которая широко описана в литературе, посвященной вопросам идентификации динамических систем, например, Д. Гропп [80], Э. П. Сэйдж и Дж. Л. Мелса [94, 95], Л. Льюнг [86], а также книги Я. З. Цыпкина [101], Н. С. Райбмана [93], Ш. Е. Штейнберга [102] и др. Современные учебники по идентификации (например, учебник издательства МГТУ им. Н. Э. Баумана [87], Санкт-Петербургского государственного электротехнического университета «ЛЭТИ»), а также монографии (например, [97]) и учебные пособия (например, [85]) содержат методы корреляции как подход к определению наличия функциональных зависимостей между объектами.

Если считать, что на «входе» в «черный ящик» есть некоторый набор переменных *X*, а на «выходе» – набор переменных *Y* и между ними существует корреляционная связь *Ф*, при которой $y_i = f(x_i)$, то задача идентификации объекта сводится к установлению такой взаимосвязи и установлению самой функции f, а в соответствии с принципом суперпозиции $\Phi = \sum f_i$.

Читатель может задаться вопросом: каково же тогда значение и место кластерного анализа в процессе управления, если достаточно идентифицировать все возможные составляющие функции f?

В этом месте мы подходим к пониманию важного факта: объекты управления «школа» отличаются друг от друга настолько, что имеют не только разные функции, но и разные значимые параметры. Доказательством этого факта является сам факт того, что введено понятие «резильентная школа», то есть школа, которая функционирует (с точки зрения результатов учащихся) хорошо, несмотря на определенные факторы неблагополучия.

То есть при анализе некоторой интегративной школы с использованием интегративной характеристики были получены параметры, обуславливающие высокую степень зависимости, и доказано, что данная зависимость отрицательно влияет на результат. Пример такого параметра – «отдаленность от административного центра региона», индекс которого можно считать различными методами. «Резильентность» при этом указывает на то, что параметр никак не повлиял на результат (подробно о резильентных школах в методике ФИОКО [78] и практике исследований PISA [63]).

Описанное выше говорит нам о том, что в формуле $Y = \Phi(X)$ и допущении принципа суперпозиции $\Phi = \sum f_i$ сделана ошибка. Возможно, эта ошибка заключается в необходимости учета масштаба (веса) параметра каждой составляющей, которая может выражаться в виде $\Phi = \sum k_i f_i$ и рассчитываться по формуле $y = k_1 x_1 + k_2 x_2 + \cdots + k_n x_n$, но и в этой формуле сделано допущение, существует линейная зависимость. А любому здравому человеку понятно, что она гораздо более высокого порядка.

Из-за отсутствия конечной математической модели, учитывающей одновременно хотя бы пару параметров, современным исследователям в области управления образованием приходится либо исследовать зависимости в «качественных» характеристиках, либо работать исключительно со статистикой и игнорировать «выбросы» (в данном случае под ними подразумеваются системы, ведущие себя не так, как среднестатистические). Важно, что именно в этих «выбросах» заключаются значимые особенности внутреннего устройства образовательной системы.

Рассмотрим пример. К внешним входным переменным относят совокупность показателей, характеризующую начальные способности учащегося, социально-экономический статус семей учащихся, часть материально-технических

ресурсов школы (размер школьных кабинетов, наличие спортивных площадок, медицинских кабинетов, зон релаксации, размеры финансирования на одного учащегося и т. д.), тогда как выходные данные чаще всего включают результаты экзаменов, оценки учебных достижений учащихся (ВПР, контрольные и т. д.) и выпускников школ (ГИА) и полный перечень освоенных ими знаний и умений (элементы требований ФГОС). Разделение школ на однородные группы позволяет применять уравнения множественной регрессии для прогнозирования результатов и применения методов управленческого воздействия. Важным становится вопрос оснований для деления на однородные группы. Кластерный анализ отвечает на данный вопрос, указывая на реально важные параметры и относя школу к одному из них.

Например, в рамках одной муниципальной системы параметр «размеры финансирования на одного учащегося» может оказаться несущественным. В 2012–2014 годах исследования ГАОУ ДПО СО «ИРО» указывали на то, что размер финансирования на одного учащегося не влияет на результаты ЕГЭ и муниципальные образования с б***о***льшим объемом финансирования имеют такую же долю низких результатов, что и образования с вчетверо меньшим финансированием.

Таким образом, можно рассматривать две составляющие – начальный уровень подготовки обучающихся (поделим, к примеру, на «низкие», «средние», «высокие») и индекс социального благополучия семей (поделим так же на 3 уровня). При прямом применении метода классификации получим 9 возможных сочетаний:

- низкие + низкие, низкие + средние,
- низкие + высокие, средние + низкие, средние + средние,
- средние + высокие,
- высокие + низкие, высокие + средние, высокие + высокие.

Кластерный анализ покажет, сколько реальных кластеров получается, и их число может быть больше или меньше указанного.

Вторая важная особенность кластерного анализа – нахождение «выбросов», в которых заключаются значимые особенности. Дополнительное изучение «выбросов» позволяет ответить на вопрос, почему они произошли. С одной стороны, это может быть связано с необъективностью и необходимостью «чистить сырые данные», с другой – с наличием неучтенных значимых для системы образования параметров системы.

1.4. РСОКО

В соответствии с федеральными подходами и рекомендациями по управлению качеством образования региональная система оценки качества образования Свердловской области разработана с учетом девяти компонентов управленческого цикла (рис. 1), восьми направлений оценки (рис. 3).

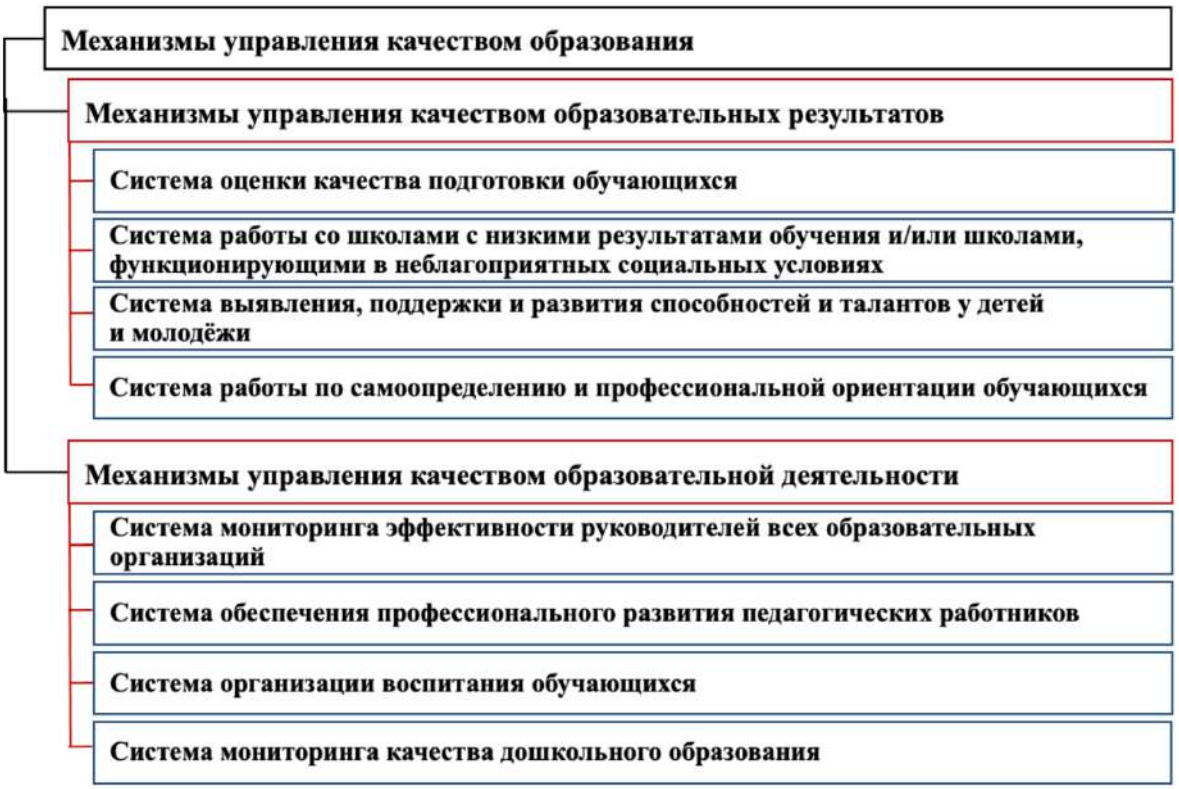

Рис. 3. Направления РСОКО

Кластерный анализ является одним из методов анализа данных, применяемых в РСОКО. Кластерный анализ встраивается в управленческий цикл на этапе работы с мониторингами по направлениям РСОКО 1.1–1.4 включительно, а также на этапах разработки адресных рекомендаций, анализа эффективности принятых мер. Главная задача кластерного анализа – выделить значимые кластеры и значимые факторы, влияющие на результаты обучающихся, и тем самым способствовать поиску более эффективных мер, мероприятий, большей эффективности управленческих решений.

Группы данных, которые включены в анализ качества образовательных результатов, определяются региональными нормативными актами, распорядительными документами, положениями о соответствующих мониторингах:

- Положение о мониторинге региональных показателей качества подготовки обучающихся в рамках реализации направления «Система оценки качества подготовки обучающихся» региональной системы оценки качества образования в Свердловской области;

- Положение о комплексном мониторинге показателей по реализации направления «Поддержка школ с низкими результатами обучения и/или школ, функционирующих в неблагоприятных социальных условиях» в рамках реализации региональной системы оценки качества образования в Свердловской области;
- Положение о проведении мониторинга качества системы выявления, поддержки и развития способностей и талантов у детей и молодежи в Свердловской области.

Глава 2. Рассматриваемые параметры кластеризации

Кластеризация проводится на основе предварительно собранных данных. В этой главе мы рассмотрим данные специфических макро- и микрохарактеристик для кластеризации и методы их сбора, выделяемые в различных источниках.

2.1. Переменные «входа»

К данным характеристикам следует относить объективные параметры, характеризующие образовательную организацию, определяемые учредителем совместно с администрацией школы (внешние факторы): нормативные основания приема, территориальные факторы, наличие конкурсного отбора при поступлении, при переходе в 5-й класс, при переходе в 8-й класс, при переходе в 10-й класс; социокультурные барьеры (образование, сфера занятости и должность родителей), «статус» школы.

2.2. Параметры системы (внутренние параметры)

Наличие специальных образовательных программ (специализация ОУ на отдельных предметах или направлениях), дополнительных услуг, договорных отношений с вузами, профессиональные характеристики преподавателей. Состояние материально-технической базы школы, библиотечного фонда, программа «Школьный автобус». Ассигнования (бюджетные и внебюджетные источники). Профилизация образовательных программ. Сведения о педагогах (стаж, возраст, квалификационная категория, уровень образования, повышение квалификации, наличие переподготовки и др.).

2.3. Зависимые переменные «выхода» (результата)

Рассматриваются результаты конкретной процедуры в виде среднего балла по школе и статистического распределения.

2.4. Определение размерности данных для кластерного анализа

Для проведения кластерного анализа применяются следующие предполагаемые иерархические кластеры:

1. По типу муниципального образования:

- сельские
- городские

2. По количеству жителей:

- менее 30 тысяч жителей
- от 30 до 100 тысяч жителей
- от 100 до 250 тысяч жителей
- от 250 до 500 тысяч жителей
- от 500 тысяч до 1 миллиона жителей
- более 1 миллиона жителей

3. По статусности ОО:

- статусные ОО
- массовые ОО
- вечерние ОО

4. По наличию отбора в классы:

- наличие отбора
- отсутствие отбора

5. По профилю класса:

- общеобразовательный (нет профиля)
- физико-математический
- физико-химический
- химико-биологический
- инженерный
- биолого-географический
- социально-гуманитарный
- социально-экономический
- филологический
- информационно-технологический
- агротехнологический
- художественно-эстетический
- оборонно-спортивный
- математический
- физический
- химический
- технологический
- медицинский
- биологический

- географический
- гуманитарный
- с углубленным изучением иностранных языков
- экономический
- технологический
- академический
- кадетский
- экологический
- смешанный
- иное
- естественно-научный
- химико-математический

6. По категории учителей, преподающих предмет:

- высшая категория
- первая категория
- соответствие занимаемой должности
- без категории

7. По педагогическому стажу учителей, преподающих предмет:

- числовое значение стажа

8. По уровню образования учителей:

- высшее образование – специалитет
- высшее образование – магистратура
- высшее образование – бакалавриат
- переподготовка
- без высшего образования

9. По количеству часов в год:

- числовое значение количества часов

10. По выбранному УМК для преподавания:

- числовой код УМК из перечня, рекомендованного Минпросвещения

11. Уровень реализуемых общеобразовательных программ:

- начальное общее
- основное общее
- среднее общее

12. По результатам оценочной процедуры:

- индекс среднего балла

2.5. Методы сбора информации

Для сбора информации используется региональная база данных подготовки и проведения ГИА (далее – РБД). В РБД уже содержится информация о муниципальных образованиях, ОО, участниках. Также в РБД происходит сбор данных о результатах участников и информации о классах, включая сведения об учителях, преподающих предметы, УМК и количестве часов в году на изучение предмета. В РБД вносятся сведения обо всех учителях и классах (без учащихся) школ.

Глава 3. Этапы кластерного анализа

Алгоритм кластерного анализа для образовательных систем содержит до шести шагов (Моуи и Сарстедт, 2011 [48]; Пастор, 2010 [54], Радмер и Аламолходей, 2014 [23]):

1) подбор переменных для кластеризации;
2) процедура кластеризации;
3) выбор меры сходства/различия;
4) выбор алгоритма кластеризации;
5) определение количества кластеров;
6) проверка и интерпретация кластерного решения.

Представим их подробно.

3.1. Подбор переменных для кластеризации

При выборе переменных для кластерного анализа исследователи стараются опираться на числовые данные (более широкий набор данных рассмотрен у Джавей Хана, 2021 [37]). Возможны данные в виде категорий (например, профили классов, учебно-методические комплекты), включая двоичные (да/нет, за/против и т. п.). Для кластерного анализа важно, чтобы такие данные были дискретными.

Данные в виде текста, мультимедийные данные, временных рядов, последовательностей, потоков, графов, гомогенных (однородных) и гетерогенных сетей описаны у Джавей Хана. Отдельное внимание можно уделить неуверенным данным («Инструментарий неопределенности для аналитиков в Правительстве» [74]) и большим данным.

Большинство данных в образовании сегодня категоризированы либо переведены в дискретные значения на числовых осях. В связи с этим мы не будем рассматривать специфичные методы обработки этих данных.

Степень коллинеарности/корреляции между параметрами
После выбора значимых переменных для анализа, подбора возможных типов данных следует проверить данные на степень их коллинеарности.

При наличии только числовых данных (наиболее частая ситуация в педагогических измерениях) исследователю достаточно убедиться в том, что между переменными невысока степень корреляции. Например, методом подсчета коэффициента корреляции по полученным наборам данных об измерениях.

Если коэффициент корреляции Стьюдента (или Пирсона) между переменными составляет более 0,9, то следует отказаться от использования пары показателей и сократить количество переменных перед проведением кластерного анализа.

Как показывал в своем исследовании Беннетт еще в 1975 году [7], сильно коррелированные переменные будут влиять на кластеризацию. И произойдет «перевес» кластерного решения (см. также [54]).

Количество переменных кластеризации
Количество группирующих переменных также влияет на минимальный размер выборки исследования (Форман, 1984) [27]. Форман предложил минимальное количество $2^x 2^x$ выборок для набора данных, где xx указывает на число переменных кластеризации.

Например, если набор данных включает 100 наблюдений (например, 100 результатов учащихся), то кластерный анализ может быть выполнен не более чем по 6 переменным (поскольку $2^6 = 64$, и $2^6 = 64$ и $2^7 = 128$) $2^7 = 128$.

Недискриминативные переменные (частотные таблицы)
Последний вопрос выбора переменных – проблема переменных с одинаковым значением для всей выборки. Если переменные не изменяются или изменяются несильно, такую переменную не следует рассматривать полезным дискриминатором (подробнее в [7], [54]).

Поэтому, прежде чем выбирать переменную для кластеризации, следует составить частотную таблицу и проверить дискриминативность переменных.

3.2. Принятие решения о способе кластеризации

На втором шаге кластерного анализа следует принять решение о способе кластеризации в зависимости от цели проведения исследования и задачи кластеризации (например, найти выбросы или создание кластера с почти одинаковыми результатами наблюдений).

Существуют две основные процедуры проведения кластерного анализа: с жестким (однозначным) делением, в котором каждое наблюдение может принадлежать только одному кластеру, и нечетким (многозначным) делением, при котором одно наблюдение может относиться к разным кластерам (Хоссейни, 2013) [30].

Кроме того, следует сделать выбор между иерархическими и неиерархическими методами проведения кластерного анализа, учитывая их специфические особенности.

У неиерархических методов достоинством является низкая зависимость от шумов и выбросов, от выбора метрики, допускается включение незначимых переменных в набор параметров кластеризации. Но требуется заранее определить количество кластеров и правило остановки итерационного процесса (и другие значимые для алгоритмов параметры).

Если количество кластеров неизвестно, следует использовать иерархические алгоритмы, которые строят полное дерево вложенных кластеров. Данные алгоритмы наглядны и позволяют получить детальное представление о структуре данных. Иерархические алгоритмы позволяют идентифицировать выбросы, что улучшает датасеты за счет исключения «случайностей» и проведения второго шага кластеризации без выбросов. Однако данные методы имеют значимые ограничения: невозможность роботы с большими объемами данных (преодолевается за счет работы с репрезентативной выборкой вместо реального объема), необходимость тщательного подбора метрик («мер близости»), негибкость полученных классификаций.

Существует довольно широкий спектр комбинированных алгоритмов. Например, очищенный после первого шага идентификации количества кластеров и вычистки от выбросов датасет можно обработать методами неиерархической кластеризации. Или можно провести ряд экспериментов с различным количеством кластеров, начиная разбиение с совокупности из двух кластеров и постепенно увеличивая их количество, считая при этом ошибку распределения в кластере, достигая необходимой гибкости кластеризации.

Отметим, что кластерный анализ для исследований в области образования в большинстве случаев рассматривает жесткую кластеризацию (например, Боронико и Чокси, 2012 [8]; Юксельтурк и Топ, 2013 [77]).

Жесткая кластеризация чаще всего использует одну из трех процедур:

- иерархическая кластеризация;
- группировка;
- кластеризация на основе плотности.

Другие методы можно посмотреть, например, у Эверитта, Ландау, Лессе и Стал, 2011 [22]; Кауфмана и Россью, 2005 [39]. Они же описывают алгоритм, подходящий для поиска данных с выбросами.

Вы можете подробнее познакомиться с алгоритмами кластеризации в приложениях 1 и 5 или в указанных изданиях.

3.3. Выбор метрики (меры близости/сходства/различия)

На третьем шаге кластерного анализа выбираются меры сходства или различия для определения близости двух объектов[3].

В соответствии с имеющимися переменными кластеризации (например, категориальными, порядковыми или непрерывными) в литературе предлагаются различные меры сходства/различия (Эверит, Ландау и др., 2011 [22]; Кауфман и Россью, 2005 [39]).

Данные тестирований и анкет, чаще всего используемых в образовании, являются непрерывными. Для их исследования используются такие меры близости, как евклидово расстояние, «городские кварталы», расстояние Минковского, Канберра, корреляции Пирсона, углового разделения). Наиболее популярным является евклидово расстояние. В той или иной степени оно используется во всех ранее упомянутых работах.

Формула для вычисления евклидова расстояния для объекта с n-измерениями: $d(P,Q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$, где $P = (p_1, p_2, \dots, p_n)$, в свою очередь $Q = (q_1, q_2, \dots, q_n)$.

Это самая распространенная мера, в двумерном пространстве сводящаяся к расчету гипотенузы.

Подробно о других метриках (мерах сходства/различия) можно прочитать в приложении 2.

3.4. Выбор алгоритма кластеризации

Алгоритмы кластеризации развиваются очень быстро, адаптируются к решению задач анализа различных данных – от изображений и программирования натуральных языков (*NLP – natural language programming*, англ.) до кластеризации валютных коридоров.

Среди различных алгоритмов иерархической кластеризации довольно часто в исследованиях в области образования применяется алгоритм одиночной связи (метод ближайшего соседа) (Флорек и др., 1951 [26]) из-за его «умения» обнаруживать выбросы (Моуи и Сарстед, 2011 [48]).

В алгоритме одиночной связи расстояние между двумя кластерами определяется как минимальное расстояние между любыми двумя наблюдениями в двух кластерах, как показано на рис. 4.

[3] «Два объекта близки, когда их различие мало или сходство велико» (Эверит, Ландау и др., 2011 [22]).

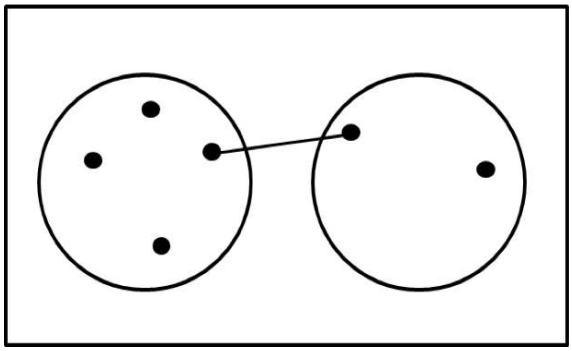

Рис. 4. Метод «одиночной связи» для расчета расстояния между кластерами

Другой популярный метод – метод *k*-средних. Алгоритм *k*-средних строит *k* кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм *k*-средних, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа *k* может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции, или можно провести процесс итераций в соответствии со следующим шагом (параграф 3.5).

Эти и другие методы подробно изложены в приложении 5.

3.5. Подбор количества кластеров

Важным вопросом, решаемым в процедуре кластерного анализа, является подбор количества кластеров: большое количество трудно интерпретировать, а небольшое снижает объем и ценность полученных результатов.

Зачастую в исследованиях заранее выбирается небольшое управляемое количество кластеров, и оно колеблется от 2 до 7 [54, с. 42].

Вместе с тем существует «метод локтя» ([46], [72]) для «графика агломерации», указывающего на накопленную ошибку каждого этапа кластеризации. Метод позволяет определить оптимальное количество кластеров.

Подробно о методе в приложении 3.

3.6. Проверка и интерпретация кластерного решения

В литературе предлагается несколько способов проверки и интерпретации окончательного кластерного решения (например, [22], [39], [48], [54] и др.). Однако для целей поиска возможностей развития системы образования и подготовки адресных рекомендаций не все методы полезны.

Например, Пастор [54] предлагает исследовать различные процедуры и алгоритмы для проверки устойчивости полученного итогового кластерного решения. Следует воспользоваться данной идеей, поскольку в стране накоплены результаты школ по ОГЭ и ВПР начиная с 2015 года, а результаты ЕГЭ – с 2010 года.

Алгоритм одинарной связи хорошо справляется с распознаванием выбросов, что подробно описано у Моуи и Сарстеда [48].

Для распознавания резильентности учащихся и школ отлично подходит методика PISA/ФИОКО, описанная в методических материалах [78] и отчете о PISA [63]. При этом следует отметить, что проверка резильентности затруднена в связи с отсутствием необходимых контекстных данных, собрать которые за прошедшие годы не представляется возможным, в связи с чем сочетание трех последних алгоритмов невозможно.

Перечисленные выше исследования предлагают проверять кластеры при помощи следующих процедур:

1) разделить набор данных на две части и провести кластерный анализ независимо на каждой части, чтобы оценить стабильность решения кластерного анализа [54];
2) изменить порядок данных в наборе данных и повторно запустить кластерный анализ [48].

Затем следует найти наблюдения, расположенные в отдалении от основного кластера и выявить его отличительные характеристики. Кроме того, если в наборе данных имеется какая-либо зависимая переменная, связанная с вопросом исследования, исследователь должен выяснить, существует ли связь между зависимыми переменными и переменными кластеризации, для чего следует изучить литературные источники.

Заключение

В трех главах рассмотрены алгоритмы, методы, способы применения и возможные цели кластерного анализа.

В главе 1 рассмотрены современные дефиниции кластерного анализа, его предмет, объект. Подробно рассказано о месте кластерного анализа в региональных системах оценки качества образования.

В главе 2 подробно разобраны исходные данные для кластерного анализа, метод их сбора и обработки в Свердловской области.

Глава 3 посвящена описанию наиболее полного из имеющихся алгоритмов проведения кластерного анализа. В приложениях подробно рассмотрены математические методы, алгоритмы и процедуры, используемые в кластерном анализе; приведен пример из литературы.

На уровне региона данные алгоритмы будут реализовываться на данных ВПР, ЕГЭ и ОГЭ за 2010–2022 годы с целью определения устойчивости кластеров, а также для определения реального (эффективного, оптимального) существующего количества кластеров, для которых следует разрабатывать методические рекомендации по итогам оценочных процедур.

Поскольку алгоритмы требуют сбора контекстной информации, применение их к данным прошлых лет будет ограничено существующими в распоряжении региона базами данных.

Приложение 1

Сравнительный анализ иерархических и неиерархических методов кластеризации

Выбор между иерархическими и неиерархическими методами следует осуществлять с учетом их особенностей.

У неиерархических методов достоинством является низкая зависимость от шумов и выбросов, от выбора метрики («меры»), допускается включение незначимых переменных в набор параметров кластеризации. Но требуется заранее определить количество кластеров и правило остановки итерационного процесса, а в некоторых алгоритмах – другие параметры кластеризации.

Если количество кластеров неизвестно, следует использовать иерархические алгоритмы, которые строят полное дерево вложенных кластеров. Данные алгоритмы наглядны и позволяют получить детальное представление о структуре данных. Иерархические алгоритмы позволяют идентифицировать выбросы в наборе данных, что позволяет повысить качество датасетов за счет исключения «случайностей» и проведения второго шага кластеризации без выбросов. Однако данные методы имеют значимые ограничения: невозможность роботы с большими объемами данных (преодолевается за счет работы с репрезентативной выборкой вместо реального объема), необходимость тщательного подбора «мер близости», негибкость полученных классификаций.

Существует довольно широкий спектр комбинированных алгоритмов. Например, очищенный после первого шага идентификации количества кластеров и вычистки от выбросов датасет можно обработать методами неиерархической кластеризации. Или можно провести ряд экспериментов с различным количеством кластеров, начиная разбиение с совокупности из двух кластеров и постепенно увеличивая их количество, считая при этом ошибку распределения в кластере, достигая необходимой гибкости кластеризации.

Критерии качества алгоритмов кластеризации
Данная глава написана в соответствии с исследованием Ганти, Деркa и Рамакришнан [79].

Еще недавно считалось необходимым уместить весь набор данных в оперативную память, это было основным критерием качества кластеризации, но появление сверхбольших баз данных изменило данное положение дел.

Кроме стандартного понятия ресурсоемкости, которая для персональных компьютеров выражается в используемом процессорном времени (либо вычислительных мощностях при наличии системы балансировки нагрузки) и памяти, появилась задача масштабируемости алгоритма.

С другой стороны, существуют качественные свойства, которым должен удовлетворять алгоритм кластеризации: независимость результатов от порядка входных данных, независимость параметров алгоритма от входных данных (универсальность алгоритма).

В последнее время появились интегрированные алгоритмы.

Иерархическая кластеризация
Иерархическая кластеризация делится на две основные группы: агломеративную и дивизимную.

Агломеративные методы AGNES (Agglomerative Nesting)
Эта группа методов начинает с того, что рассматривает каждое наблюдение как отдельный кластер. Затем производятся последовательные объединения похожих элементов с соответствующим уменьшением числа кластеров, и так до тех пор, пока не получится один кластер со всеми наблюдениями.

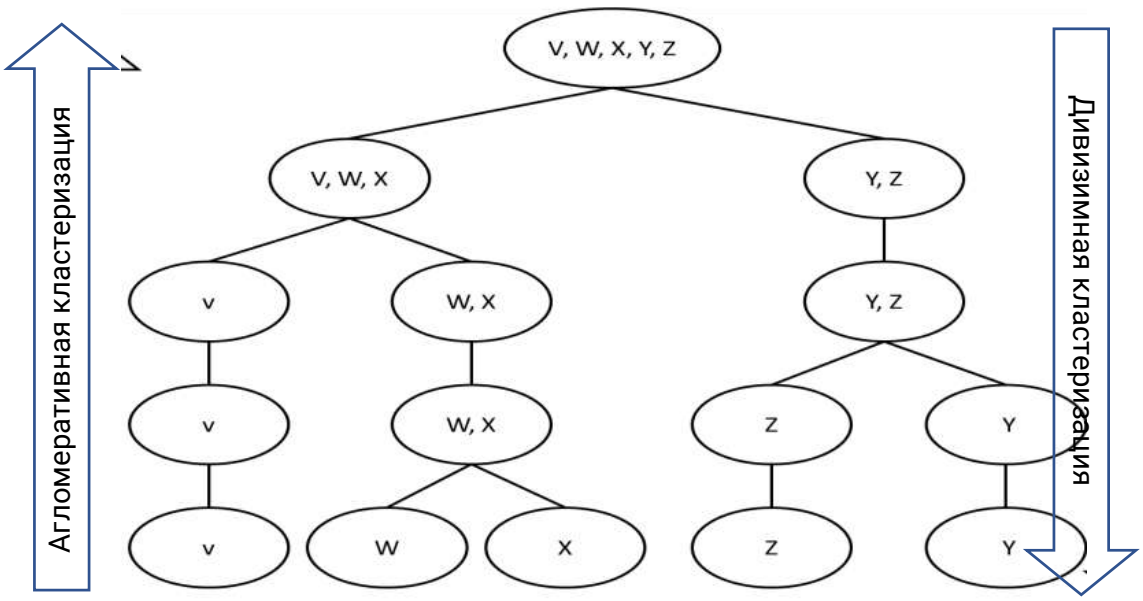

Рис. 5. Агломеративная и дивизимная кластеризации

Довольно часто используемый для нахождения выбросов данных алгоритм одиночной связи «ближайший сосед» (Флорек, Люкасцевик, Перкал, Стейнхаус и Зубрицки, 1951 [26]) относится к агломеративной кластеризации.

Дивизимные методы DIANA (Divisive Analysis)
Дивизимные (разделительные) методы начинаются с того, что все наблюдения объединены в один кластер, затем они рекурсивно разделяются на отдельные,

максимально непохожие кластеры, с соответствующим увеличением числа кластеров на каждом шаге. В результате образуется последовательность расщепленных групп (см. рис. 5).

Процесс завершается, когда каждое наблюдение помещается в отдельный кластер (Эверит и др., 2011 [22]; Кауфман и Россью, 2005 [39]).

Неиерархическая кластеризация
У неиерархических методов повышенная устойчивость к шумам и выбросам, некорректному выбору метрики («меры») и включению незначимых переменных в датасеты (нет необходимости «предварительной чистки данных»).

При этом следует понимать, что сама идея анализа образовательных результатов, например для поиска выбросов, в данном случае неприменима. Кроме того, аналитик должен заранее определить количество кластеров, количество итераций и правило остановки, а также другие параметры кластеризации. Начинающие специалисты не смогут с этим справиться.

Если число кластеров неизвестно, использовать следует иерархические алгоритмы кластеризации.

С другой стороны, если число кластеров и их параметры уже определены на региональном или федеральном уровне, то на муниципальном уровне уже можно исследовать данные объекты неиерархическими методами.

Итеративные методы
При большом количестве наблюдений иерархические методы кластерного анализа практически непригодны. Так, задача расчета данных о кластерах для всего объема результатов Всероссийских проверочных работ в Свердловской области довольно затруднительна (около 1 миллиона результатов).

В таких случаях предпочтительно использовать неиерархические методы, основанные на итеративной сегментации исходной совокупности измерений до момента выполнения «правила остановки».

Процесс неиерархической кластеризации всегда итеративен и содержит несколько значимых элементов (параметров), которые следует определить заранее:

- стартовая точка;
- правило формирования новых кластеров;
- правило остановки.

По сути, существуют две большие группы подходов к неиерархической кластеризации: определение границ кластеров на основе плотности распределения в многомерном пространстве исходных данных, т. е. создание кластера там, где «сгусток точек», и минимизация меры различия между объектами.

Алгоритм k-средних (k-means)
Это наиболее распространенный среди неиерархических методов, также называемый «быстрым кластерным анализом». Требует гипотезы о наиболее вероятном количестве кластеров, поэтому базируется на результатах предыдущих исследований или экспертных оценках.

Алгоритм *k*-средних строит *k* кластеров, средние в которых расположены на возможно больших расстояниях друг от друга.

Для этого на первом шаге выбирается *k* первоначальных центров кластеров (с учетом максимизации расстояния или даже случайным способом). Затем каждый последующий объект присоединяется к тому кластеру, расстояние до которого наименьшее. И далее начинается итерационный процесс, во время которого рассчитываются средние значения показателя в кластере, оно берется за центр кластера, и происходит перераспределение объектов.

Соответственно, итерации можно продолжать, пока не прекратится перераспределение или пока не будет достигнуто какое-то максимальное количество итераций.

Метод можно проиллюстрировать следующим рисунком:

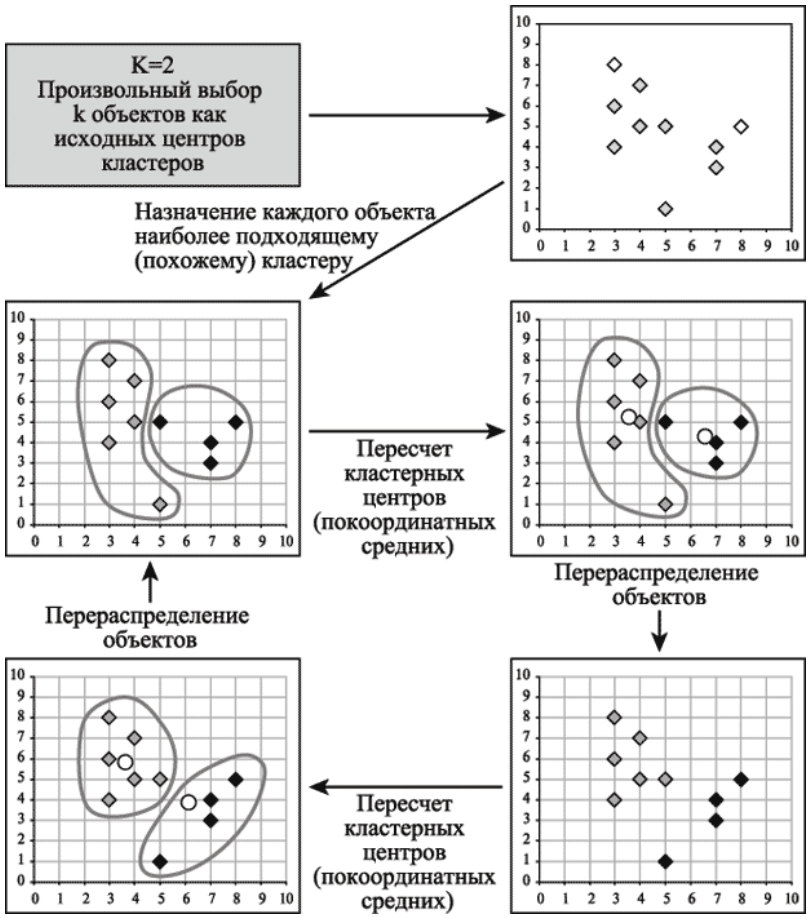

Рис. 6. Пример работы алгоритма *k*-средних (*k* = 2)

Понятно, что для хорошей кластеризации нужно, чтобы значения средних в кластере максимально отличались, а ошибка в кластере была минимальной.

Безусловное достоинство метода – простота (и понятность) алгоритма и быстрота использования. Недостаток заключается в чувствительности к выбросам, которые могут искажать среднее.

Также алгоритм может чересчур долго работать на больших данных, в связи с чем возможно использование репрезентативной выборки.

Приложение 2

Метрики

Формула для вычисления евклидова расстояния для объекта с n-измерениями: $d(P,Q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$, где $P = (p_1, p_2, \dots, p_n)$, в свою очередь $Q = (q_1, q_2, \dots, q_n)$.

Проиллюстрируем, как евклидова мера определяет близость, для чего выберем два наблюдения из набора данных, представленных в табл. 1.

Таблица 1

Два наблюдения из полного датасета

	Отношение к математике	Уровень тревоги	Внимательность	Балл за тестирование	Рабочая память
Наблюдение 1	195	77	107	8	4
Наблюдение 2	187	61	86	7	3

Формула Евклида для указанного датасета, используемого для кластеризации по пяти переменным: $d(P,Q) = \sqrt{\sum_{i=1}^{5}(q_i - p_i)^2}$. Для наблюдений, указанных в таблице 1, формула будет такой:

$$d(P,Q) = \sqrt{(187-195)^2 + (61-77)^2 + (86-107)^2 + (7-8)^2 + (3-4)^2}$$
$$= \sqrt{64 + 256 + 441 + 1 + 1} = \sqrt{763} \approx 27{,}62.$$

Безусловно, все указанные переменные находятся в разных масштабах и диапазонах (см. раздел «Иллюстрация метода одиночной связи (ближайшего соседа) для выборки»), и для них следует предусмотреть преобразование значений до проведения кластерного анализа. Это необходимо, чтобы избежать давления (перевеса) переменных с б**о**льшим масштабом на кластерное решение (смотрите, например, Джонсон и Вичер, 2007 [38], Пастор, 2010 [54]).

Для указанных в таблице 1 переменных кластерный анализ без преобразования имеет сильную зависимость от отношения к математике, внимательности и тревожности. Без преобразования (масштабирования) результат оценки и объем рабочей памяти оказывают незначительное влияние на окончательное кластерное решение.

Среди различных методов, предлагаемых для транспортации, можно выделить такие, как простая *z*-стандартизация и стандартизация по диапазону (например, преобразование в диапазон от 0 до 1 или от -1 до 1) (так, например, поступают в своих исследованиях Моуи и Сарстед, 2011 [48], Пастор, 2010 [54]).

При *z*-стандартизации для каждой переменной выбирают масштаб таким образом, чтобы среднее значение равнялось нулю, а стандартное отклонение равнялось единице.

Такой формат масштабирования широко применялся в исследованиях в области социальных наук (например, Кетчен и Шук, 1996 [40]). Например, для указанных выше выборок из датасета стандартизированная таблица будет выглядеть так, как показано в табл. 2.

Таблица 2

Два наблюдения из полного датасета после стандартизации

	Отношение к математике	**Уровень тревоги**	**Внимательность**	**Балл за тестирование**	**Рабочая память**
Наблюдение 1	1,10	-0,06	1,98	-0.03	-0,15
Наблюдение 2	0,72	-1,06	0,20	-0,31	-1,25

Евклидово расстояние для стандартизированных переменных, вычисленное по той же формуле:

$$d(P,Q) = \sqrt{(0{,}72-1{,}10)^2+\left(-1{,}06-(-0,06)\right)^2+(0{,}20-1{,}98)^2+\left(-0{,}31-(-0{,}03)\right)^2+(1{,}25-(0{,}15))^2}$$
$$= \sqrt{0{,}14+1+3{,}16+0{,}07+1{,}21} = \sqrt{5{,}58} \approx 2{,}36.$$

Сравнивая $\sqrt{0{,}14+1+3{,}16+0{,}07+1{,}21}$ и $\sqrt{64+256+441+1+1}$, видим, что после стандартизации показатель «балл за тестирование» и объем «рабочей памяти» имеют такой же вес, что и другие переменные, для влияния на кластерное решение.

Меры расстояния и сходства между объектами

Исходные данные об объектах можно представить в виде **матрицы измеренных значений признаков**, размером $n \times k$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Аналогичным образом **расстояния между парой векторов** $d(X_i, X_j)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}.$$

Заметим, что диагональные элементы $d_{ii} = 0$ для всех $i = 1{,}2..,n$.

Понятие, противоположное понятию расстояния между X_i и X_j, – **понятие сходства** между объектами (однородности объектов) I_i и I_j.

Пары значений мер сходства можно объединить в матрицу сходства:

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ \vdots & \vdots & \dots & \vdots \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}.$$

При этом величину s_{ij} называют коэффициентом сходства (однородности), а если каждый вектор измерения X_i состоит из нулей и единиц, то величину называют коэффициентом ассоциации или парным коэффициентом сопряженности.

Существует ряд видов коэффициентов ассоциации, значения которых лежат в пределах от -1 до +1. К этой группе принадлежит фи-коэффициент, известный также под названием «четырехпольный коэффициент корреляции». Подробный разбор коэффициентов сходства можно найти в [68].

Таблица 3

Некоторые функции расстояния

Название функции и формула расчета	**Использование**
Евклидово расстояние $d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_{ki} - x_{kj})^2}$	– исходные признаки однородны по физическому смыслу и одинаково важны для классификации; – наблюдения берутся из генеральной совокупности, имеющей многомерное нормальное распределение, т. е. исходные признаки взаимно независимы и имеют близкие значения дисперсий
«Взвешенное» евклидово расстояние $d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} w \cdot (x_{ki} - x_{kj})^2}$	– применяется, когда каждой *k*-й компоненте удается приписать некоторый «вес», пропорциональный степени важности признака в задаче классификации; – определение весов осуществляется экспертным методом, определение по данным выборки может привести к ложным выводам
«Квадрат» евклидова расстояния $d(X_i, X_j) = \sum_{k=1}^{p} (x_{ki} - x_{kj})^2$	– придает больший вес расстояниям между отдаленными объектами
Манхэттенское расстояние (хэммингово расстояние, «сити-блок», l_i-норма, «расстояние городских кварталов») $d(X_i, X_j) = \sum_{k=1}^{p} \lvert x_{ki} - x_{kj} \rvert$	– наиболее простое вычисление; – в большинстве случаев приводит к тем же результатам, что и мера евклидова расстояния; – меньшее влияние отдельных выбросов, чем в евклидовом расстоянии, поскольку нет возведения разницы в квадрат; – если все признаки дихотомические, то расстояние равно количеству несовпадений в признаках объектов *i* и *j*.

Расстояние Чебышева («равномерная метрика», «sup-метрика», «супреум-норма», бокс-метрика, «метрика решетки», «метрика шахматной доски», «метрика хода короля», «8-метрика», l_∞-норма) $d(X_i, X_j) = max(\lfloor x_{ki} - x_{kj} \rfloor)$	– расстояние Чебышева в пространстве следует представлять «шаром»; – расстояние стоит использовать, когда необходимо определить два объекта как «различные», если они отличаются хотя бы по какому-то одному измерению;
Расстояние Минковского («l_p-норма») $d(X_i, X_j) = \sqrt[r]{\sum_{k=1}^{p} \lfloor x_{ki} - x_{kj} \rfloor^r}$	– показатель *r* выбирается от 1 до 4; при 2 получаем евклидово расстояние, при 1 – «расстояние городских кварталов», при ∞ метрика превращается в расстояние Чебышева
Процент несогласия $P = VALUE\lfloor I_i \neq I_j \rfloor$	– используется, если свойства объекта являются категориальными

Менее употребительные расстояния: расстояния Махаланобиса [44], Джеффриса – Матуситы [18], коэффициент дивергенции [15].

Наиболее часто встречаемые меры сходства для интервальных данных и особенности их применения представлены в таблице.

Таблица 4

Некоторые меры сходства для интервальных данных

Название меры сходства и формула расчета	**Использование**
Критерий согласия Пирсона («критерий хи-квадрат, χ^2») $\chi_n^2 = \sum_{i=1}^{n} \frac{(x_{ki} - x_{kj})^2}{x_{kj}}$ В качестве меры сходства используется $\sqrt{\chi_n^2}$.	– наиболее часто употребляемый критерий для проверки гипотез принадлежности выборки некоторому закону распределения; – сравниваются не значения, а частоты выпадения значений; – позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей)
Коэффициент корреляции, или «косинус» $r_{ij} = \frac{\sum_{i=1}^{n} x_{ki} x_{kj}}{\sqrt{\sum_{i=1}^{n} {x_{ki}}^2 \sum_{i=1}^{n} {x_{kj}}^2}}$ Предполагается, что $\sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} x_{kj} = 0$ Если рассматривать X_i и X_j как координаты двух точек в многомерном пространстве, являющихся концами векторов с началом в начале координат, то $r_{ij} = cos(\theta)$, где θ – угол между векторами	Из уравнения для векторов следует, что если объекты I_i и I_j направлены в одну сторону, то угол между ними близок к 0° и r_{ij} близок к 1. Если же векторы отрицательно направлены по отношению друг к другу, то r_{ij} близок к -1. Если вектора не сходны, то угол близок к 0. Несмотря на то, что коэффициент корреляции часто используется для доказательства сходства, он не является функцией сходства в связи с тем, что строго математически s_{ij} не может быть отрицательным числом

Название меры сходства и формула расчета	Использование
Мера «фи-квадрат» $\frac{\sqrt{\chi_n^2}}{\sqrt{\sum_{i=1}^{n} x_{ki} + \sum_{i=1}^{n} x_{kj}}}$	Эта мера представляет собой попытку нормализации меры «хи-квадрат». Для этого она делится на квадратный корень общей суммы частот

Если же объекты состоят из бинарных данных (то есть содержат только 0 и 1, так называемая дихотомическая шкала), то существует довольно большой набор коэффициентов сходства. Для заданных векторов X_i и X_j размерностью *p*:

n_{IJ} – число характеристик, соответствующих единицам в векторах X_i и X_j

n_{ij} – число характеристик, соответствующих нулям в векторах X_i и X_j

n_{iJ} – число характеристик, соответствующих нулю в векторе X_i и единице в X_j

n_{Ij} – число характеристик, соответствующих единице в X_i и нулю в X_j

Таким образом,

$n_J = n_{IJ} + n_{iJ}$ – число единиц в X_j , а $n_j = n_{Ij} + n_{ij}$ – число нулей в X_j .

Таблица 5

Некоторые коэффициенты сходства для бинарных данных

Название функции и формула расчета	Ссылка на описание коэффициента сходства и его использования
$\frac{n_{IJ}}{n_{IJ} + n_{Ij} + n_{iJ}}$	[32], [66]
$\frac{n_{IJ} + n_{iJ}}{p}$	[67]
$\frac{n_{IJ}}{p}$	[59]
$\frac{2n_{IJ}}{2n_{IJ} + n_{Ij} + n_{iJ}}$	[19], [69]
$\frac{2(n_{IJ} + n_{ij})}{p + n_{IJ} + n_{ij}}$	
$\frac{n_{IJ}}{n_{IJ} + 2(n_{Ij} + n_{iJ})}$	
$\frac{n_{IJ} + n_{ij}}{p + n_{Ij} + n_{iJ}}$	[58]

Приложение 3

Подбор количества кластеров

Решение о количестве кластеров является одним из важнейших вопросов кластерного анализа и влияет на знания, полученные в результате его выполнения.

Большое количество кластеров трудно интерпретировать, а небольшое количество кластеров снизит объем и ценность полученных в результате выполнения исследования знаний.

Вместо того, чтобы сосредоточиться на всех возможных решениях, исследователи обычно выбирают небольшое управляемое количество кластеров на основе вопроса исследования и целей кластерного анализа.

В исследовании Пастора указывается, что «всерьез можно рассматривать только решения от 2 до 7 кластеров» [54, с. 42]. О. Мэтт в своих практических исследованиях свободных датасетов [46] приходит к выводам о наличии в них от 2 до 7 реальных кластеров и делится опытом по оптимальному подбору количества кластеров.

При этом О. Мэтт в своих исследованиях производит расчет для «метода локтя» (см. Торндайк, 1953 [72]) при помощи Python, хотя в иностранных источниках чаще встречается метод, основанный на «графике агломерации», построенном в программном обеспечении SPSS.

График агломерации показывает накопленную ошибку, возникающую на каждом этапе кластеризации. При объединении двух разных объектов в один кластер у кластера возрастает «ошибка» описания (грубо говоря, при объединении чисел 1 и 3 в один кластер среднее число в кластере станет 2, а ошибка станет 1). В зависимости от того, какой показатель мы будем брать за расчет «ошибки» кластера, ошибка будет считаться по-разному.

Для демонстрации возьмем описание Дэвида Брина [10] с примером обработки датасета в программе SPSS.

Таблица 6

Пример расчета для графика агломерации

Стадия кластеризации	Количество кластеров	Коэффициент ошибки
83 (-11)		146,118
84 (-10)		190,286
85 (-9)		210,676
86 (-8)		317,640
87 (-7)		342,596
88 (-6)		641,646
89 (-5)		729,812
90 (-4)	Четыре --	815,925
91 (-3)	Три ---	1024,923
92 (-2)	Два --	4126,820
93 (-1)	Все --	5327,993

Полный график агломерации (с учетом всех 93 шагов) выглядит так, как показано на рис. 7.

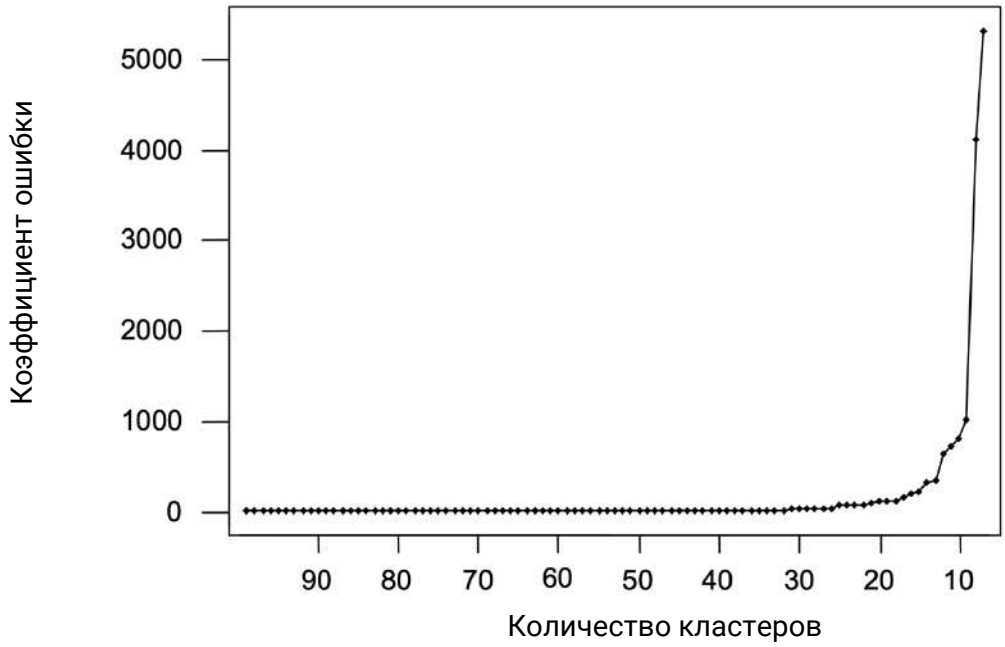

Рис. 7. График агломерации

На рисунке хорошо видно, что на последних 10 шагах происходит резкий скачок коэффициента ошибки. Увеличим его.

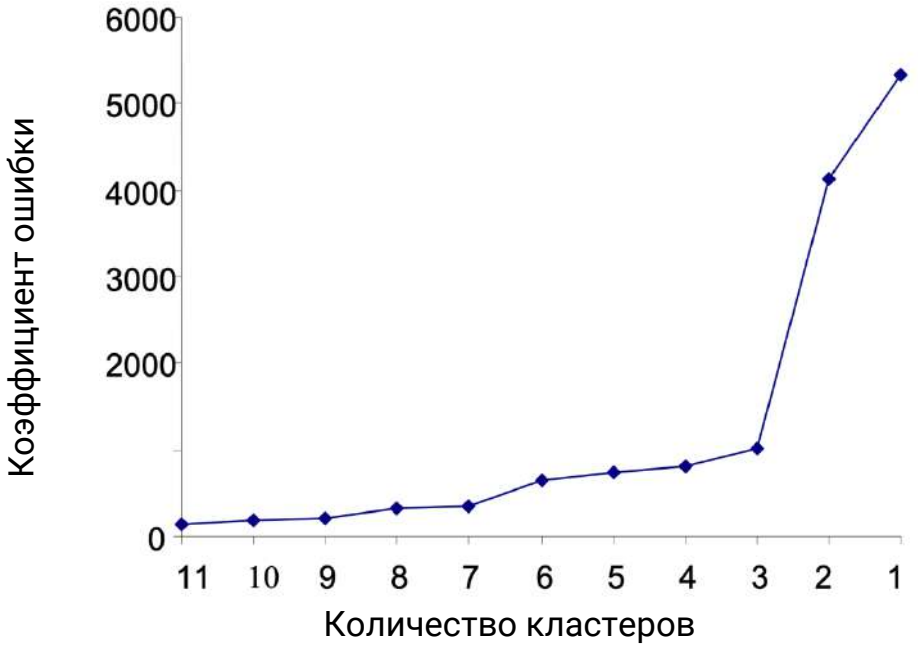

Рис. 8. Детализированный график агломерации

Для указанного набора данных о заболеваниях, рассматриваемых Дэвидом Брином [10], хорошо виден скачок коэффициента ошибки почти вчетверо, когда три кластера уменьшаются до двух.

Следующее же объединение, как мы видим, не сильно влияет на значение накопленной ошибки.

На этом графике хорошо видно, где происходит характерный излом («локоть»). В общем виде шаг объединения, на котором появляется «локоть», вычитается из выборки исследования, а в результате получается количество кластеров. Например, если изгиб произошел на 100-м шаге и в наборе данных было 106 наблюдений, количество кластеров, предлагаемых этим методом, равно шести (106 – 100 = 6).

Приложение 4

Использование метода ближайшего соседа для кластерного анализа влияния психологических параметров на математические результаты обучающихся

Для иллюстрации того, как методом ближайшего соседа (одиночной связи) находить выбросы, используем пример из набора данных о влиянии уровня тревожности, внимания, отношения к предмету, объема рабочей памяти из исследования Хаджибаба, Радмера и Аламоходая [28].

Задача состоит в том, чтобы найти потенциальных кандидатов для исследования неучтенных значимых контекстных факторов и изучения их связи с математическими успехами.

Опишем все шесть шагов для указанного набора данных с целью ознакомления исследователей, заинтересованных в применении кластерного анализа.

Описание выборки
Выборочная группа состояла из 169 учениц 11-х классов средней школы (в возрасте 17–18 лет) из Ирана. Из этой выборки на все вопросы анкет и тестов ответили только 112 учащихся, поэтому в SPSS загружены 112 наблюдений для кластерного анализа.

Шаг 1. Принятие решения о переменных для кластеризации
В исследование включены несколько важных для кластеризации параметров, описанных ранее в литературных источниках.

Таблица 7

Параметры из литературных источников

Параметр	Влияние на математические способности
Внимание	Положительное. Описано у Амани, Аламоходай и Радмер, 2011 [4]
Отношение к математике	Положительное. Описано у Саха, 2007 [61]
Объем рабочей памяти	Положительное. Описано у Рагубар, Барнс и Хечт, 2010 [57]
Оценка за тест групповых встроенных фигур (GEFT)	Положительное. Описано у Мусави, Радмер и Аламоходай, 2012 [49]; Лихи и Залатимо, 1985 [42]; Лис и Пауерс, 1979 [41] и др.
Математическая тревожность	Отрицательное. Описано у Аламоходай, 2009 [3]

Все эти переменные были выбраны для кластеризации и проведения исследования.

Для проверки высокой коллинеарности рассчитана корреляция Пирсона, поскольку переменные непрерывны. Полученные результаты показали, что они имеют умеренную значимую корреляцию друг с другом (от -0,311 до 0,346). Поэтому их можно рассматривать для кластеризации. Более того, поскольку для кластеризации выбрано 5 переменных, то минимальная выборка должна составлять $2^5 = 32$, и это условие выполнено за счет наличия 112 наблюдений в выборке.

Далее было проведено исследование переменных на уровень дискриминативности. Для этого в SPSS были построены частотные таблицы для каждого вопроса в анкете об отношении к математике в модифицированных шкалах отношения Феннемы – Шерман (Докин, Лоуски и Падва, 2004 [20]), внимания (тест на математическое внимание был разработан Хаджибаба и др., 2011 [28]) и опросник тревожности (математическая шкала оценки тревожности, основанная на Фергусон, 1986 [25]) для проверки разнообразия ответов. Эти три анкеты составлены в соответствии с пятью шкалами Лайкерта (подробнее в [92], [99], [75]) и содержали 47, 25 и 32 вопроса соответственно. Что касается математической тревожности и внимания, все ответы (от 1, означающей «очень мало», до 5, означающей «слишком много») были выбраны участниками для каждого вопроса. Однако в первом вопросе анкеты об отношении к математике никто не выбрал «категорически не согласен» или «не согласен», все остальные ответы в анкете были выбраны. Следовательно, эти три анкеты могут выступать в роли дискриминатора.

Что касается GEFT (Олтман, Раскин и Виткин, 1971 [53]), они продемонстрировали разные результаты в тесте и получили баллы от 1 до 20.

Для проверки объема рабочей памяти использовался обратный тест диапазона цифр (Digit Span Backward Test), подробно описанный Рагхубаром и др., 2010 [57]. Его результаты могут быть от 3 до 7, и у учащихся оказался довольно разный объем в этом диапазоне. Следовательно, эти переменные можно рассматривать как дискриминатор.

Шаги 2–4. Процедура кластеризации, выбор меры сходства/несходства и выбор алгоритма кластеризации

Поскольку целью кластерного анализа в исследовании было найти выбросы в выборке, в исследовании использовался метод ближайшего соседа (единственная связь) в рамках иерархической кластеризации.

Более того, в качестве меры различия было выбрано евклидово расстояние, поскольку переменные кластеризации непрерывны.

Шаг 5. Принятие решения о количестве кластеров, проверка и интерпретация кластерного решения

Расчет кластеров методом ближайшего соседа был выполнен на стандартизированных переменных в наборе данных, а в качестве меры различия использовалось евклидовое расстояние.

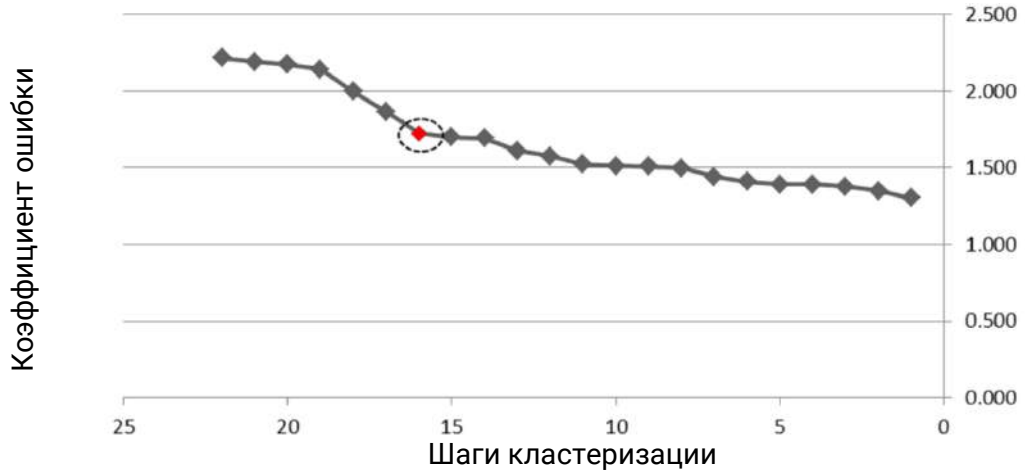

Рис. 9. Линейный график для коэффициентов последних 22 шагов кластеризации одиночной связи

На рисунке представлены коэффициенты для последних 22 шагов кластеризации для того, чтобы визуализировать, где находится «локоть». Как видно на рисунке, «преломление» происходит на 105-м шаге, а в наборе данных было 112 наблюдений. Поэтому количество кластеров по методу «локтя» предлагается равным семи $(112 - 105 = 7)$.

Шаг 6. Результаты

После повторного запуска кластерного анализа для определения количества кластеров количество кластеров для каждого наблюдения было сохранено $k = 7$.

Таблица 8

Среднее значение переменных кластеризации и математической производительности на основе одиночной связи $(k = 7)$

	Количество наблюдений в кластере	Уровень тревоги	Внимательность	Отношение к предмету	Отметка за GEFT	Рабочая память	Математическая успеваемость
Кластер 1	106	77,47	83,05	171,28	8,06	4,03	43,04
Кластер 2	1	40	61	198	14	5	37,50
Кластер 3	1	83	72	183	20	7	40,25
Кластер 4	1	41	99	191	18	7	54,50

Кластер 5	1	69	104	178	10	6	69,50
Кластер 6	1	44	108	178	10	6	70,75
Кластер 7	1	50	114	230	12	5	73
Общая выборка	112	76,24	83,58	172,51	8,43	4,10	43,91

В табл. 8 представлены средние значения переменных кластеризации и математические характеристики наблюдений в каждом кластере. Как видно из этой таблицы, одиночная связь (ближний сосед) образует кластер с большинством наблюдений $(n = 106)$, а выбросы образуют 6 других кластеров.

Эти выбросы являются потенциальными кандидатами на изучение причин отклонений. Приведем лишь некоторые из представленных в работе выводов.

Наблюдение в кластере 2 имеет самый низкий показатель тревожности. И наименьшую математическую успеваемость, наименьшее значение отношения к предмету. При этом объем памяти и балл за GEFT больше, чем среднее по выборке. Для высоких баллов и объема памяти характерно ответственное отношение к математике. Однако ответственное отношение участника к математике было одним из самых низких в выборке.

Подробное изучение этого случая позволит выявить, как эти факторы могут влиять на математическую успеваемость, особенно – роль влияния ответственного отношения к математике на математическую успеваемость.

Мы не будем приводить остальные выводы, хотя они довольно интересны для изучения психологии и педагогики.

Проверка кластерного решения
Для проверки кластерного решения данные переупорядочили по возрастанию на основе оценки GEFT и повторно провели кластерный анализ. Затем создали линейную диаграмму, которая подтвердила местоположение «перелома».

Приложение 5

Описание и сравнение некоторых методов кластеризации

Методы кластеризации развиваются очень быстро, адаптируются к решению задач анализа различных данных – от изображений и программирования натуральных языков (NLP – *natural language programming*, англ.) до кластеризации валютных коридоров.

Мы приведем лишь некоторые из описанных в литературе методов.

Таблица 9

Методы кластеризации

Метод	Достоинства и недостатки
Иерархические агломеративные методы AGNES (Agglomerative Nesting)	
CURE	Основное применение: кластеризация объемных числовых низкоразмерных датасетов. Достоинства: высокий уровень кластеризации, кластеры сложной формы и размеров, линейные требования к памяти и временным затратам в зависимости от размерности датасетов. Недостатки: необходимость определения пороговых значений и количества кластеров, работает только с числовыми данными.
ROCK	
CHAMELEON	
Иерархические дивизимные методы DIANA (Divisive Analysis)	
BIRCH	Достоинства: двухступенчатая кластеризация, кластеризация больших объемов данных, работает на ограниченном объеме памяти, является локальным алгоритмом, может работать при одном сканировании входного набора данных, использует тот факт, что данные неодинаково распределены по пространству, и обрабатывает области с большой плотностью как единый кластер. Недостатки: работа с только числовыми данными, хорошо выделяет только кластеры выпуклой или сферической формы, есть необходимость в задании пороговых значений.
MST	Достоинства: выделяет кластеры произвольной формы, в т. ч. кластеры выпуклой и вогнутой формы, выбирает из нескольких оптимальных решений самое оптимальное. Недостатки: чувствителен к выбросам.
Неиерархические итеративные методы	
Метод *k*-средних (k-means)	Достоинства: простота использования, быстрота использования, понятность и прозрачность алгоритма Недостатки: алгоритм слишком чувствителен к выбросам, которые могут искажать среднее; медленная работа на больших базах данных; необходимо задавать количество кластеров; невозможность применения алгоритма на данных, где имеются пересекающиеся кластеры.

PAM (k-means + k-medoids)	Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма, алгоритм менее чувствителен к выбросам в сравнении с k-means. Недостатки: необходимо задавать количество кластеров; медленная работа на больших базах данных.
CLOPE	Достоинства: высокие масштабируемость и скорость работы, а также качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Он легко рассчитывается и интерпретируется. Во время работы алгоритм хранит в RAM небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. CLOPE автоматически подбирает количество кластеров, причем это регулируется одним-единственным параметром – коэффициентом отталкивания.
LargeItem	

Алгоритм CURE (Clustering Using Representatives)
Основное использование – кластеризация больших числовых низкоразмерных датасетов. К достоинствам относится низкая зависимость от выбросов и умение выделять кластеры сложной формы различных размеров. Требования к памяти и времени линейно зависят от размерности датасета.

Приведем краткое описание алгоритма в соответствии с [70]:

1. Строится дерево кластеров, в котором кластер создается для каждой строки исходного датасета.

2. Для каждого получившегося кластера рассчитывается расстояние до ближайшего кластера. Формируется массив из строк данных и расстояния до ближайшего кластера (по двум ближайшим элементам кластеров) с указанием на ближайший кластер, производится сортировка по возрастанию расстояния до ближайшего кластера.

3. Производится слияние ближайших кластеров в один. Производится расчет расстояния между остальными кластерами и вновь образованным кластером.

Кластеры делятся на две группы: те, у которых ближайшими были точки, вошедшие в новый кластер, и остальные.

Кластеры, у которых расстояние до ближайшего кластера, рассчитанное на шаге 2, больше, чем расстояние до новообразованного кластера, включаются в новообразованный кластер. В противном случае ищется новый ближайший кластер, но при этом не берутся кластеры, расстояния до которых больше, чем до новообразованного кластера. Для кластеров второй группы выполняется следующее: если расстояние до новообразованного кластера ближе, чем предыдущий ближайший кластер, то ближайший кластер меняется. В противном случае ничего не происходит.

4. Если еще не получено предельное необходимое количество кластеров, то вновь выполняется шаг 3.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
В этом алгоритме предусмотрен двухэтапный процесс кластеризации.

Назначение: кластеризация очень больших наборов числовых данных.

Ограничения: работа с только числовыми данными.

Достоинства: двухступенчатая кластеризация, кластеризация больших объемов данных, работает на ограниченном объеме памяти, является локальным алгоритмом, может работать при одном сканировании входного набора данных, использует тот факт, что данные неодинаково распределены по пространству, и обрабатывает области с большой плотностью как единый кластер.

Недостатки: работа с только числовыми данными, хорошо выделяет только кластеры сферической формы, есть необходимость в задании пороговых значений.

Подробное описание алгоритма можно найти в [73].

Фаза 1: загрузка данных в память.

Построение начального кластерного дерева (CF Tree) по данным (первое сканирование набора данных) в памяти.

Подфазы основной фазы происходят быстро, точно, практически нечувствительны к порядку. Алгоритм построения кластерного дерева (CF Tree): кластерный элемент представляет собой тройку чисел (N, LS, SS), где N – количество элементов входных данных, входящих в кластер, LS – сумма элементов входных данных, SS – сумма квадратов элементов входных данных.

Кластерное дерево – это взвешенно сбалансированное дерево с двумя параметрами: B – коэффициент разветвления, T – пороговая величина. Каждый нелистьевой узел дерева имеет не более чем B вхождений узлов следующей формы: $[CF_i, Child_i]$, где $i = 1, 2, ..., B$; $Child_i$ – указатель на i-й дочерний узел.

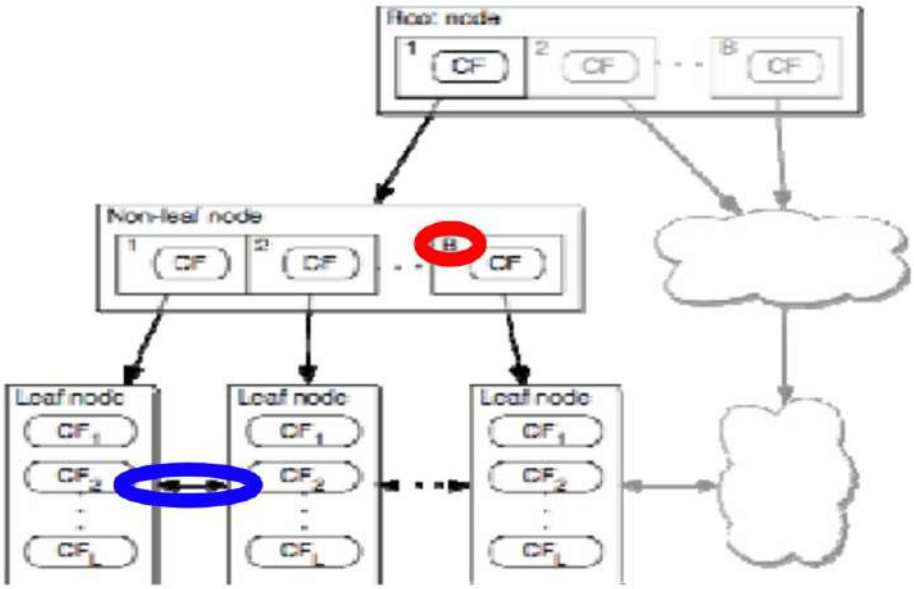

Рис. 10. Построение кластерного дерева

Каждый листьевой узел имеет ссылку на два соседних узла. Кластер, состоящий из элементов листьевого узла, должен удовлетворять следующему условию: диаметр или радиус полученного кластера должен быть не более пороговой величины T.

Фаза 2 (необязательная): сжатие (уплотнение) данных.

Сжатие данных до приемлемых размеров с помощью перестроения и уменьшения кластерного дерева с увеличением пороговой величины T.

Фаза 3: глобальная кластеризация.

Применяется выбранный алгоритм кластеризации на листьевых компонентах кластерного дерева.

Фаза 4 (необязательная): улучшение кластеров.

Использует центры тяжести кластеров, полученные в фазе 3, как основы.

Перераспределяет данные между «близкими» кластерами. Данная фаза гарантирует попадание одинаковых данных в один кластер.

Алгоритм MST (Algorithm based on Minimum Spanning Trees)
Назначение: кластеризация больших наборов произвольных данных.

Достоинства: выделяет кластеры произвольной формы, в т. ч. кластеры выпуклой и вогнутой формы, выбирает из нескольких оптимальных решений самое оптимальное.

Описание алгоритма [17]

Шаг 1. Построение минимального остовного дерева

Связный, неориентированный граф с весами на ребрах G(V, E), в котором V – множество вершин (контактов), а E – множество их возможных попарных соединений (ребер), для каждого ребра (u, v) однозначно определено некоторое вещественное число w(u, v) – его вес (длина или стоимость соединения).

Алгоритм Борувки:

1. Для каждой вершины графа находим ребро с минимальным весом.

2. Добавляем найденные ребра к остовному дереву при условии их безопасности.

3. Находим и добавляем безопасные ребра для несвязанных вершин к остовному дереву.

Общее время работы алгоритма: O(ELogV).

Алгоритм Крускала. Обход ребер по возрастанию весов. При условии безопасности ребра добавляем его к остовному дереву.

Общее время работы алгоритма: O(ELogE).

Алгоритм Прима:

1. Выбор корневой вершины.

2. Начиная с корня добавляем безопасные ребра к остовному дереву.

Общее время работы алгоритма: O(ELogV).

Шаг 2. Разделение на кластеры

Дуги с наибольшими весами разделяют кластеры.

Алгоритм k-средних (k-means)
Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Ограничения: небольшой объем данных.

Достоинства: простота использования, быстрота использования, понятность и прозрачность алгоритма.

Недостатки: алгоритм слишком чувствителен к выбросам, которые могут искажать среднее; медленная работа на больших базах данных; необходимо задавать количество кластеров.

Описание алгоритма [17]

Этап 1. Первоначальное распределение объектов по кластерам

Выбирается число k, и на первом шаге эти точки считаются центрами кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом: выбор k-наблюдений для максимизации начального расстояния; случайный выбор k-наблюдений; выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

Этап 2. Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т. е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т. д., сравнивая полученные результаты.

Алгоритм PAM (Partitioning Around Medoids)
Ограничения: небольшой объем данных.

Достоинства: простота использования, быстрота использования, понятность и прозрачность алгоритма, алгоритм менее чувствителен к выбросам в сравнении с k-means.

Недостатки: необходимо задавать количество кластеров, медленная работа на больших базах данных.

Описание алгоритма [5]

Данный алгоритм берет на вход множество S и число кластеров k, на выходе алгоритм выдает разбиение множества S на k кластеров: S1, S2, ..., Sk. Этот алгоритм аналогичен алгоритму k-средних, только при работе алгоритма перераспределяются объекты относительно медианы кластера, а не его центра.

Алгоритм CLOPE
Назначение: кластеризация огромных наборов категорийных данных.

Достоинства: высокие масштабируемость и скорость работы, а также качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Он легко рассчитывается и интерпретируется. Во время работы алгоритм хранит в RAM небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. CLOPE автоматически подбирает количество кластеров, причем это регулируется одним-единственным параметром – коэффициентом отталкивания.

Описание алгоритма [2], [88]

Пусть имеется база транзакций D, состоящая из множества транзакций $\{t_1,t_2,...,t_n\}$. Каждая транзакция есть набор объектов $\{i_1,...,i_m\}$. Множество кластеров $\{C_1,...,C_k\}$ есть разбиение множества $\{t_1,...,t_n\}$, такое, что $C_1 U ... U Ck=\{t_1,...,t_n\}$ и $C_i <> \emptyset$ и $C_i \cap C_j = \emptyset \ \forall \ i \geq 1, k \geq j$. Каждый элемент C_i называется кластером, а *n, m, k* – количество транзакций, количество объектов в базе транзакций и число кластеров соответственно.

Каждый кластер *C* имеет следующие характеристики:

D(C) – множество уникальных объектов;

Occ(*i,C*) – количество вхождений (частота) объекта *i* в кластер *C*;

$$S(C) = \sum_{i \in D(C)} Occ(i,C) = \sum_{t_i \in C} |t_i|$$

W(C)= D(C),H(C)=S(C)/W(C)

Функция стоимости:

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) * |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} * |C_i|}{\sum_{i=1}^{k} |C_i|}$$

где $|C_i|$ – количество объектов в i-м кластере, k – количество кластеров, r – коэффициент отталкивания ($0 < r \leq 1$).

С помощью параметра r регулируется уровень сходства транзакций внутри кластера и, как следствие, финальное количество кластеров. Этот коэффициент подбирается пользователем. Чем больше r, тем ниже уровень сходства и тем больше кластеров будет сгенерировано.

Формальная постановка задачи кластеризации алгоритмом CLOPE выглядит следующим образом: для заданных D и r найти разбиение C: $\text{Profit}(C, r) \rightarrow \max$.

Самоорганизующиеся карты Кохонена

Назначение: кластеризация многомерных векторов, разведочный анализ данных, обнаружение новых явлений.

Достоинства: используется универсальный апроксиматор – нейронная сеть, обучение сети без учителя, самоорганизация сети, простота реализации, гарантированное получение ответа после прохождения данных по слоям.

Недостатки: работа только с числовыми данными, минимизация размеров сети, необходимо задавать количество кластеров.

Описание алгоритма [89]

Самоорганизующаяся карта Кохонена – нейронная однослойная сеть прямого распространения.

Этап 1. Подготовка данных для обучения

Обучающая выборка должна быть представительной, не должна быть противоречивой; преобразование и кодирование данных, нормализация данных.

Этап 2. Начальная инициализация карты

На этом этапе выбирается количество кластеров и, соответственно, количество нейронов в выходном слое.

Перед обучением карты необходимо проинициализировать весовые коэффициенты нейронов сети. Существуют два способа инициализации весовых коэффициентов:

- инициализация случайными значениями, когда всем весам даются малые случайные величины;
- инициализация примерами, когда в качестве начальных значений задаются значения случайно выбранных примеров из обучающей выборки.

Этап 3. Обучение сети

Карта Кохонена представляет собой нейронную сеть, состоящую из двух слоев: входного и выходного.

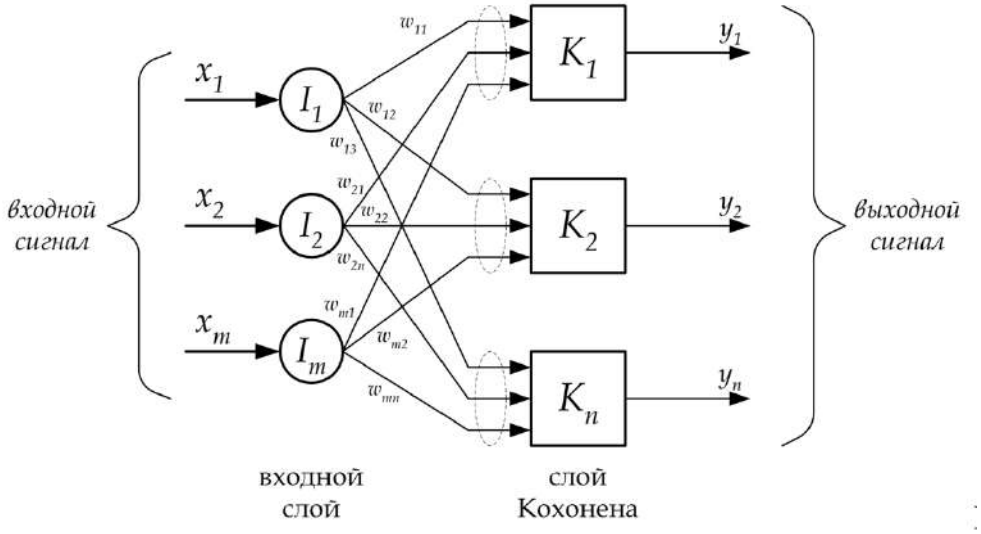

Рис. 11. Карта Кохонена

Обучение состоит из последовательности коррекций векторов, представляющих собой нейроны. На каждом шаге обучения из исходного набора данных случайно выбирается один из векторов, а затем производится поиск наиболее похожего на него вектора коэффициентов нейронов. При этом выбирается нейрон-победитель, который наиболее похож на вектор входов. Под похожестью в данной задаче понимается расстояние между векторами, обычно вычисляемое в евклидовом пространстве. Таким образом, если обозначить нейрон-победитель как *c*, то получим: $\|x - w_c\| = \min_i \{\|x - w_i\|\}$.

После того, как найден нейрон-победитель, производится корректировка весов нейросети. При этом вектор, описывающий нейрон-победитель, и векторы, описывающие его соседей в сетке, перемещаются в направлении входного вектора.

Для модификации весовых коэффициентов используется формула:

$$w_i(t+1) = w_i(t) + h_{ci}(t) * [x(t) - w(t)],$$

где t – номер эпохи, вектор x(t) выбирается случайно из обучающей выборки на итерации t, функция h(t) – функция соседства нейронов.

Функция соседства нейронов представляет собой невозрастающую функцию от времени и расстояния между нейроном-победителем и соседними нейронами в сетке. Эта функция разбивается на две части: функцию расстояния и функцию скорости обучения от времени.

$$h(t) = h(\|r_c - r_i\|, t) * a(t),$$

где r определяет положение нейрона в сетке, a(t) – функция скорости обучения сети.

Функции от расстояния применяются двух видов: простая константа и Гауссова функция.

Простая константа:

$$h(d,t) = \begin{cases} const, d \le \delta(t) \\ 0, d > \delta(t) \end{cases}$$

Гауссова функция:

$$h(d,t) = e^{-\frac{d^2}{2*\delta^2(t)}}$$

Функция скорости обучения сети:

$$a(t) = \frac{A}{t+B},$$

где A и B – константы.

Обучение состоит из двух основных фаз: на первоначальном этапе выбирается достаточно большое значение скорости обучения и радиуса обучения, что позволяет расположить векторы нейронов в соответствии с распределением примеров в выборке, а затем производится точная подстройка весов, когда значения параметров скорости обучения много меньше начальных. В случае использования линейной инициализации первоначальный этап грубой подстройки может быть пропущен.

Этап 4. Использование карты

При использовании карты входной вектор предъявляется на вход, после чего на выходе активизируется нейрон или группа нейронов, которые соответствуют тому или кластеру, полученному в процессе обучения сети.

Алгоритм HCM (Hard C-means)
Назначение: кластеризация больших наборов числовых данных.

Достоинства: легкость реализации, вычислительная простота.

Недостатки: задание количества кластеров, отсутствие гарантии в нахождении оптимального решения.

Описание алгоритма [35]

Шаг 1. Инициализация кластерных центров c_i (i = 1, 2, ..., c). Это можно сделать, выбрав случайным образом c векторов из входного набора.

Шаг 2. Вычисление рядовой матрицы M. Матрица M состоит из элементов m_{ik}:

$$m_{ik} = \begin{cases} 1, \|u_k - c_i\|^2 \le \|u_k - c_j\|^2 \\ 0, остальное \end{cases}$$

для всех $i \neq j$, i= 1, 2,..., c, k= 1, 2, ..., K, где K – количество элементов во входном наборе данных.

Матрица M обладает следующими свойствами:

$$\sum_{i=1}^{c} m_{ij} = 1, \sum_{j=1}^{K} m_{ij} = K$$

Шаг 3. Расчет объектной функции:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, u_k \in C_i} \|u_k - c_i\|^2 \right)$$

На этом шаге происходит остановка и выход из цикла, если полученное значение ниже пороговой величины или полученное значение не сильно отличается от значений, полученных на предыдущих циклах.

Шаг 4. Пересчет кластерных центров

Пересчет кластерных центров выполняется в соответствии со следующим уравнением:

$$c_i = \frac{1}{|Ci|} * \sum_{k, u_k \in C_i} u_k ,$$

где $|C_i|$ – количество элементов в i-м кластере.

Шаг 5. Переход на шаг 2.

Нечеткая кластеризация. Алгоритм Fuzzy C-means

Назначение: кластеризация больших наборов числовых данных.

Достоинства: нечеткость при определении объекта в кластер позволяет определять объекты, которые находятся на границе, в кластеры.

Недостатки: вычислительная сложность, задание количества кластеров, возникает неопределенность с объектами, которые удалены от центров всех кластеров.

Описание алгоритма [35]

Пусть нечеткие кластеры задаются матрицей разбиения:

$$F = [\mu_{ki}], \mu_{ki} \in [0,1], k = \overline{1,M}, i = \overline{1,c},$$

где μ_{ki} – степень принадлежности объекта *k* к кластеру *i*, *c* – количество кластеров, M – количество элементов.

При этом:

$$\sum_{i=1}^{c} \mu_{ki} = 1, k = \overline{1,M}; 0 < \sum_{k=1}^{M} \mu_{ki} < M, i = \overline{1,c}. \quad (1)$$

Этап 1. Установить параметры алгоритма: c – количество кластеров; m – экспоненциальный вес, определяющий нечеткость, размазанность кластеров (m∈[1,∞)); ε – параметр останова алгоритма.

Этап 2. Генерация случайным образом матрицы нечеткого разбиения с учетом условий (1).

Этап 3. Расчет центров кластеров:

$$V_i = \frac{\sum_{k=1}^{M} \mu_{ki}^{\ m} * |X_k|}{\sum_{k=1}^{M} \mu_{ki}^{\ m}}, i = \overline{1,c}$$

Этап 4. Расчет расстояния между объектами X и центрами кластеров:

$$D_{ki} = \sqrt{\|X_k - V_i\|^2}, k = \overline{1,M}, i = \overline{1,c}$$

Этап 5. Пересчет элементов матрицы разбиения с учетом следующих условий:

если $D_{ki} > 0$:

$$\mu_{kj} = \frac{1}{\left(D_{jk}^{\ 2} * \sum_{j=1}^{c} \frac{1}{D_{jk}^{\ 2}} \right)^{1/(m-1)}}, j = \overline{1,c}$$

если $D_{ki} = 0$:

$$\mu_{kj} = \begin{cases} 1, j = i \\ 0, j \neq i \end{cases}, j = \overline{1,c}$$

Этап 6. Проверить условие $\|F - F^*\| < \varepsilon$, где F* – матрица нечеткого разбиения на предыдущей итерации алгоритма. Если «да», то переход к этапу 7, иначе к этапу 3.

Этап 7. Конец.

Библиографический список

1. A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, Clustering for improving educational process mining, in Proc. Fourth Int. Conf. Learn. Anal. Knowl. - LAK 14, 2014, pp. 11-15.
2. A. Salazar, J. Gosalbez, I. Bosch, R. Miralles, and L. Vergara, A 45 case study of knowledge discovery on academic achievement, student desertion and student retention, in Proc. ITRE 2004 2nd Int. Conf. Inf. Technol. Res. Educ., 2004, pp. 150-154.
3. Alamolhodaei, H. (2009). A working memory model applied to mathematical word problem solving. Asia Pacific Education Review, 10(2), 183-192.
4. Amani, A., Alamolhodaei, H., & Radmehr, F. (2011). A gender study on predictive factors of mathematical performance of University students. Educational Research, 2(6), p. 1179–1192.
5. Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. Clustering Algorithms Applied in Educational Data Mining. International Journal of Information and Electronics Engineering, Vol. 5, №2, March 2015
6. B. J. P. Campbell, P. B. Deblois, and D. G. Oblinger, Academic analytics: A new tool for a new era, Educause Review, vol. 42, pp. 40-57, 2007.
7. Bennett, N. (1975). Cluster analysis in educational research: a non-statistical introduction. Research Intelligence, 1(1), 64-70.
8. Boronico, Jess, Choksi Shail S. Identifying Peer Institutions Using Cluster Analysis. - American Journal Of Business Education – May/June 2012, Volume 5, Number 3.
9. Brock, G., Pihur, S., Datta, S., & Datta, S. (2008). ClValid: an R package for cluster validation. Journal of Statistical Software, 25(4), 1-22.
10. Byrne, D.S. - 'Class and Ethnicity in Complex Cities' 1998 Environment and Planning A 30 703-720. Режим доступа: https://www.restore.ac.uk/ qualquanres/papers/Cluster%20Analysis%20example.pdf
11. C. Romero and S. Ventura, Educational data mining: A review of the state of the art, IEEE Transactions on Systems Man and Cybernetics Part C, vol. 40, no. 6, pp. 601-618, 2010.
12. C. Romero, S. Ventura, J. A. Delgado, and P. De Bra, Personalized links recommendation based on data mining in adaptive educational hypermedia systems, Ectel 2007. Lcns 4753, vol. 4753, pp. 292-306, 2007.
13. Chakrapani, C. (2004). Statistics in Market Research. Arnold, London.
14. Chignell, M. H., & Stacey, B. G. (1981). The classification of patients into diagnostic groups using cluster analysis. Journal of Clinical Psychology, 37(1), 151-153.
15. Clark P.J. An extension of the coefficient of divergence for use with multiple characters, Copria 2, 1952. p. 61-64.
16. Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. British Journal of Health Psychology, 10(3), 329-358.
17. Daniel Fasulo. An Analysis Of Recent Work on Clustering Algorithms. June 1999. Режим доступа: https://www.researchgate.net/publication/ 2279437_An_Analysis_of_Recent_Work_on_Clustering_Algorithms
18. Dempster A.P. Elements of Continuous Multivariate Analysis. Reading. Massachusetts: Addison-Wesley Publishing Co., 1969.
19. Dice L.R. Measures of the amount of ecological association between species, Ecology, 26. – 1945, p. 297-302.
20. Doepken, D., Lawsky, E., & Padwa, L. (2004). Modified fenne-ma-sherman attitude scales. Режим доступа: https://woodrow.org/teachers/ math/gender/08scale.html.
21. E. García, C. Romero, S. Ventura, and C. de Castro, A collaborative [37] educational association rule mining tool, Internet High. Educ., vol. 14, no. 2, pp. 77-88, Mar. 2011.

22. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis (5th ed). Wiley Series in Probability and Statistics.
23. F. Radmehr, H. Alamolhodaei. Who Should be Interviewed? A Response from Cluster Analysis, Practice in Clinical Psychology, 2014
24. Faghani Seyedeh Zahra, Y. Jamshid, Nikpour Abolfazl. Geographic pattern of population aging in Mazandaran province during 1986-2011 using hierarchical cluster analysis
25. Ferguson, R.D. (1986). Abstraction anxiety: A factor of mathematics anxiety. Journal for Research in Mathematics Education, 17(145), 145-150.
26. Florek, K., Łukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzy- cki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. Colloquium Mathematicae, 2, №3-4, 282-285.
27. Formann, A. K. (1984). Die Latent-Class-Analyse: Einführung in Theorie und Anwendung. Beltz.
28. Hajibaba, M., Radmehr, F., & Alamolhodaei, H. (2013).A psychological model for mathematical problem solving based on revised Bloom taxonomy for high school girl students. Journal of the Korean Society of Mathematical Education Series D: Research in Mathematical Education. Vol. 17, No. 3, 199–220. Режим доступа: https://www.researchgate.net/publication/264170399_A_Psychological Model_for_Mathematical_Problem_Solving_based_on_Revised_Bloom Taxonomy_for_High_School_Girl_Students
29. Hillhouse, J. J., & Adler, C. M. (1997). Investigating stress effect patterns in hospital staff nurses: results of a cluster analysis. Social Science & Medicine, 45(12), 1781-1788.
30. Hosseini, R. (2013). Topics in data mining (Master's thesis, Ferdowsi University of Mashhad, Mashhad, Iran).
31. Ian H, Witten Eibe, Frank Mark, A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. – 2011.
32. Isaacson E. and Keller H.B. Alanysis of Numerical Methods, John Wiley and Sons, Inc., New York, 1966
33. J. Beck and B. Woolf, High-level student modeling with machine learning, Intell. Tutoring Syst., pp. 584-593, 2000.
34. J. Luan, Data Mining and Knowledge Management in Higher Education, Toronto, Canada, 2002.
35. Jan Jantzen. Neurofuzzy Modelling. Oct. 1998. Режим доступа: http://faculty.petra.ac.id/resmana/private/fuzzy/nfmod.pdf
36. Jiawei Han, Micheline Kamber, Pei Jian. Data Mining: Concepts and Techniques. – 2012.
37. Jiawei Han. Cluster Analysis in Data Mining. Режим доступа: https://www.coursera.org/learn/cluster-analysis/home/welcome
38. Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River. // Open Journal of Statistics, Vol.5 No.7, December 17, 2015. Режим доступа: https://www.scirp.org/pdf/OJS_2015121715550753.pdf
39. Kaufman, L., & Rousseeuw, P. J. (2005). Finding groups in data: An introduction to cluster analysis. Hoboken, NJ: Wiley.
40. Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal, 17(6), 441-458.
41. Lis DJ, Powers JE. Reliability and validity of the Group Embedded Figures Test for a grade school sample. 1979. Режим доступа: https://pubmed.ncbi.nlm.nih.gov/461067/
42. M D Leahy, S D Zalatimo. Group embedded figures test: psychometric data for a sample of high school students. 1985. Режим доступа: https://pubmed.ncbi.nlm.nih.gov/4094865/
43. M. Zorrilla, E. Menasalvas, and D. Marin, Web usage mining project for improving web-based learning sites, in Proc. 10th [39] International Conference on Computer Aided Systems Theory, 2005, pp. 205-210.
44. Mahalanobis P.C. On the generalized distance in statistics. Proc. Natl. Inst. Sci. (India), Vol. 12, 1936, p. 49-55.
45. Maloney, E. A., Risko, E. F., Ansari, D., & Fugelsang, J. (2010). Mathematics anxiety affects counting but not subitizing during visual enumeration. Cognition, 114(2), 293-297.

46. Matt O. 10 Tips for Choosing the Optimal Number of Clusters. Режим доступа: https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92
47. Merceron and Yacef, A web-based tutoring tool with mining facilities to improve learning and teaching, presented at the 11th International Conference on Artificial Intelligence in Education, 2003.
48. Mooi, E., & Sarstedt, M. (2011). A concise guide to market research: The process, data, and methods using IBM SPSS statistics. Springer.
49. Mousavi, S., Radmehr, F., & Alamolhodaei, H. (2012). The role of mathematical homework and prior knowledge on the relationship between students' mathematical performance, cognitive style and working memory capacity. Electronic Journal of Research in Educational Psychology, 10(3), 1223- 48.
50. O. C. Santos and J. G. Boticario, Modeling recommendations for the educational domain, Procedia Comput. Sci., vol. 1, no. 2, pp. 2793-2800, Jan. 2010.
51. O. Zaíane, Building a recommender agent for e-learning systems, in Proc. International Conference on Computers in Education, 2002, pp. 55-59.
52. O. Zaıane and J. Luo, Web usage mining for a better web-based learning environment, presented at Conf. Adv. Technol., 2001.
53. Oltman, P. K., Raskin, E., & Witkin, H. A. (1971). A manual for the embedded figures test. Palo Alto, CA: Consulting Psychologists Press, Inc.
54. Pastor, D. A. (2010). Cluster Analysis. In G. R. Hancock & R. O. Mueller (Eds.). The reviewer's guide to quantitative methods in the social sciences. (pp.41-54). Routledge.
55. R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, Off-task behavior in the cognitive tutor classroom: When students „game the system", 2004.
56. R. S. J. D. Baker and K. Yacef, The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining, vol. 1, no. 1, 2009.
57. Raghubar, K. P., Barnes, M. A., & Hecht, S. A.(2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. Learning and Individual Differences, 20(2), 110–122.
58. Rogers D.J. and Tanimoto T.T. A computer program for classifying plants. Science. Vol. 132. - Oct. 21, 1960, p. 1115 – 1118.
59. Russel P.F. and Rao T.R. On habitat and association of species of anopheline larvae in Southeaster Madras. J. Malor. Inst. India 3. – 1940, p. 153-178.
60. S. Lin, Data mining for student retention management, J. Comput. Sci. Coll., vol. 27, no. 4, pp. 92-99, 2012.
61. Saha, S. (2007). A study of Gender Attitude to Mathematics, Cognitive style and Achievement in mathematics. Experiments in Education, 35, 61–67.
62. Sampogna, F., Sera, F., & Abeni, D. (2004). Measures of clinical severity, quality of life, and psychological distress in patients with psoriasis: a cluster analysis. Journal of Investigative Dermatology, 122(3), 602-607.
63. Scaling procedures and construct validation of context questionnaire data. Режим доступа: https://www.oecd-ilibrary.org/docserver/9789264048096-17-en.pdf
64. Selinski, S. and Ickstadt, K. (2008) Cluster analysis of genetic and epidemiological data in molecular epidemiology. Journal of Toxicology and Environmental Health, 71, 835–844.
65. Shavelson, R. J. (1979). Applications of cluster analysis in educational research: looking for a needle in a haystack. British Educational Research Journal, 5(1), 45-53.
66. Sneath P.H. A. The application of computers to taxonomy. J. General Microbiology, 17. – 1957, p. 201-226.
67. Sokal R.R. and Michener C.D. A statistical method for evaluation systematic relationships, University of Kansas Sci. Bull., - Mar., 20. – 1958, p. 1409-1438.
68. Sokal R.R. and Sneath P.H. A Principles of Numberical Taxonomy, San Francisco: W.H. Freeman and Company, 1963.

69. Sorenson T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application for analyses of the vegetation on Danish commons, Biol., Skr. 5. – 1968, p. 1- 34.
70. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. CURE: An Efficient Clustering Algorithm for Large Databases. 1998. Режим доступа: https://dl.acm.org/doi/abs/10.1145/276305.276312
71. Sutton, M. Q., & Reinhard, K. J. (1995). Cluster analysis of the coprolites from Antelope House: implications for Anasazi diet and cuisine. Journal of Archaeological Science, 22(6), 741-750.
72. Thorndike, R. L. (1953). Who belongs in the family?. Psychometrika, 18(4), 267-276.
73. Tian Zhang, Raghu Ramakrishnan, Miron Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. 1996. Режим доступа: https://dl.acm.org/doi/10.1145/235968.233324
74. Uncertainty Toolkit for Analysts in Government. Режим доступа: https://analystsuncertaintytoolkit.github.io/UncertaintyWeb/index.html
75. Wuensch, Karl L. (October 4, 2005). «What is a Likert Scale? and How Do You Pronounce 'Likert?'». East Carolina University. Retrieved April 30, 2009. Режим доступа: http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm
76. Y. Wang and H.-C. Liao, Data mining for adaptive learning in a [38] TESL-based e-learning system, Expert Syst. Appl., vol. 38, no. 6, pp. 6480-6485, Jun. 2011.
77. Yukselturk, E., & Top, E. (2013). Exploring the link among entry characteristics, participation behaviors and course out- comes of online learners: An examination of learner profile using cluster analysis. British Journal of Educational Technology, 44(5),716–728.
78. Анализ резильентности российских школ. ФИОКО. Режим доступа: https://fioco.ru/Media/Default/Documents/ШНОР/Анализ%20резильентности%20российских%20школ_.pdf
79. Венкатеш Ганти, Йоханнес Герке, Раджу Рамакришнан. Добыча данных в сверхбольших базах данных. Открытые системы, 1999, № 9-10
80. Гроп Д. Методы идентификации систем. – М.: Мир, 1979. – 302 с.
81. Данилов, С. В. Логистика педагогических инноваций на основе кластерного подхода // Диссертация на соискание ученой степени доктора педагогических наук. – 2021. – Ульяновск.
82. Демидов, А. А. Креативные кластеры для Петербурга [Текст] / А. А. Демидов, И. И. Комарова // Современные производительные силы. Теория и практика кластерной политики и науки. – 2014. – № 4. – С. 124–160.
83. Джон Форман. Много цифр: Анализ больших данных при помощи Excel. – М.: Альпина Паблишер, 2016. – 464 с.
84. Дюран Б. и Оделл П. Кластерный анализ. Пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского. Предисловие А. Я. Боярского. – М., «Статистика», 1977. – 128 с.: ил.
85. Идентификация и диагностика систем: учеб. для студ. высш. учеб. заведений / А. А. Алексеев, Ю. А. Кораблев, М. Ю. Шестопалов. – М.: Издательский центр «Академия», 2009. – 352 с.
86. Льюнг Л. Идентификация систем. Теория для пользователя. – М.: Наука, 1991. – 432 с.
87. Методы классической и современной теории автоматического управления: Учебник в 5 Т. 2-е изд., перераб. и доп. Т. 2: Статистическая динамика и идентификация систем автоматического управления / под ред. К. А. Пупкова и Н. Д. Егупова. – М.: Издательство МГТУ им. Н. Э. Баумана, 2004. – 646 с.
88. Паклин Н. Кластеризация категорийных данных: масштабируемый алгоритм CLOPE. Режим доступа: https://loginom.ru/blog/clope
89. Паклин Н. Алгоритмы кластеризации на службе Data Mining. Режим доступа: https://loginom.ru/blog/data-mining-clustering
90. Отбор ШНОР на основе комплексного анализа данных о качестве, ФИОКО. Режим доступа: https://fioco.ru/Media/Default/Documents/ШНОР /Отчет%20по%20комплексному%20анализу%20данных%20о%20ШНОР.pdf
91. Портер, М. Э. Конкуренция [Текст] / М. Э. Портер. – М.: Издательский дом «Вильямс», 2005. – 608 с. – ISBN: 5-8459-0794-2.

92. Практиче Малхотра, Нэреш К. Маркетинговые исследования. Практическое руководство. Пер. с англ. – 3-е изд. – М.: Издательский дом «Вильямс», 2002. – 960 с. – ISBN 5-8459-0357-2.
93. Райбман Н. С. Что такое идентификация? – М.: Наука, 1970. – 118 с.
94. Сейдж Э. П., Мелса Дж. Л. Идентификация систем управления. – М.: Наука, 1974. – 248 с.
95. Сейдж Э. П., Мелса Дж. Л. Теория оценивания и ее применение в связи и управлении. – М.: Связь, 1976. – 496 с.
96. Меньков, А. В., Острейковский, В. А. Теоретические основы автоматизированного управления. Учебник. – М.: Оникс, 2005. – 640 с.: ил.
97. Теория управления / Алексеев А. А., Имаев Д. Х., Кузьмин Н. Н., Яковлев В. Б. – СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 1999. – 435 с.
98. Тим Филлипс. Управление на основе данных. Как интерпретировать цифры и принимать качественные решения в бизнесе. – М.: Манн, Иванов и Фербер, 2017. – 192 с.
99. Толстова Ю. Н. Измерение в социологии. – М., 1998
100. ФИОКО. Методические рекомендации по организации и проведению оценки механизмов управления качеством образования органов местного самоуправления муниципальных районов, городских и муниципальных округов и иных органов, реализующих данные полномочия. Режим доступа: https://fioco.ru/Media/Default/Методики/ Методические%20рекомендации%20по%20организации%20и%20проведению%20МУМ-2021-1.pdf
101. Цыпкин Я. З. Основы информационной теории идентификации. – М.: Наука, 1984. – 320 с.
102. Штейнберг Ш. Е. Идентификация в системах управления. – М.: Энергоатомиздат, 1987. – 80 с.