Science for Education Today

2024. Том 14. № 2 http://sciforedu.ru ISSN 2658-6762

УДК 37.012.4+159.9.072
DOI: 10.15293/2658-6762.2402.06

Научная статья / Research Full Article
Язык статьи: русский / Article language: Russian

Анализ эффективности алгоритмов кластеризации мультимодальных выборок с помощью компьютерного моделирования педагогического эксперимента

Р. Н. Абитов[1], Р. С. Сафин[1]

[1] Казанский государственный архитектурно-строительный университет, Казань, Россия

Проблема и цель. *Статья посвящена проблеме первичной обработки данных педагогических экспериментов, имеющих мультимодальный характер. Целью публикации является выявление наиболее эффективных и универсальных алгоритмов кластеризации данных педагогических экспериментов.*

Методология. *В исследовании использовался метод моделирования педагогического эксперимента. Представлен анализ 5 алгоритмов кластеризации. Оценка эффективности алгоритмов кластеризации проводилась по доле наблюдений с ошибками кластеризации на различных уровнях допустимости и коэффициенту подобия Жаккара. Для оценки влияния параметров моделирования педагогического эксперимента и показателей описательной статистики на эффективность алгоритмов кластеризации использовался регрессионный анализ.*

Результаты. *Дана оценка эффективности различных алгоритмов кластеризации данных, а также проведен корреляционный и регрессионный анализ факторов, влияющих на показатели эффективности кластеризации.*

Наиболее эффективными алгоритмами кластеризации мультимодальных выборок являются алгоритм К-средних и агломеративный иерархический алгоритм.

Заключение. *Результаты, полученные в данной публикации, могут использоваться для статистического анализа данных педагогических, психологических, социологических, биологических и медицинских исследований.*

Ключевые слова: *моделирование педагогического эксперимента; алгоритмы кластеризации данных; мультимодальные выборки; педагогический анализ данных.*

Библиографическая ссылка: Абитов Р. Н., Сафин Р. С. Анализ эффективности алгоритмов кластеризации мультимодальных выборок с помощью компьютерного моделирования педагогического эксперимента // Science for Education Today. – 2024. – Т. 14, № 2. – С. 125–151. DOI: http://dx.doi.org/10.15293/2658-6762.2402.06

Автор для корреспонденции: Р. Н. Абитов, rouslan.abitov@gmail.com

Постановка проблемы

Все новые педагогические технологии обучения, воспитания и развития навыков базируются на проверках гипотез на основе логико-математического аппарата. Многие алгоритмы проведения эксперимента в педагогике заимствованы из экспериментальной психологии, социологии и других социально-гуманитарных наук. Однако если мы попробуем привести прямую аналогию между различными научными дисциплинами, то самым лучшим вариантом окажется доказательная медицина, где, как правило, сравниваются «традиционные» и «новые» терапии и на основе проведения эксперимента и математической обработки данных высказывается суждение о том, что «новая» терапия доказала собственную эффективность. Несмотря на то, что ошибочное подтверждение (ошибка первого рода) или опровержение гипотез (ошибка второго рода) в клинических медицинских исследованиях обходятся дороже для общества, авторы настоящей статьи считают, что педагогические исследования должны приближаться к уровню качества медицинских и биологических исследований. Прежде всего, это касается правильного понимания, какие математические инструменты нужно использовать в определенной ситуации: размер выборок, дизайн эксперимента, насколько сильные/слабые изменения мы хотим измерить, для чего используются различные статистические тесты значимости разницы между выборками.

На основе анализа диссертаций, защищенных в 2023 г., мы можем судить, какие статистические тесты используются для проверки достоверности разницы между контрольной и экспериментальной выборками (табл. 1). В целом методы анализа, используемые для определения достоверности разницы выборок в педагогических исследованиях, можно разделить на 5 групп: тесты Стьюдента/Уэлча; хи-квадрат; тесты Манна – Уитни/Уилкоксона; сочетание различных видов тестов; другие виды, не попадающие ни в одну из вышеперечисленных категорий. Анализировались авторефераты диссертаций по педагогическим научным специальностям, вынесенных на защиту с 01.05.2023 по 31.10.2023.

Таблица 1

Использование тестов достоверности разницы между выборками в диссертациях

Table 1

Using tests of probability between samples in Russian PhD theses

Тип теста	Для каких выборок используется	% вхождений в диссертациях
Тест Стьюдента/Уэлча	Для линейных показателей, подчиняющихся закону нормального распределения	34
Тест хи-квадрат	Для номинативных и ранговых показателей	42
Тест Манна – Уитни /Уилкоксона	Для ранговых показателей	20
Другие тесты	–	4

Таким образом, исходя из результатов анализа защищенных диссертаций, существенную долю экспериментальных педагогических исследований составляют линейные или ранговые квалиметрические показатели,

для которых проверка статистической гипотезы осуществляется тестами Стьюдента либо непараметрическими тестами Манна – Уитни/Уилкоксона. Однако все эти алгоритмы для статистической проверки достоверности работают лишь при определенных «идеальных» условиях постановки эксперимента.

В нашей предыдущей работе [1] приводятся особенности выборок педагогических экспериментов:

– трудность получения параметрической выборки на всех этапах эксперимента;

– частая мультимодальность педагогических выборок;

– малые выборки по сравнению с другими дисциплинами;

– ориентация на оценку качества;

– необходимость анализа всего спектра выборки, а не среднего результата [1].

Авторы настоящей статьи хотят обратить внимание на вторую особенность педагогических выборок: их частую мультимодальность, т. е. когда одна большая выборка состоит из множества более мелких. Подобное явление нередко усложняет проведение педагогического эксперимента. Во-первых, мультимодальная выборка имеет характер непараметрической, и к ней не подходят критерии проверки достоверности разницы Стьюдента/Уэлча. Во-вторых, всячески теряется смысл среднего арифметического и медианы как основных характеристик такой выборки, т. е. получается, что основные выводы, полученные в результате эксперимента, хотя и обобщают результаты совокупной выборки, но не отражают особенности ни одной из подвыборок, из которых эта совокупная выборка состоит. Решением данной проблемы является разделение таких выборок с помощью различных алгоритмов кластеризации данных.

Исходя из вышеизложенного, *целью* публикации является выявление наиболее эффективных и универсальных алгоритмов кластеризации данных педагогических экспериментов. Для достижения поставленной цели исследования потребуется решить ряд *задач*:

– провести анализ предыдущих публикаций по оценке алгоритмов кластеризации данных, а также выявить основные показатели, по которым можно оценить их эффективность;

– смоделировать с помощью алгоритмов генерации данных мультимодальные выборки с максимально возможным охватом комбинаций размера, математического ожидания, среднеквадратичного отклонения и асимметричности распределения подвыборок;

– с помощью различных алгоритмов кластеризации данных определить наиболее оптимальный алгоритм разделения (кластеризации) выборок на подвыборки;

– определить, по каким явным признакам выборки педагог-исследователь может судить о целесообразности кластеризации выборок.

Обзор работ по оценке эффективности алгоритмов кластеризации данных

Сравнению алгоритмов кластеризации с точки зрения их эффективности посвящены исследования T. Kinnunen с соавторами [11], Y. G. Jung с соавторами [18], A. P. Reynolds с соавторами [10], K. M. Bataineh с соавторами[1],

[1] Bataineh K. M., Naji M., Saqer M. A Comparison Study between Various Fuzzy Clustering Algorithms // Jordan Journal of Mechanical & Industrial Engineering. – 2011. – Vol. 5 (4). – P. 335–343. URL: https://jjmie.hu.edu.jo/files/v5n4/v5n4.pdf#page=58

Syakur M. A., Khotimah B. K., Rochman E. M. S., Satoto B. D. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster // IOP Conference Series: Materials Science and Engineering, 2018. vol. 336. P. 012017. DOI: https://10.1088/1757-899X/336/1/012017

G. A.Wilkin и X. Huang [27], К. С. Ершова и Т. Н. Романовой [2], С. Л. Подвального с соавторами [3], и Е. В. Сивоголовко [4]. Во всех представленных работах проверяли эффективность алгоритмов кластеризации либо уже на готовых данных, либо рассматривали частные случаи с одной выборкой или одной повторностью.

Статья D. Xu и Y. Tian [28] уделяет основное внимание подробному обзору алгоритмов кластеризации. Также авторы рассмотрели основные параметры кластеризации, такие как измерение расстояния или сходства, и другие индикаторы. Кроме того, в этой работе описаны основные методы оценки точности алгоритмов кластеризации. Авторы классифицируют алгоритмы следующим образом: алгоритмы разделения (*K-средние*, *CLARA*) [2]; иерархические алгоритмы (*BIRCH*, *CURE*, *ROCK*, *Chameleon*)[3] [15; 16; 19; 29]. В статье также даются подробные рекомендации по применению алгоритмов кластеризации в зависимости от типов данных их размерности, а также классифицируемых категорий [28].

Наиболее близкой к проблематике настоящей публикации является статья “Clustering algorithms: a comparative approach”, посвященная эффективности различных методов кластеризации данных в научных исследованиях [9]. Авторы статьи рассматривают эффективность девяти алгоритмов кластеризации: *алгоритм К-средних*, *CLARA*, *иерархический алгоритм*, *EM*, *HCModel*, *спектральный алгоритм*, *OPTICS*, DBSCAN[4] [5; 7; 21; 24; 26]. Авторы статьи проверяют эффективность алгоритмов кластеризации на сгенерированных случайных двумерных данных по параметрам классов, особенностей, количества объектов, параметров смешения подвыборок. Для оценки качества работы алгоритмов кластеризации они используют коэффициенты подобия: Жаккара, ARI (Adjusted Rand Index), NMI (Normalized Mutual Information), FMI (Fowlkes – Mallows Index)[5] [14; 22]. Количество смоделированных вариантов смешанных сгенерированных выборок равнялось 400. Для оценки эффективности алгоритмов авторы статьи выбрали непараметрический тест Краскела – Уоллиса[6] [20]. В результате совокупного анализа всех алгоритмов в различных условиях авторы пришли к выводу, что в одних случаях наиболее эффективным методом кластеризации является *спектральный алго-*

[2] Steinhaus H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV: 801–804. – cf, 1956.

MacQueen J. Some methods for classification and analysis of multivariate observations. T. 1. – Oakland, CA, USA, 1967.

Partitioning Around Medoids (Program PAM). – Wiley Series in Probability and Statistics, 1990. – P. 68–125. Portico. DOI: http://dx.doi.org/10.1002/9780470316801.ch2

Kaufman L., Rousseeuw P. J. Partitioning Around Medoids (Program PAM) // Finding Groups in Data. – John Wiley & Sons, Inc., 2008. – P. 68–125.

[3] Guha S., Rastogi R., Shim K. ROCK: A robust clustering algorithm for categorical attributes // Information systems. – 2000. – Vol. 25 (5). – P. 345–366.

Guha S., Rastogi R., Shim K. CURE: An efficient clustering algorithm for large databases // ACM Sigmod record. – 1998. – Vol. 27 (2). – P. 73–84.

[4] Ester M., Kriegel H.-P., Sander J., Xiaowei Xu A density-based algorithm for discovering clusters in large spatial databases with noise // KDD–96 Proceedings. – 1996. – P. 226–231. URL: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

Hastie T., Tibshirani R., Friedman J. The EM algorithm // The Elements of Statistical Learning. – 2003. – P. 236–243.

[5] Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines // Bull Soc Vaudoise Sci Nat. – 1901. – Vol. 37. – P. 241–272.

[6] Kruskal W. H., Wallis W. A. Use of ranks in one-criterion variance analysis // Journal of the American statistical Association. – 1952. – Vol. 47 (260). – P. 583–621.

ритм, в других случаях – *иерархический алгоритм, EM и OPTICS*. Тем не менее в этой статье рассматривалась лишь эффективность работы алгоритмов на двумерных данных. Отметим, что в статье не рассматривается ни один из параметров вариативности выборки, кроме ее размера, а также величина стандартного отклонения, асимметричности, расстояния между величинами стандартного отклонения.

Методология исследования

Поскольку авторам статьи необходимо выработать рекомендации для педагога-исследователя в случаях, если есть подозрение, что данные, полученные на начальном (констатирующем) этапе эксперимента, имеют мультимодальную природу, прежде всего следует объяснить алгоритм построения такой выборки и ее параметров, состоящий из 4 этапов (рис. 1):

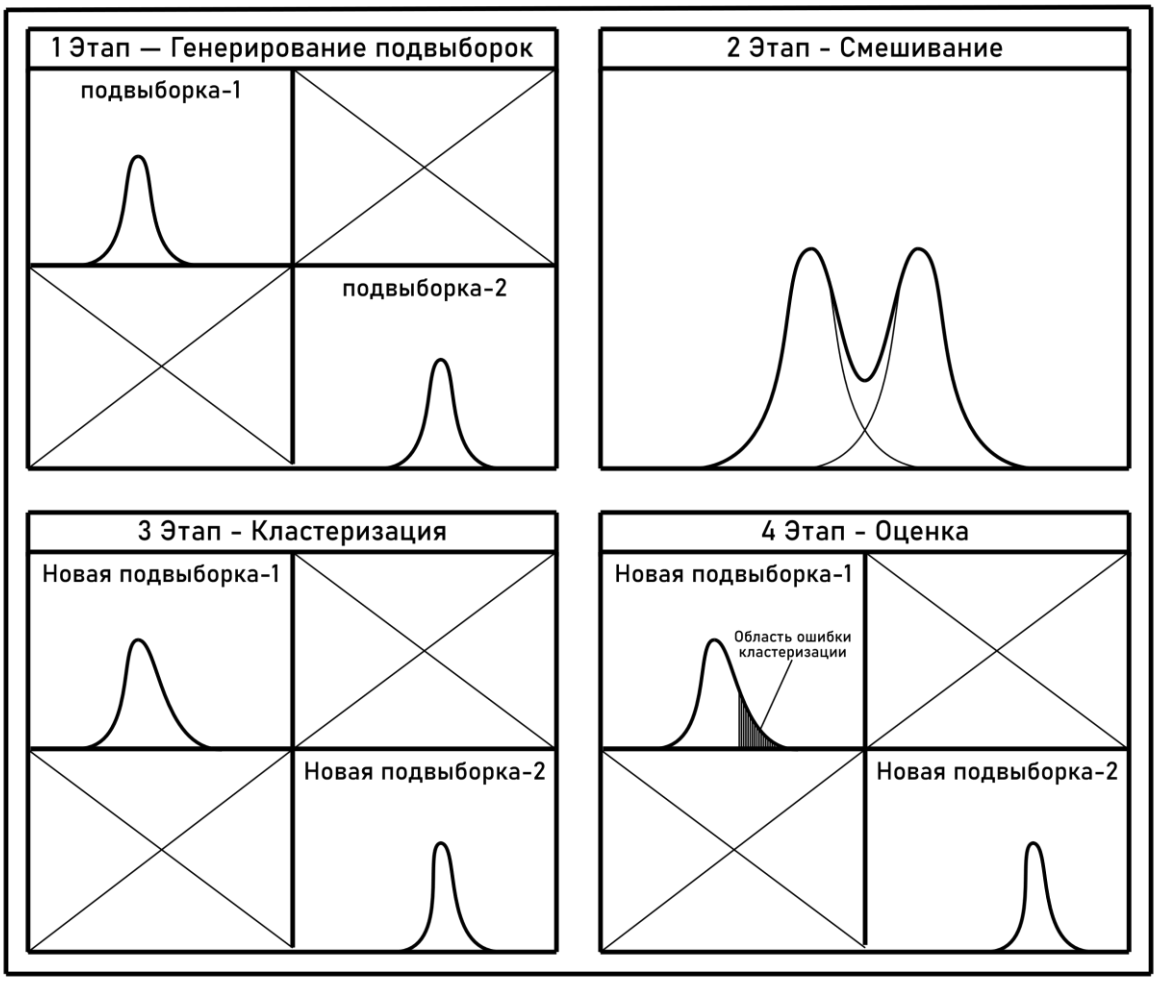

Рис. 1. Основные этапы моделирования педагогического эксперимента

Fig. 1. The main stages of modeling a pedagogical experiment

1. *Генерирование подвыборок*: по заданным параметрам генерируются две малые подвыборки.

2. *Смешивание*: две подвыборки смешиваются в одну большую выборку.

3. *Кластеризация*: сформировавшаяся выборка проверяется посредством алгоритмов кластеризации на принадлежность каждого наблюдения к определенному кластеру (в нашем случае к одному из двух).

4. *Оценка*: делается вывод об эффективности алгоритмов кластеризации на основе ошибочного или правильного отнесения наблюдения в первоначальную подвыборку.

Для всеохватывающего представления различных вариантов смешивания подвыборок были выбраны следующие их параметры (табл. 2).

Таблица 2

Параметры моделирования педагогических экспериментов

Table 2

Parameters of computer modeling of experiments

Параметр выборки	Варианты параметров выборки
Математическое ожидание подвыборки, баллы	От 100 до 145 с шагом 5 (100, 105, 110, 115, 120, 125, 130, 135, 140, 145)
Стандартное отклонение подвыборки, баллы	5, 10, 15
Асимметрия подвыборки, коэффициент-α	-3, -2, -1, 0, 1, 2, 3
Количество наблюдений в подвыборке	100, 200, 300

Рассмотрим каждый из этих параметров подробнее. *Математическое ожидание выборки* представляет собой количественный (линейный) квалиметрический показатель, отражающий определенный уровень компетенции испытуемого. Диапазон выбирался исходя из необходимости избежать генерации отрицательных значений с тем учетом, что расстояние в трех стандартных отклонениях уложится в диапазон от 0 до 100 баллов (самая большая величина стандартного отклонения в нашем случае – 15 баллов), а также показателя асимметричности распределения. Именно поэтому самое низкое значение математического ожидания было выбрано как 100 баллов. При генерации данных показатель математического ожидания изменялся только у первой подвыборки, у второй подвыборки он оставался неизменным (145 баллов), в то время как показатель математического ожидания у первой подвыборки двигался в сторону увеличения с шагом 5 баллов (рис. 2). Величина шага определялась как величина минимального показателя стандартного отклонения подвыборки – 5 баллов. Для величин *стандартного отклонения* были выбраны варианты показателя в 5, 10 и 15 баллов.

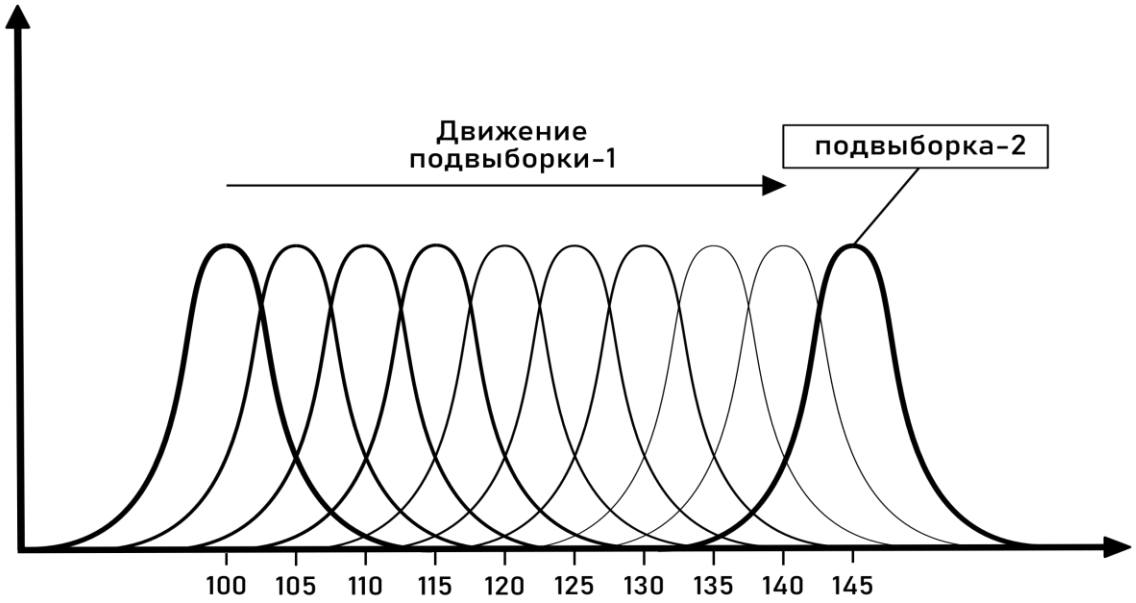

Рис. 2. Схема изменения математического ожидания в процессе моделирования педагогического эксперимента

Fig. 2. Scheme of expected value movement in the process of modeling

Также была учтена вероятность несимметричного распределения выборки. Для генерации подобных распределений было выбрано несимметричное нормальное распределение [6]. В качестве *параметра асимметричности* этого распределения выступает параметр α. Диапазон показателя асимметричности был выбран от -3 до 3 с шагом 1 (рис. 3; распределение положительных показателей α зеркально симметричны отрицательным). Нулевой показатель-α дает симметричное гауссово распределение. *Количество наблюдений* (100, 200, 300) выбиралось исходя из типичных размеров выборок в педагогических исследованиях. При моделировании использовались две повторности.

При генерации отбрасывались симметричные варианты сочетания подвыборок: например, если у первой подвыборки число вхождений равнялось 100 человек, а у второй – 300, то симметричный вариант (300 человек в первой подвыборке, 100 – во второй) исключался. Тот же принцип применялся ко всем другим параметрам подвыборок (стандартное отклонение, коэффициент-α). Совокупное количество сформированных выборок – 20 160. Размер каждой выборки варьировался от 200 до 600 наблюдений.

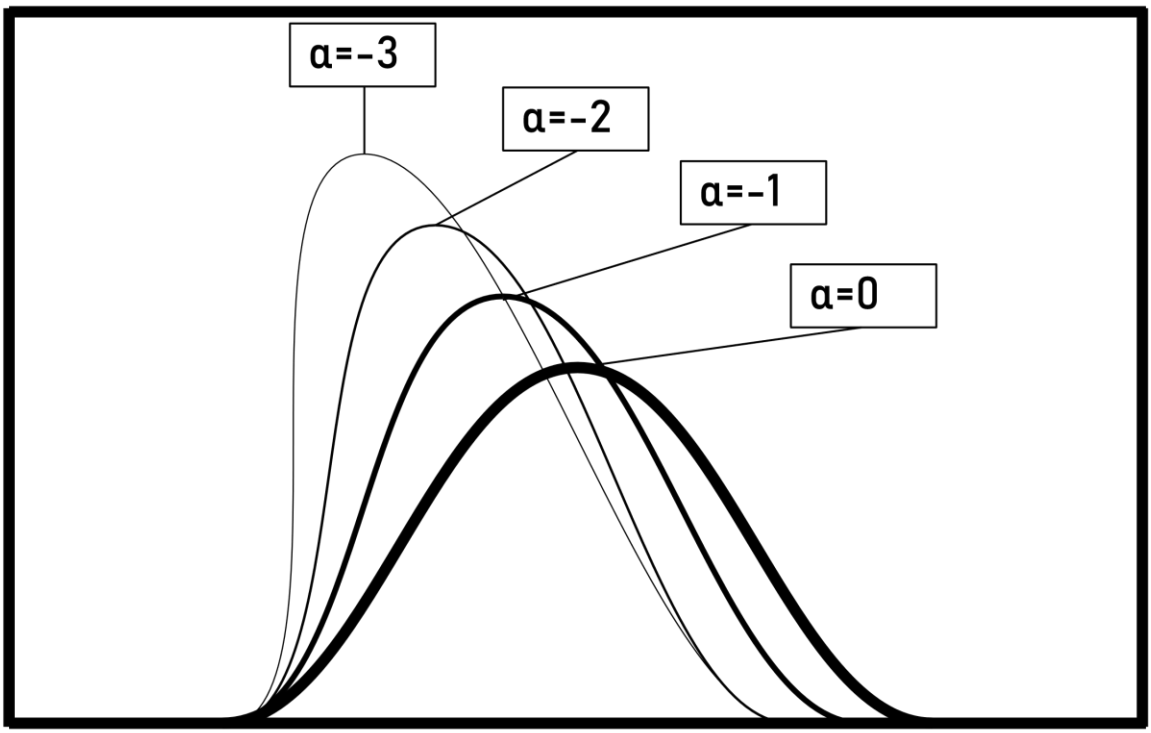

Рис. 3. Визуальное представление коэффициента-α ассиметричного нормального распределения

Fig. 3. Visual representation of α-coefficient of asymmetric normal distribution

Генерация данных производилась с помощью авторского кода на языке программирования Python[7]. В качестве алгоритмов кластеризации были выбраны: *алгоритм К-средних*, *алгоритм К-средних с минимальными выборками* [23], *агломеративный иерархический алгоритм*, *алгоритм BIRCH*, *спектральный алгоритм*. Для генерирования выборок использовались научные математические библиотеки языка Python – NumPy и ScikitLearn. Для регрессионного и корреляционного анализа полученных результатов использовалось программное обеспечение Microsoft Excel, а также пакет статистической обработки данных на основе языка программирования R – RStudio.

Результаты исследования

Для первичной оценки алгоритмов авторы данной статьи взяли две отсечки допустимости ошибок кластеризации 5 % и 32 % (величины пересечения плотностей двух нормальных распределений на расстоянии двух и одного стандартных отклонений соответственно). В таблице 3 и на рисунке 4 показано

[7] Авторский код на языке программирования Python доступен по ссылке: https://disk.yandex.ru/d/1H0cX6UDsNYhRA

число случаев, когда процент отнесения к «чужому кластеру» не превышает 5 % и 32 % по количеству выборок и проценту от общего числа выборок (20 160 наблюдений).

Таблица 3

Доля наблюдений с ошибками кластеризации на различных уровнях допустимости

Table 3

Proportions of observations with clustering errors at different tolerance levels

Алгоритмы кластеризации	Менее 5 % ошибок, число наблюдений / %	Менее 32 % ошибок, число наблюдений / %
К-средние	7358 / 36,5	15675 / 77,8
К-средние с мин. выб.	7356 / 36,5	15657 / 77,7
Иерархический	8007 / 39,7	15164 / 75,2
BIRCH	6036 / 29,9	13986 / 69,4
Спектральный	5647 / 28	8995 / 44,6

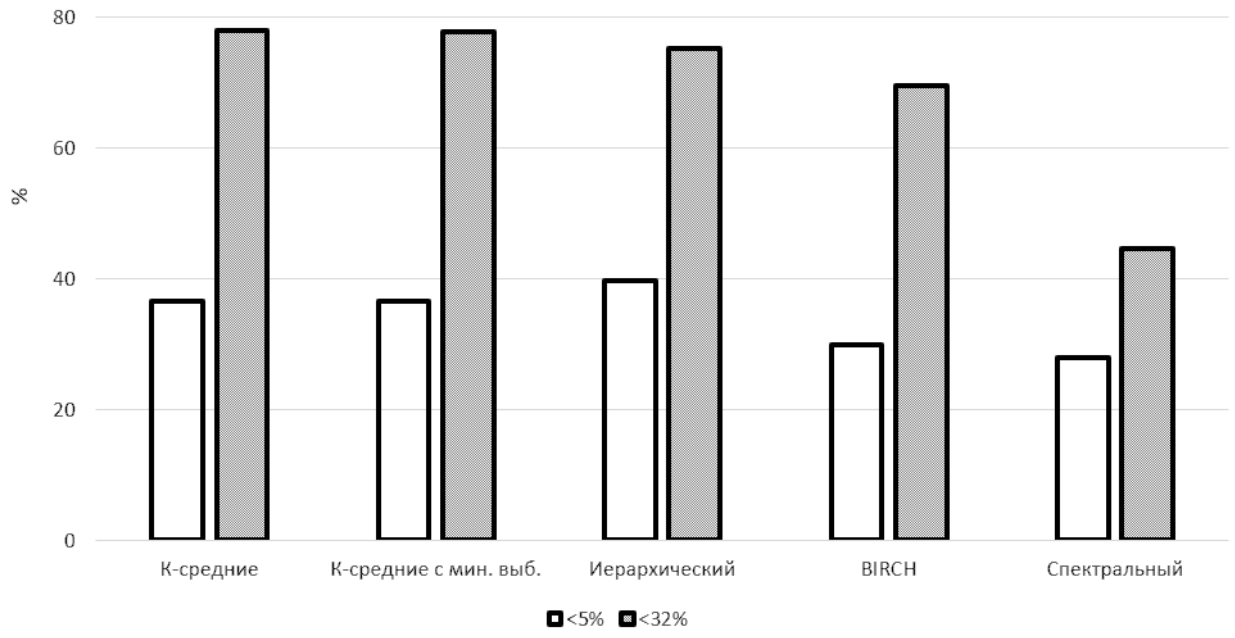

Рис. 4. Диаграмма доли наблюдений с ошибками кластеризации на различных уровнях допустимости

Fig. 4. Plot of the proportion of observations with clustering errors at different tolerance levels

Наилучший результат на отсечке 5 % ошибок показывает *иерархический алгоритм* кластеризации с результатом 39,7 %. За ним практически с идентичным результатом идут оба алгоритма *К-средних* (36,5 %). *BIRCH* и *спектральный алгоритм* показывают близкие результаты: 29,9 % и 28 % соответственно.

На отсечке 32 % ошибок наилучший результат показывают два варианта алгоритма *К-средних* с почти идентичными показателями (77,8 % – классический алгоритм, 77,7 % – алгоритм с малыми выборками). Чуть худший результат показал *Иерархический алгоритм* (75,2 %) и *алгоритм BIRCH* (69 %). *Спектральный алгоритм* показал неудовлетворительный результат – 44,6 %.

Для более широкого представления эффективности алгоритмов кластеризации в таблице 4 и на рисунке 5 представлено распределение процента ошибок кластеризации.

Таблица 4

Показатели описательной статистики доли ошибок кластеризации

Table 4

Indicators of descriptive statistics of clustering errors percentage

	К-средние, % ошибок	**К-средние с мин. выб., % ошибок**	**Иерархический, % ошибок**	**BIRCH, % ошибок**	**Спектральный, % ошибок**
Среднее арифм.	18,5	18,5	18,7	22,1	32,8
Первый квартиль	1,7	1,7	0,8	2,25	3
Медиана	11	10,8	10,75	17,75	39,6
Третий квартиль	29	29	31,6	36,5	49,8

Представленные данные в целом коррелируют с предыдущими выводами об эффективности алгоритмов кластеризации. Показатели среднего арифметического у несимметричных распределений следует рассматривать в комплексе с другими показателями. В частности, у всех алгоритмов, кроме *спектрального*, показатель среднего арифметического выше медианы. Это значит, что асимметрия распределения отклонена в сторону меньшего процента ошибок кластеризации. На рисунке 5 также видно, что *спектральный алгоритм* показывает наихудшие результаты по всем показателям (медиана, квартили, максимум и минимум без учета статистических выбросов). Немного лучшие результаты демонстрирует *алгоритм BIRCH*, при этом, несмотря на довольно высокий разброс значений ошибок кластеризации в последней четверти, алгоритм довольно неплохо себя показывает на оставшихся трех четвертях – с результатом 36,5 % по третьему квартилю. Наилучшие результаты показали оба алгоритма *К-средних* и *иерархический алгоритм*. При этом *иерархический алгоритм* лучше всего показывает себя на отсечке первой четверти (первый квартиль) и половины (медиана) выборки, а оба алгоритма *К-средних* – на отсечке трех четвертей.

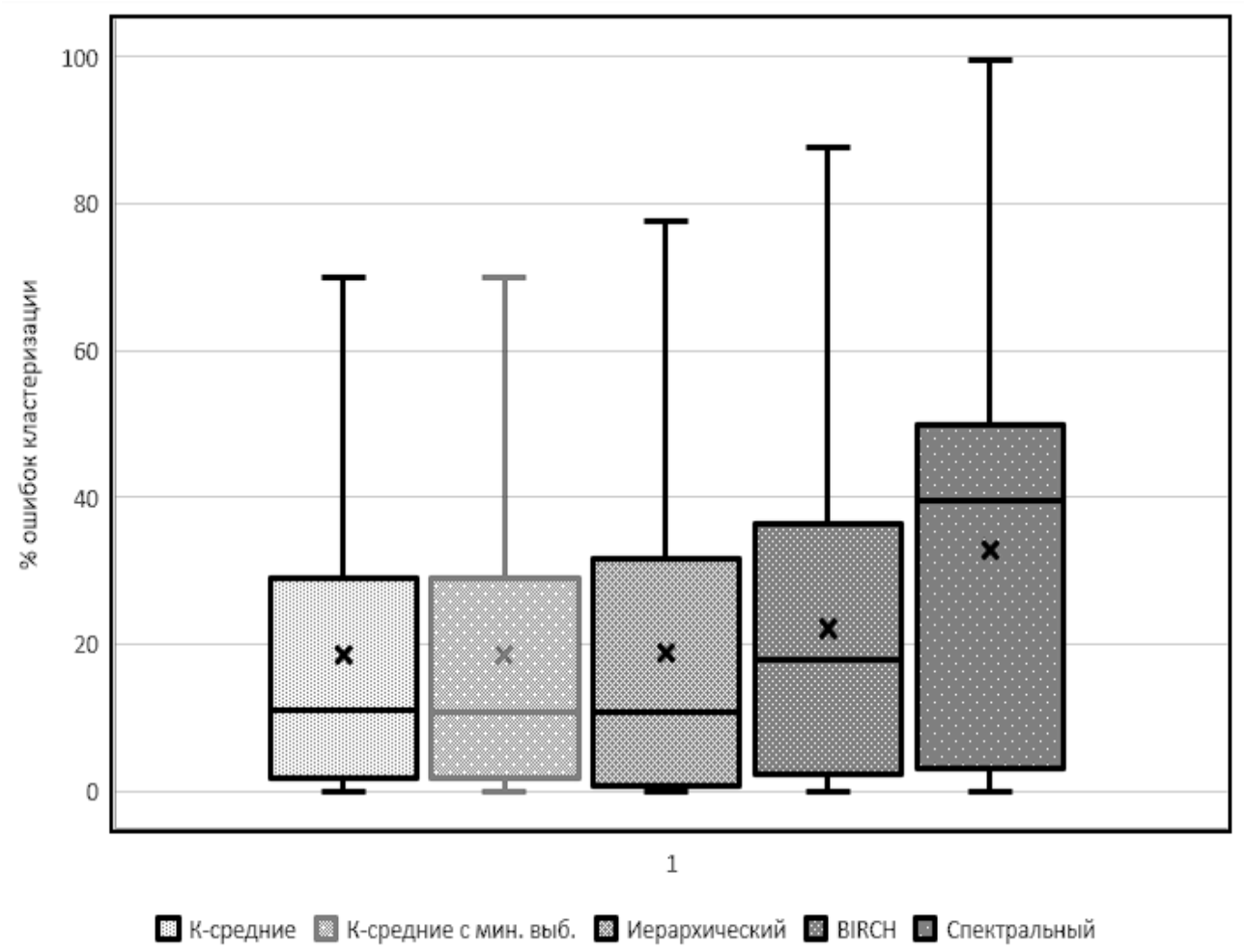

Рис. 5. Диаграммы размаха доли ошибок кластеризации

Fig. 5. Boxplots of clustering error rate

В публикации M. Z. Rodriguez с соавторами для оценки эффективности алгоритмов кластеризации использовались различные коэффициенты подобия [9]. Для похожей оценки авторами данной статьи был использован *коэффициент Жаккара* (коэффициент, равный нулю, говорит о полном несовпадении; коэффициент, равный единице, соответствует полной идентичности). Результаты обработки данных представлены в таблице 5 и на рисунке 6.

Таблица 5

Показатели описательной статистики коэффициента подобия Жаккара

Table 5

Indicators of descriptive statistics of the Jaccard similarity coefficient

	К-средние, коэфф J	К-средние с мин. выб., коэфф J	Иерархический, коэфф J	BIRCH, коэфф J	Спектральный, коэфф J
Среднее арифм.	0,75	0,75	0,74	0,68	0,5
Первый квартиль	0,56	0,56	0,53	0,48	0,01
Медиана	0,82	0,82	0,82	0,72	0,53
Третий квартиль	0,97	0,97	0,99	0,96	0,95

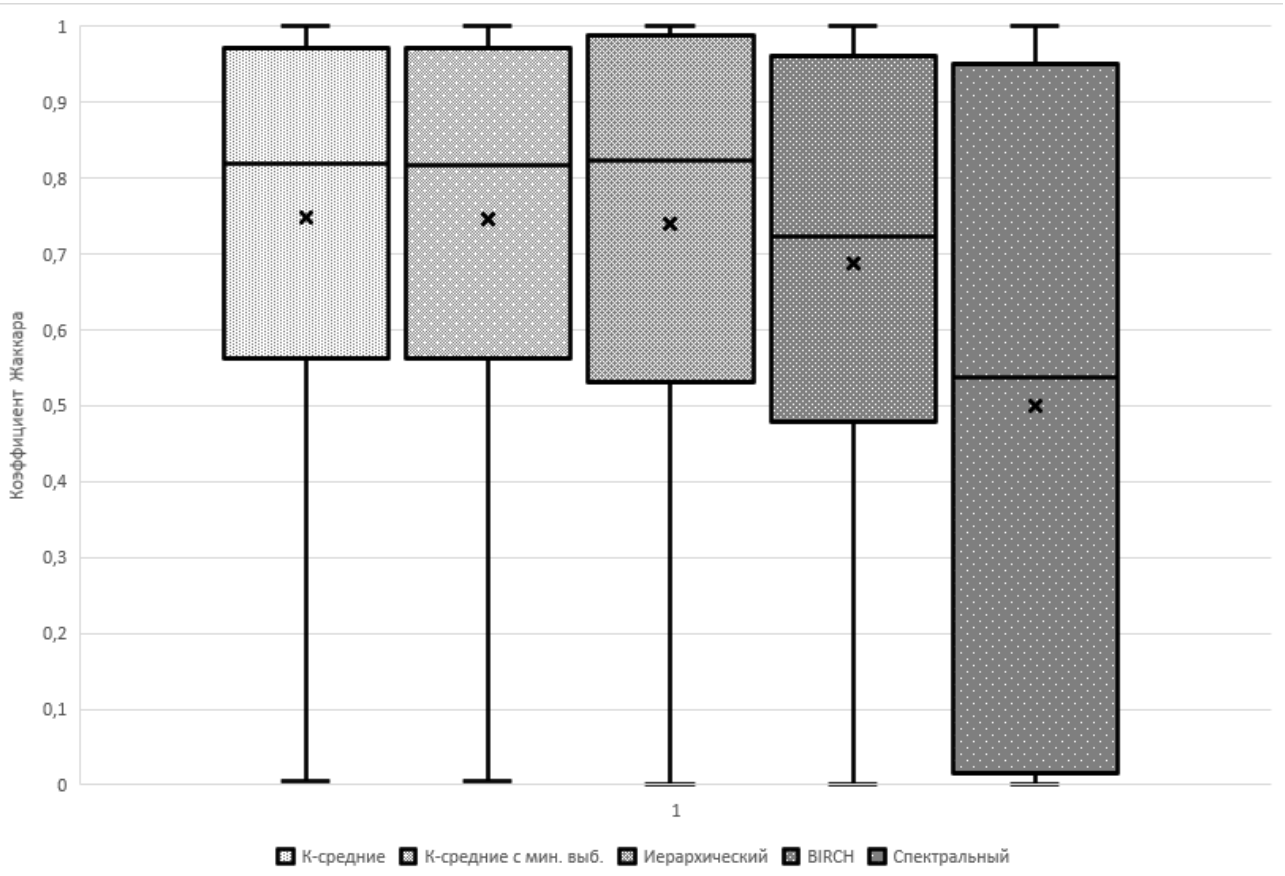

Рис. 6. Диаграмма размаха коэффициента подобия Жаккара

Fig. 6. Boxplots of Jaccard similarity coefficient

Результаты проверки эффективности алгоритмов кластеризации полностью соответствуют выводам, полученным в результате их оценки по проценту ошибочного отнесения выборок к «чужому кластеру». Самым неэффективным алгоритмом оказался *спектральный*. Все остальные алгоритмы показали достаточно хорошие результаты.

В целом на основе вышеизложенного анализа данных можно сделать следующие выводы:

– добиться разделения гетерогенной выборки с достаточно высокой точностью (на уровне не более 5 % ошибочного отнесения к «чужой» подвыборке) с помощью существующих алгоритмов кластеризации практически невозможно – подобный уровень достигается только в 29–40 % случаев в зависимости от выбранного алгоритма;

– если для педагогического исследования требуется очень точное разделение выборки, то педагогу-исследователю следует внимательно визуально проанализировать плотность распределения первоначальной выборки: распределение должно быть явно мультимодальным, моды распределения должны находиться на значительном расстоянии друг от друга, в этом случае лучше всего использовать *иерархический алгоритм* кластеризации как наиболее эффективный на высоком уровне точности кластеризации;

– в случае если высокой точности кластеризации не требуется (на уровне не более 32 % ошибочного отнесения к «чужой» подвыборке), то педагог-исследователь может выбирать либо оба алгоритма *К-средних*, либо *иерархический алгоритм* – все они предлагают примерно одинаковую точность кластеризации;

– *алгоритм BIRCH*, хоть и показывает чуть худшую эффективность в целом, тоже может использоваться для разделения мультимодальных выборок;

– *спектральный алгоритм* демонстрирует полную неэффективность по всем аналитическим показателям – по *коэффициенту Жаккара* алгоритм показал максимальную энтропию, близкую к случайному распределению результатов выборки к одному из двух кластеров.

Помимо оценки эффективности алгоритмов кластеризации, нами также был проведен анализ влияния параметров моделирования педагогического эксперимента на показатели эффективности алгоритмов кластеризации (коэффициент подобия Жаккара). Под параметрами моделирования понимаются прежде всего: математическое ожидание подвыборок; стандартное отклонение подвыборок; асимметрия подвыборок; количество наблюдений в подвыборках; расстояние между подвыборками. Однако в качестве аргументов уравнения регрессии нами были выбраны не сами параметры моделирования, а их математические отношения: расстояния между величинами математического ожидания подвыборок, сумма стандартных отклонений, сумма количества наблюдений, модуль разницы коэффициента асимметричности распределения. В качестве зависимой переменной использовался *коэффициент Жаккара иерархического алгоритма* кластеризации. Уравнение регрессии представлено в следующей формуле:

$$J = F(x_1, x_2, ..., x_4)$$
$$x_1 = \mu_2 - \mu_1$$
$$x_2 = \sigma_1 + \sigma_2$$
$$x_3 = N_1 + N_2$$
$$x_4 = |\alpha_2 - \alpha_1| \quad (1);$$

где

J – коэффициент подобия Жаккара;

μ – математическое ожидание подвыборки, баллы;

σ – величина стандартного отклонения подвыборки, баллы;

N – число наблюдений в подвыборке;

α – коэффициент асимметричности подвыборки.

Корреляционный анализ, представленный в таблице 6, показывает, что наиболее сильно влияющим на показатель эффективности кластеризации является расстояние между подвыборками. Вторым параметром по степени влияния выступает величина стандартного отклонения подвыборок. Размер выборок и величина асимметрии показали почти нулевую корреляцию.

Таблица 6

Коэффициенты корреляции параметров моделирования эксперимента

Table 6

Correlation coefficients of experimental simulation parameters

	J	**$\sigma_1 + \sigma_2$**	**$\mu_2 - \mu_1$**	**$N_1 + N_2$**	**$\lvert\alpha_2 - \alpha_1\rvert$**
J	1	–	–	–	–
$\sigma_1 + \sigma_2$	-0,24	1	–	–	–
$\mu_2 - \mu_1$	0,72	0	1	–	–
$N_1 + N_2$	$-6{,}9 \cdot 10^{-3}$	0	0	1	–
$\lvert\alpha_2 - \alpha_1\rvert$	0,04	$-9{,}3 \cdot 10^{-20}$	0	$5 \cdot 10^{-20}$	1

Таким образом, для расчета уравнения регрессии нами учитывались только переменные расстояния между математическими ожиданиями подвыборок и сумма стандартных отклонений выборок. Окончательное уравнение регрессии приняло следующий вид:

$$J = F(x_1, x_2) = 0,014x_1 + 0,018x_2$$
$$x_1 = \mu_2 - \mu_1$$
$$x_2 = \sigma_1 + \sigma_2$$

$$R^2 = 0,9$$
$$SE_J = 0,24 \quad (2);$$

где

J – коэффициент подобия Жаккара;

μ – математическое ожидание подвыборки, баллы;

σ – величина стандартного отклонения подвыборки;

R^2 – коэффициент детерминации уравнения;

SE_J – величина стандартной ошибки предсказанного значения J.

Коэффициент детерминации (R^2) полученного уравнения составил 0,9, что говорит о том, что выбранные аргументы функции уравнения на 90 % определяют коэффициент подобия Жаккара. Величина стандартной ошибки составила – 0,24.

Проведенный анализ влияния параметров компьютерного моделирования эксперимента позволяет прийти к следующим выводам:

– наибольшее влияние на эффективность кластеризации оказала разница между математическими ожиданиями двух подвыборок, поэтому педагогу-исследователю в первую очередь следует обращать внимание на визуальную форму распределения, получившегося в результате: если на распределении присутствует несколько пиков (мод) и несколько локальных минимумов между ними, то имеет смысл разделить большую выборку на несколько подвыборок;

– вторым по важности фактором, влияющим на эффективность кластеризации, является величина стандартного отклонения;

– величины асимметрии и количества наблюдений выборки никак не влияют на эффективность кластеризации.

Также был проведен регрессионный анализ влияния статистических параметров полученной смешанной выборки на коэффициент подобия Жаккара. Этот анализ наиболее приближен к реальной ситуации педагогического эксперимента, когда получены сырые данные эксперимента и необходимо провести их первичную статистическую обработку, а также предсказать, насколько выборка гомогенна или мультимодальна. В качестве параметров описательной статистики нами использовались: показатели вероятности принадлежности распределения к нормальному (*p-показатель*); усеченные показатели минимума (P_5) и максимума (P_{95}) выборки, а также медиана (P_{50}) и квартили (P_{25}, P_{75}) выборок (рис. 7).

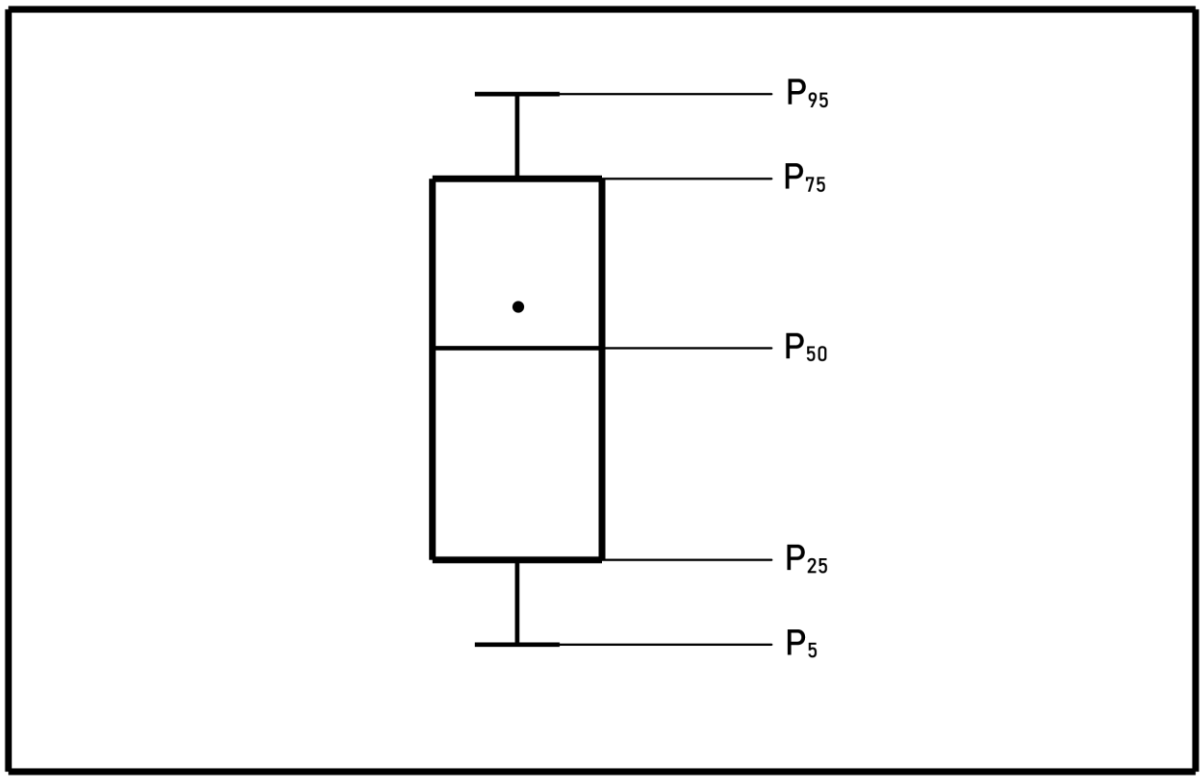

Рис. 7. Схема определения квартилей выборки на диаграмме размаха

Fig. 7. The scheme for determining sample quartiles on a boxplot

В уравнении регрессии использовались не сами показатели квартилей, а отношение межквартильных диапазонов к диапазону между минимумом (пятым процентилем) и максимумом (девяносто пятым процентилем) выборки. В качестве зависимой переменной использовался *коэффициент Жаккара иерархического алгоритма* кластеризации. Уравнение регрессии представлено в следующей формуле:

$$J = F(p, x_1, x_2, ..., x_4)$$

$$x_1 = \frac{P_{25} - P_5}{P_{95} - P_5}$$

$$x_2 = \frac{P_{50} - P_{25}}{P_{95} - P_5}$$

$$x_3 = \frac{P_{75} - P_{50}}{P_{95} - P_5}$$

$$x_4 = \frac{P_{95} - P_{75}}{P_{95} - P_5} \quad (3);$$

где

J – коэффициент подобия Жаккара;

p – показатель вероятности принадлежности распределения к нормальному;

P_5 – минимум (пятый процентиль) выборки, баллы;

P_{95} – максимум (девяносто пятый процентиль) выборки, баллы;

P_{25} – первый квартиль (двадцать пятый процентиль) выборки, баллы;

P_{75} – третий квартиль (семьдесят пятый процентиль) выборки, баллы;

P_{50} – медиана (пятидесятый процентиль) выборки, баллы.

Для более полной и точной оценки взаимосвязи аргументов функции использовались, кроме межквартильных расстояний, еще и их произведение между собой, чтобы учесть не только сами переменные, но и то, как они влияют друг на друга. Корреляционный анализ представлен в таблице 7. В данном случае мы не исключали переменные из уравнения регрессии из-за низких одиночных значений коэффициента корреляции, а считали их комплексом переменных (p – показатель, x-переменные, произведение x-переменных), и для каждого комплекса в качестве основного коэффициента корреляции рассматривался максимальный. Самая низкая корреляция с коэффициентом Жаккара наблюдалась у показателя вероятности принадлежности выборки к нормальному (-0,28), вторыми по степени влияния являются величины отношения межквартильных диапазонов к диапазону между минимумом и максимумом выборки (-0,58), самый высокий коэффициент корреляции показало произведение отношения межквартильных диапазонов (-0,82).

Таблица 7

Коэффициенты корреляции параметров описательной статистики

Table 7

Correlation coefficients of descriptive statistics parameters

	J	***p***	***x1***	***x2***	***x3***	***x4***	***x1x2***	***x1x3***	***x1x4***	***x2x3***	***x2x4***	***x3·x4***
J	1	–	–	–	–	–	–	–	–	–	–	–
p	-0,28	1	–	–	–	–	–	–	–	–	–	–
x1	-0,52	0,13	1	–	–	–	–	–	–	–	–	–
x2	0,54	-0,15	-0,29	1	–	–	–	–	–	–	–	–
x3	0,31	-0,10	-0,64	-0,35	1	–	–	–	–	–	–	–
x4	-0,58	0,19	0,02	-0,68	0,02	1	–	–	–	–	–	–
x1x2	0,05	-0,02	0,67	0,44	-0,76	-0,52	1	–	–	–	–	–
x1x3	-0,43	0,13	0,46	-0,71	0,23	0,18	-0,10	1	–	–	–	–
x1x4	-0,82	0,27	0,72	-0,61	-0,47	0,65	0,11	0,47	1	–	–	–
x2x3	0,63	-0,18	-0,71	0,40	0,62	-0,53	-0,32	-0,23	-0,81	1	–	–
x2x4	0,07	0,05	-0,27	0,45	-0,40	0,20	0,04	-0,56	-0,05	0,00	1	–
x3x4	-0,08	0,02	-0,47	-0,66	0,72	0,65	-0,84	0,22	0,01	0,03	-0,17	1

Ввиду того, что включение в уравнение переменной *p* и группы переменных *x* ничтожно влияло на коэффициент детерминации и показатель стандартной ошибки, при этом сильно усложняя уравнение, в конечном варианте уравнения регрессии использовались лишь произведения переменных *x:*

$$J = F(x_1, x_2, ..., x_4) = 6{,}3x_1x_2 - 1{,}06x_1x_3 - 4{,}3x_1x_4 + 3{,}83x_2x_3 + 1{,}6x_2x_4 + 5{,}05x_3x_4$$

$$x_1 = \frac{P_{25} - P_5}{P_{95} - P_5}$$

$$x_2 = \frac{P_{50} - P_{25}}{P_{95} - P_5}$$

$$x_3 = \frac{P_{75} - P_{50}}{P_{95} - P_5}$$

$$x_4 = \frac{P_{95} - P_{75}}{P_{95} - P_5}$$

$$R^2 = 0,965$$
$$SE_J = 0,14 \qquad (4)$$

Коэффициент детерминации (R^2) полученного уравнения составил 0,96, что говорит о том, что выбранные аргументы функции уравнения на 96 % определяют коэффициент

подобия Жаккара, что даже выше, чем у уравнения, содержащего параметры моделирования выборок. Величина стандартной ошибки составила – 0,14; 95 % доверительный интервал – 0,28.

Проведенный анализ влияния явных факторов позволяет прийти к следующим выводам:

– единственными факторами, которые с помощью уравнений регрессии позволяют предсказать показатели эффективности кластеризации, оказались произведения соотношений межквартильных размахов;

– никакие другие параметры описательной статистики, например, показатели нормальности распределения или его асимметрии, существенным образом не влияли на конечную точность уравнения;

– можно сделать вывод, что, математически анализируя межквартильные или даже межпроцентильные расстояния, можно достаточно точно с помощью уравнений регрессии предсказывать свойства полученных распределений – в нашем случае вероятность ошибки кластеризации распределения.

Обсуждение. Заключение

В результате моделирования педагогического эксперимента на 20160 вариантах выборки, различающихся по таким параметрам, как размер подвыборок, величины стандартного отклонения, расстояния между математическими ожиданиями, коэффициента асимметрии подвыборок, мы пришли к следующим выводам.

1. Добиться высокой точности (до 5 % ошибок) разделения гетерогенной выборки с помощью алгоритмов кластеризации довольно сложно, это удается лишь в 29–40 % случаев. Для точного разделения в педагогическом исследовании выборка должна иметь явно мультимодальное распределение, и в этом случае рекомендуется использовать *иерархический алгоритм*. Если точность до 32 % ошибок приемлема, подойдут и *алгоритмы К-средних*, и *иерархический алгоритм*.

2. Среди параметров моделирования педагогического эксперимента на эффективность кластеризации наибольшее влияние оказывает разница между математическими ожиданиями подвыборок, поэтому при наличии нескольких пиков на распределении рекомендуется разделить выборку на подгруппы. Второстепенным фактором влияния является стандартное отклонение, тогда как асимметрия и количество наблюдений на результаты кластеризации не влияют.

3. Для предсказания эффективности кластеризации ключевыми являются произведения соотношений межквартильных размахов. Другие параметры описательной статистики, включая показатели нормальности распределения и его асимметрии, не оказали влияния на точность уравнения. Таким образом, анализируя межквартильные расстояния, можно через уравнения регрессии предсказать вероятность ошибки в кластеризации распределения.

Благодаря этим выводам мы можем вывести следующий алгоритм действий для педагога-исследователя.

1. Прежде всего следует проверить полученную выборку на предмет принадлежности к нормальному распределению любым из статистических тестов (тест Колмогорова, тест Шапиро – Вилка). Высокий показатель вероятности (*p-показатель* > 0,5), что выборка принадлежит к нормальному распределению, однозначно исключает мультимодальность выборки.

2. Даже если существует априорное знание, что выборка действительно мультимодальна, но при этом достоверно принадлежит к нормальному распределению, то разделение

выборки будет сопровождаться большим количеством ошибок кластеризации.

3. Если все же есть подозрение, что полученная выборка имеет мультимодальный характер, то необходимо визуально оценить распределение. Наиболее удобная форма представления – это скрипичная диаграмма (*violin plot*).

4. Если на скрипичной диаграмме просматривается несколько пиков (мод), то можно воспользоваться статистическими тестами, направленными на выявление количества мод, например, таких, как метод Сильвермана, метод ACR , метод Чена – Холла, метод Хартигана, метод Фишера – Маррона, метод Холла – Йорка[8] [8; 12; 13; 17; 25], при этом нельзя однозначно полагаться только на визуальный либо только на статистический анализ количества мод, а делать суждение на основе комплексного анализа.

5. Для оценки вероятности ошибок кластеризации двухмодальной выборки можно воспользоваться уравнением регрессии (4), однако к любым выводам, полученным с помощью этого уравнения, следует относиться с осторожностью, так как 95 % доверительный интервал коэффициента Жаккара этого уравнения равняется 0,28.

6. После определения количества мод педагог-исследователь может воспользоваться одним из следующих алгоритмов кластеризации: *К-средние*, *К-средние с минимальными выборками*, *иерархический алгоритм*. Оба *алгоритма К-средних* являются наиболее универсальными; *иерархический алгоритм* рекомендуется использовать, если визуально распределение имеет несколько явно выраженных мод.

Представленная научная работа не является всеобъемлющей, поэтому необходимо упомянуть основные проблемы и границы применения результатов, полученных в данном исследовании.

1. Настоящее исследование посвящено изучению эффективности кластеризации только двухмодальных выборок, поэтому результаты ограничено применимы для случаев, когда количество мод превышает число «два». Особенно эти ограничения касаются уравнения регрессии (4) – вычислять коэффициент Жаккара для трех- или четырехмодальных выборок с помощью этого уравнения не имеет смысла, так как он применим для сравнения только двух выборок. Определение количества мод, а также алгоритм исследования выборок с количеством мод больше двух будет предметом наших дальнейших исследований.

2. Существует неразрешимое противоречие между точностью алгоритма кластеризации и возможностью предсказания его точности посредством уравнения регрессии – чем больше точность алгоритма, тем меньше случаев с высокой долей ошибок и, соответственно, меньше предсказательная сила уравнения для таких случаев. Увеличение же количества параметров (децили и квинтили вместо квартилей) сильно усложнит конечное уравнение при небольшом увеличении точности. Использование других способов анализа и предсказания результатов (в том числе нейросетевых моделей) также станет предметом наших дальнейших исследований.

[8] Crujeiras-Casais R. M., Alonso J. A., Casal A. R. Mode testing, critical bandwidth and excess mass // XXXV Congreso Nacional SEIO: IX Jornadas de Estadística Pública: Universidad Pública de Navarra, Pamplona, del 26 al 29 de mayo de 2015. – Departamento de Estadística e Investigación Operativa. Universidad de Navarra, 2015. – С. 60–60.

Hall P., York M. On the calibration of Silverman's test for multimodality // Statistica Sinica. – 2001. – P. 515–536.

СПИСОК ЛИТЕРАТУРЫ

1. Абитов Р. Н. Пути повышения валидности и повторяемости экспериментальных педагогических исследований // Казанский педагогический журнал. – 2022. – № 4. – С. 79–90. DOI: https://10.51379/kpj.2022.154.4.009 URL: https://elibrary.ru/item.asp?id=49482910
2. Ершов К. С., Романова Т. Н. Анализ и классификация алгоритмов кластеризации // Новые информационные технологии в автоматизированных системах. – 2016. – № 19. – С. 274–279. URL: https://elibrary.ru/item.asp?id=25864070
3. Подвальный С. Л., Плотников А. В., Белянин А. М. Сравнение алгоритмов кластерного анализа на случайном наборе данных // Вестник Воронежского государственного технического университета. – 2012. – Т. 8, № 5. – С. 4–6. URL: https://elibrary.ru/item.asp?id=17743528
4. Сивоголовко Е. В. Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании. – 2011. – № 4. – С. 14–31. URL: https://elibrary.ru/item.asp?id=21786023
5. Xiaowei Xu, Ester M., Kriegel H.-P., Sander J. A distribution-based clustering algorithm for mining in large spatial databases // Proceedings 14th International Conference on Data Engineering. DOI: https://doi.org/10.1109/icde.1998.655795
6. Azzalini A., Valle A. D. The multivariate skew-normal distribution // Biometrika. – 1996. – Vol. 83 (4). – P. 715–726. DOI: https://doi.org/10.1093/biomet/83.4.715
7. Banfield J. D., Raftery A. E. Model-based Gaussian and non-Gaussian clustering // Biometrics. – 1993. – Vol. 49 (3). – P. 803–821. DOI: https://doi.org/10.2307/2532201
8. Cheng M.-Y., Hall P. Calibrating the excess mass and dip tests of modality // Journal of the Royal Statistical Society: Series B: Statistical Methodology. – 1998. – Vol. 60 (3). – P. 579–589. DOI: https://doi.org/10.1111/1467-9868.00141
9. Rodriguez M. Z., Comin C. H., Casanova D., Bruno O. M., Amancio D. R., Costa L. da F., Rodrigues F. A. Clustering algorithms: A comparative approach // PloS ONE. – 2019. – Vol. 14 (1). – P. e021023. DOI: https://doi.org/10.1371/journal.pone.0210236
10. Reynolds A. P., Richards G., de la Iglesia B., Rayward-Smith V. J. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms // Journal of Mathematical Modelling and Algorithms. – 2006. – Vol. 5 (4). – P. 475–504. DOI: https://doi.org/10.1007/s10852-005-9022-1
11. Kinnunen T., Sidoroff I., Tuononen M., Fränti P. Comparison of clustering methods: A case study of text-independent speaker modeling // Pattern Recognition Letters. – 2011. – Vol. 32 (13). – P. 1604–1617. DOI: https://doi.org/10.1016/j.patrec.2011.06.023
12. Ameijeiras-Alonso J., Crujeiras R. M., Rodríguez-Casal A. Mode testing, critical bandwidth and excess mass // TEST. – 2018. – Vol. 28 (3). – P. 900–919. DOI: https://doi.org/10.1007/s11749-018-0611-5
13. Fisher N. I., Marron J. S. Mode testing via the excess mass estimate Fisher N.I. Mode testing via the excess mass estimate // Biometrika. – 2001. – Vol. 88 (2). – P. 499–517. DOI: https://doi.org/10.1093/biomet/88.2.499
14. Fowlkes E. B., Mallows C. L. A method for comparing two hierarchical clusterings: Rejoinder // Journal of the American statistical association. – 1983. – Vol. 78 (383). – P. 584. DOI: https://doi.org/10.2307/2288123
15. Guha S., Rastogi R., Shim K. Cure: an efficient clustering algorithm for large databases. Information Systems. – 2001. – Vol. 26 (1). – P. 35–58. DOI: https://doi.org/10.1016/s0306-4379(01)00008-4

16. Guha S., Rastogi R., Shim K. ROCK: a robust clustering algorithm for categorical attributes // Proceedings 15th International Conference on Data Engineering. 1999. (Cat. No.99CB36337). DOI: https://doi.org/10.1109/icde.1999.754967
17. Hartigan J. A., Hartigan P. M. The dip test of unimodality // The annals of Statistics. – 1985. – Vol. 13 (1). – P. 70–84. DOI: https://doi.org/10.1214/aos/1176346577
18. Jung Y. G., Kang M. S., Heo J. Clustering performance comparison using K-means and expectation maximization algorithms // Biotechnology & Biotechnological Equipment. – 2014. – Vol. 28 (sup1). – P. S44–S48. DOI: https://doi.org/10.1080/13102818.2014.949045
19. Karypis G., Eui-Hong Han, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling // Computer. – 1999. – Vol. 32 (8). – P. 68–75. DOI: https://doi.org/10.1109/2.781637
20. Kruskal W. H., Wallis W. A. Errata: Use of Ranks in One-Criterion Variance Analysis // Journal of the American Statistical Association. – 1953. – Vol. 48 (264). – P. 907. DOI: https://doi.org/10.2307/2281082
21. Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering points to identify the clustering structure // ACM Sigmod record. – 1999. – Vol. 28 (2). – P. 49–60. DOI: https://doi.org/10.1145/304181.304187
22. Rand W. M. Objective criteria for the evaluation of clustering methods // Journal of the American Statistical association. – 1971. – Vol. 66 (336). – P. 846–850. DOI: https://doi.org/10.1080/01621459.1971.10482356
23. Sculley D. Web-scale k-means clustering // Proceedings of the 19th international conference on World wide web. – 2010. – P. 1177–1178. DOI: https://doi.org/10.1145/1772690.1772862
24. Shi J., Malik J. Normalized cuts and image segmentation // IEEE Transactions on pattern analysis and machine intelligence. – 2000. – Т. 22. – № 8. – С. 888-905. DOI: https://doi.org/10.1109/cvpr.1997.609407
25. Silverman B. W. Using kernel density estimates to investigate multimodality // Journal of the Royal Statistical Society: Series B (Methodological). – 1981. – Vol. 43 (1). – P. 97–99. DOI: https://doi.org/10.1111/j.2517-6161.1981.tb01155.x
26. Ward J. H. Hierarchical grouping to optimize an objective function // Journal of the American statistical association. – 1963. – Vol. 58 (301). – P. 236–244. DOI: https://doi.org/10.1080/01621459.1963.10500845
27. Wilkin G. A., Huang X. K-means clustering algorithms: implementation and comparison // Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007). – 2007. – P. 133–136. DOI: https://doi.org/10.1109/imsccs.2007.51
28. Xu D., Tian Y. A comprehensive survey of clustering algorithms // Annals of Data Science. – 2015. – Vol. 2 (2). – P. 165–193. DOI: https://doi.org/10.1007/s40745-015-0040-1
29. Zhang T., Ramakrishnan R., Livny M. BIRCH: an efficient data clustering method for very large databases // ACM sigmod record. – 1996. – Vol. 25 (2). – P. 103–114. DOI: https://doi.org/10.1145/235968.233324

Поступила: 09 января 2024 Принята: 11 марта 2024 Опубликована: 30 апреля 2024

Заявленный вклад авторов:

Абитов Руслан Назилович: организация исследования, концепция и дизайн исследования, моделирование эксперимента, статистическая обработка данных, интерпретация результатов, обсуждение и выводы, написание текста статьи.

Сафин Раис Семигуллович: литературный обзор, обсуждение концепции исследования, определение математического аппарата исследования, написание текста статьи.

Все авторы ознакомились с результатами работы и одобрили окончательный вариант рукописи.

Информация о конфликте интересов:

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи

Информация об авторах

Абитов Руслан Назилович
кандидат педагогических наук, доцент,
кафедра иностранных языков,
Казанский государственный архитектурно-строительный университет,
ул. Зеленая, д. 1, 420043, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0003-4219-9815
E-mail: rouslan.abitov@gmail.com

Сафин Раис Семигуллович
доктор педагогических наук, профессор, заведующий кафедрой,
кафедра профессионального обучения педагогики и социологии,
Казанский государственный архитектурно-строительный университет,
ул. Зеленая, д. 1, 420043, г. Казань, Россия.
ORCID ID: https://orcid.org/0000-0003-1864-7876
E-mail: safin@kgasu.ru

Science for Education Today

2024, vol. 14, issue 2 http://en.sciforedu.ru/ ISSN 2658-6762

DOI: 10.15293/2658-6762.2402.06 Research Full Article / Article language: Russian

Analysis of the effectiveness of clustering algorithms for multimodal samples using computer simulation of an educational experiment

Ruslan N. Abitov [1], Rais S. Safin[1]

[1] Kazan State University of Architecture and Engineering, Kazan, Russian Federation

Abstract

Introduction. *The article is devoted to the problem of primary data processing of pedagogical experiments having a multimodal character. The purpose of the study is to identify the most effective and universal clustering algorithms for pedagogical experiments.*

Materials and Methods. *The study used the method of modeling a pedagogical experiment. The analysis of 5 clustering algorithms is conducted. The effectiveness of clustering algorithms was evaluated based on the proportion of observations with clustering errors at various tolerance levels and the Jacquard similarity coefficient. Regression analysis was used to assess the influence of modeling parameters of a pedagogical experiment and indicators of descriptive statistics on the effectiveness of clustering algorithms.*

Results. *The assessment of the effectiveness of various data clustering algorithms is provided, as well as a correlation and regression analysis of factors affecting clustering efficiency indicators was carried out.*

Conclusions. *The most effective clustering algorithms for multimodal samples include the K-means algorithm and the agglomerative hierarchical algorithm. The results obtained in this research can be used for statistical analysis of pedagogical, psychological, sociological, biological and medical research data.*

Keywords

Educational experiment modeling; Data clustering algorithms; Multimodal samples; Data analysis in education.

For citation

Abitov R. N., Safin R. S. Analysis of the effectiveness of clustering algorithms for multimodal samples using computer simulation of an educational experiment. *Science for Education Today*, 2024, vol. 14 (2), pp. 125–151. DOI: http://dx.doi.org/10.15293/2658-6762.2402.06

Corresponding Author: Ruslan N. Abitov, rouslan.abitov@gmail.com

REFERENCES

1. Abitov R. N. On the ways to increase the validity and repeatability of experimental pedagogical research. *Kazan Pedagogical Journal,* 2022, no. 4, pp. 79–90. (In Russian) DOI: https://10.51379/kpj.2022.154.4.009 URL: https://elibrary.ru/item.asp?id=49482910
2. Ershov K. S., Romanova T. N. Analysis and classification of clustering algorithms. *New Information Technologies in Automated Systems*, 2016, no. 19, pp. 274–279. (In Russian) URL: https://elibrary.ru/item.asp?id=25864070
3. Podvalny S. L., Plotnikov A. V., Belyanin A. M. Comparison of cluster analysis of algorithms random set of data. *Bulletin of Voronezh State Technical University*, 2012, vol. 8 (5), pp. 4–6. (In Russian) URL: https://elibrary.ru/item.asp?id=17743528
4. Sivogolovko E. V. Methods for assessing the quality of clear clustering. *Computer Tools in Education*, 2011, no. 4, pp. 14–31. (In Russian) URL: https://elibrary.ru/item.asp?id=21786023
5. Xiaowei Xu, Ester M., Kriegel H.-P., Sander J. A distribution-based clustering algorithm for mining in large spatial databases. *Proceedings 14th International Conference on Data Engineering*. DOI: https://doi.org/10.1109/icde.1998.655795
6. Azzalini A., Valle A. D. The multivariate skew-normal distribution. *Biometrika*, 1996, vol. 83 (4), pp. 715–726. DOI: https://doi.org/10.1093/biomet/83.4.715
7. Banfield J. D., Raftery A. E. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 1993, vol. 49 (3), pp. 803–821. DOI: https://doi.org/10.2307/2532201
8. Cheng M.-Y., Hall P. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 1998, vol. 60 (3), pp. 579–589. DOI: https://doi.org/10.1111/1467-9868.00141
9. Rodriguez M. Z., Comin C. H., Casanova D., Bruno O. M., Amancio D. R., Costa L. da F., Rodrigues F. A. Clustering algorithms: A comparative approach. *PloS One*, 2019, vol. 14 (1), pp. e021023. DOI: https://doi.org/10.1371/journal.pone.0210236
10. Reynolds A. P., Richards G., de la Iglesia B., Rayward-Smith V. J. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modeling and Algorithms*, 2006, vol. 5 (4), pp. 475–504. DOI: https://doi.org/10.1007/s10852-005-9022-1
11. Kinnunen T., Sidoroff I., Tuononen M., Fränti P. Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters*, 2011, vol. 32 (13), pp. 1604–1617. DOI: https://doi.org/10.1016/j.patrec.2011.06.023
12. Ameijeiras-Alonso J., Crujeiras R. M., Rodríguez-Casal A. Mode testing, critical bandwidth and excess mass. *TEST*, 2018, vol. 28 (3), pp. 900–919. DOI: https://doi.org/10.1007/s11749-018-0611-5
13. Fisher N. I., Marron J. S. Mode testing via the excess mass estimate. *Biometrika*, 2001, vol. 88 (2), pp. 499–517. DOI: https://doi.org/10.1093/biomet/88.2.499
14. Fowlkes E. B., Mallows C. L. A method for comparing two hierarchical clusterings: Rejoinder. *Journal of the American Statistical Association*, 1983, vol. 78 (383), pp. 584. DOI: https://doi.org/10.2307/2288123
15. Guha S., Rastogi R., Shim K. Cure: an efficient clustering algorithm for large databases. *Information Systems*, 2001, vol. 26 (1), pp. 35–58. DOI: https://doi.org/10.1016/s0306-4379(01)00008-4
16. Guha S., Rastogi R., Shim K. ROCK: a robust clustering algorithm for categorical attributes. *Proceedings 15th International Conference on Data Engineering,* 1999. Cat. No.99CB36337. DOI: https://doi.org/10.1109/icde.1999.754967

17. Hartigan J. A., Hartigan P. M. The dip test of unimodality. *The Annals of Statistics*, 1985, vol. 13 (1), pp. 70–84. DOI: https://doi.org/10.1214/aos/1176346577
18. Jung Y. G., Kang M. S., Heo J. Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment*, 2014, vol. 28 (sup1), pp. S44–S48. DOI: https://doi.org/10.1080/13102818.2014.949045
19. Karypis G., Eui-Hong Han, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 1999, vol. 32 (8), pp. 68–75. DOI: https://doi.org/10.1109/2.781637
20. Kruskal W. H., Wallis W. Errata: Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 1953, vol. 48 (264), pp. 907. DOI: https://doi.org/10.2307/2281082
21. Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Record*, 1999, vol. 28 (2), pp. 49–60. DOI: https://doi.org/10.1145/304181.304187
22. Rand W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971, vol. 66 (336), pp. 846–850. DOI: https://doi.org/10.1080/01621459.1971.10482356
23. Sculley D. Web-scale k-means clustering. *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178. DOI: https://doi.org/10.1145/1772690.1772862
24. Shi J., Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, vol. 22 (8), pp. 888–905. DOI: https://doi.org/10.1109/cvpr.1997.609407
25. Silverman B. W. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological),* 1981, vol. 43 (1), pp. 97–99. DOI: https://doi.org/10.1111/j.2517-6161.1981.tb01155.x
26. Ward J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 1963, vol. 58 (301), pp. 236–244. DOI: https://doi.org/10.1080/01621459.1963.10500845
27. Wilkin G. A., Huang X. K-means clustering algorithms: Implementation and comparison. *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007),* 2007, pp. 133–136. DOI: https://doi.org/10.1109/imsccs.2007.51
28. Xu D., Tian Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2015, vol. 2 (2), pp. 165–193. DOI: https://doi.org/10.1007/s40745-015-0040-1
29. Zhang T., Ramakrishnan R., Livny M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod Record*, 1996, vol. 25 (2), pp. 103–114. DOI: https://doi.org/10.1145/235968.233324

Submitted: 09 January 2024 Accepted: 10 March 2024 Published: 30 April 2024

The authors' stated contribution:

Ruslan Nazilovich Abitov

Contribution of the co-author: organization of the research, concept and design of the research, modeling of the experiment, statistical data processing, interpretation of results, discussion and conclusions, writing of the text of the article

Rais Semigullovich Safin

Contribution of the co-author: literary review, discussion of the concept of the research, definition of the mathematical apparatus of the research, writing of the text of the article.

All authors reviewed the results of the work and approved the final version of the manuscript.

Information about competitive interests:

The authors declare no apparent or potential conflicts of interest in connection with the publication of this article

Information about the Authors

Ruslan Nazilovich Abitov
Candidate of Pedagogical Sciences, Associate Professor,
Department of Foreign Languages,
Kazan State University of Architecture and Engineering,
Zelenaya str., 1, 420043, Republic of Tatarstan, Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0003-4219-9815
E-mail: rouslan.abitov@gmail.com

Rais Semigullovich Safin
Doctor of Pedagogical Sciences, Professor, Head of the Department,
Department of Professional Education of Pedagogy and Sociology,
Kazan State University of Architecture and Engineering,
Zelenaya str., 1, 420043, Republic of Tatarstan, Kazan, Russian Federation.
ORCID ID: https://orcid.org/0000-0003-1864-7876
E-mail: safin@kgasu.ru