УДК 004.622

Теоретические основы методов кластеризации данных в интеллектуальном анализе

Кучумов И. В., Яндекс, г. Москва

В данной статье рассматриваются теоретические основы и принципы работы основных методов кластерного анализа данных, используемых в задачах интеллектуальной аналитики. Подробно анализируются работы ряда исследователей в области кластеризации, описан прогресс в разработке и применении классических и новейших подходов к группировке структурно сложных, разнородных данных с использованием аппарата статистики, нейронных сетей, математического моделирования. Рассмотрены математические основания иерархических, вероятностных, плотностных, графовых и других методов кластеризации, теоретически доказана эффективность их применения на разных типах данных в зависимости от поставленных аналитических целей. Отдельное внимание уделено проблематике кластеризации больших объемов разнородной информации в условиях возрастания скорости поступающих данных и требований к оперативности их обработки. Продемонстрирован потенциал гибридных нейросетевых и распределенных методов кластеризации для эффективного масштабируемого анализа Big Data в высокопроизводительных вычислительных системах. Показано, что несмотря на значительный прогресс, ряд фундаментальных вопросов в данной области остается открытым и требует дальнейших междисциплинарных исследований на стыке статистики, математики и компьютерных наук.

Введение

В эпоху беспрецедентного роста объёмов генерируемой цифровой информации остро встаёт вопрос её структуризации и извлечения полезных знаний.

Кластеризация, или группировка массивов неупорядоченных многомерных данных на однородные подмножества, относится к числу важнейших инструментов интеллектуальной аналитики. При этом решение сложных прикладных задач требует комплексного применения богатого арсенала разработанных математических методов. В результате многолетних усилий исследователей в области кластеризации сложился широкий спектр подходов, включающих иерархические, вероятностные, плотностные, графовые алгоритмы и другие. Как показано в проанализированных работах, каждый из этих методов обладает своей областью эффективного применения, определяемой природой данных, наличием априорной информации, требованиями к ресурсоёмкости обработки.

Вместе с тем в современных условиях возникают качественно новые вызовы: необходимость анализа колоссальных, быстро прирастающих, зашумлённых, неоднородных массивов Big Data и требует разработки масштабируемых распределённых алгоритмов кластеризации, эффективно использующих возможности суперкомпьютеров, графических процессоров, облачных и квантовых платформ. Перспективны гибридные решения, интегрирующие различные математические подходы.

Таким образом, несмотря на впечатляющий прогресс в области методов интеллектуального анализа, задача кластеризации структурно сложных, быстро растущих объёмов разнородных данных остаётся одним из наиболее актуальных научно-практических вызовов современности. Требуются дальнейшие скоординированные усилия исследователей для создания нового поколения высокопроизводительных алгоритмов кластеризации, отвечающих запросам Big Data era.

Описание методов и материалов

В основе кластеризации лежит понятие меры сходства объектов, позволяющей количественно оценить, насколько два объекта похожи друг на друга. Часто используют евклидово расстояние для числовых данных или коэффициенты корреляции. Для сложных объектов применяют метрики, учитывающие особенности их структуры.

Проведенное исследование опирается на комплекс взаимодополняющих теоретических и экспериментальных подходов для решения задач кластеризации данных в интеллектуальной аналитике.

Вначале выполнен подробный критический анализ научной литературы по разработке новых алгоритмов кластерного анализа. Далее с помощью математического моделирования изучены теоретические основы широкого спектра детерминированных и вероятностных методов кластеризации.

На следующем этапе реализованы в программном коде как классические, так и предложенные за последние годы алгоритмы с последующей их эмпирической проверкой на реальных данных.

Кроме того, в контексте концепции ансамблевого обучения предложены оригинальные гибридные подходы на базе объединения базовых алгоритмов.

Существует несколько больших классов алгоритмов кластеризации. Рассмотрим их теоретические основы:

- Иерархические алгоритмы последовательно объединяют похожие объекты во всё более крупные кластеры (агломеративные методы) или, наоборот, разбивают один кластер на более мелкие (дивизимные методы). При этом, результат в большинстве случаев представляется в виде дендрограммы.
- Метод к-средних относит каждый объект к ближайшему из заранее заданного числа кластерных центров, которые итеративно пересчитываются как центр масс объектов в кластере.
- Плотностные методы основаны на распределении плотности объектов в пространстве признаков, при этом, высокоплотные области соответствуют кластерам.
- Алгоритмы частиционирования разбивают данные на заданное число кластеров, минимизируя внутрикластерную дисперсию.

- Спектральная кластеризация использует спектральные свойства матрицы сходства объектов для их группировки.
- Вероятностные модели, например скрытые Марковские модели, описывают кластеры как реализацию случайного процесса.
- Для оценки качества кластеризации используют внешние (при наличии эталонных меток классов) и внутренние критерии, основанные на характеристиках самих данных.

Таким образом, за десятилетия исследований в области кластерного анализа накоплен обширный арсенал теоретически обоснованных математических методов. Однако по-прежнему остаются открытыми вопросы автоматического определения оптимального числа кластеров, обработки шумных и выбросных данных, масштабируемости алгоритмов. Активно развиваются гибридные и ансамблевые методы. Таким образом, несмотря на достигнутые успехи, теория и практика кластерного анализа остаётся быстро развивающейся актуальной областью на стыке статистики, искусственного интеллекта и анализа данных.

В исследовании M. Omran и соавторов даётся достаточно лапидарный обзор как классических (иерархический анализ, метод k-средних), так и относительно новых на тот момент (нейросетевые модели, плотностная кластеризация) алгоритмов кластеризации [10]. Производится сравнение различных техник по ключевым критериям - требуемым априорным знаниям о данных, возможности определения оптимального числа кластеров, устойчивости к выбросам и шумам. К сожалению, представленный анализ ограничивается в основном констатацией сильных и слабых сторон того или иного подхода без сравнения эффективности рассматриваемых алгоритмов на конкретных наборах данных.

Mimi Zhang в своей статье подробный анализ как классических, так и новейших методов машинного обучения применительно к кластеризации функциональных данных - временных рядов, пространственно-временных полей, и для каждого из 24 рассмотренных алгоритмов приводится описание базовых принципов, достоинств и недостатков, возможных областей применения [15]. Особо выделяются этапы предварительной обработки данных, признаваемые важнейшими для достижения качественных результатов кластеризации.

Весьма детально в статье Ying Yang и соавторов проанализированы проблемы применения методов кластеризации к неполным данным, содержащим значительное число пропусков [14].

Carlos Casanova с соавторами акцентируют внимание на выявлении групп схожих парето-оптимальных альтернатив в задачах многокритериальной оптимизации с помощью иерархических алгоритмов кластеризации [8]. В исследовании предлагаются оригинальные решения в области визуализации и аналитики получаемых дендрограмм для поддержки принятия решений лицом, ответственным за выбор. Экспериментально доказаны высокая скорость работы и интерпретируемость результатов кластеризации при решении задач планирования IT-проектов.

В работе Gao с соавторами анализируются возможности использования различных алгоритмов кластеризации в исследованиях психического здоровья [9]. Рассматриваются такие методы как иерархическая кластеризация, k-средних, кластеризация на основе графов, нечеткие алгоритмы.

Обсуждаются их преимущества и недостатки применительно к анализу данных в сфере здравоохранения. В обзорной статье Ш. Хечми поднимается важная проблема анализа больших данных с использованием методов кластеризации, автор справедливо отмечает разнообразие существующих алгоритмов кластеризации, что затрудняет выбор наиболее эффективных решений для конкретных задач обработки больших данных [6].

В связи с этим в работе ставится задача всесторонней оценки различных алгоритмов кластерного анализа, ориентированных на работу с большими объёмами данных, с акцентом на производительность и масштабируемость. Проводится экспериментальное тестирование ряда наиболее известных методов на шести крупных наборах реальных данных из открытых источников.

Результаты убедительно демонстрируют преимущества тех или иных подходов для конкретных типов данных. Так, алгоритмы на основе метода опорных векторов и случайных лесов показывают лучшую точность, в то время как кластеризация к-средними, несмотря на простоту, даёт приемлемые результаты при значительно меньшем времени работы.

Таким образом, работа вносит ценный вклад в систематизацию знаний о современных алгоритмах интеллектуального анализа больших данных, позволяя специалистам обоснованно подбирать инструментарий под конкретные прикладные задачи. Иная проанализированная статья С.С. Исакова посвящена многоэтапной кластеризации текстовых документов. Автор отмечает эффективность объектно-ориентированного подхода, когда документы представляются в виде структурированных объектов с набором характеристик, а предлагаемый многоступенчатый анализ, включающий лингвистическую обработку, извлечение признаков и собственно кластеризацию, позволяет группировать научные тексты с учётом тематической близости, оптимизируются возможности эффективного структурирования и навигации в больших массивах научных публикаций [1].

В статье Е.В. Ширинкиной речь идёт об использовании кластеризации, классификации, регрессионного анализа и других методов в образовательной аналитике [7]. Как справедливо отмечает автор, подобные инструменты позволяют оценить текущее состояние и эффективность обучающих программ, спрогнозировать результаты обучения при разных сценариях, оптимизировать учебный процесс.

Особый интерес представляет использование методов интеллектуального анализа для персонализации и адаптации обучения под конкретного студента, с учётом его способностей, предпочтений и динамики освоения материала и открывает путь к построению по-настоящему эффективных образовательных траекторий.

В исследовании Е.Д. Пуговкиной рассматривается актуальная задача применения кластеризации текстов при разработке рекомендательных систем, столь востребованных в условиях лавинообразного роста объёмов пользовательских данных [3].

Группировка текстовых описаний, комментариев, отзывов по тематическому и смысловому признаку открывает новые возможности в построении персонализированных алгоритмов рекомендаций для конкретных пользователей. При этом особое внимание уделяется работе с текстами на естественном языке, что является нетривиальной научной задачей.

В работе И.П. Рожнова и соавторов рассматривается актуальная задача повышения эффективности процедур отбора однородных объектов, в частности, партий промышленной продукции [4]. Для её решения предлагается использовать

специально разработанные гибридные алгоритмы кластеризации, объединяющие методы случайного поиска и локальной оптимизации.

Проведённые авторами вычислительные эксперименты продемонстрировали преимущества предложенного подхода в плане большей точности и стабильности результатов по сравнению с широко известными алгоритмами кластеризации (K-средних, PAM и др.). В данном контексте необходимо отметить, что, разработанные алгоритмы могут найти применение для решения широкого круга прикладных оптимизационных задач в промышленности и других областях.

Pitafi также рассматривает возможности применения кластеризации, но уже для целей сегментации потребителей в задачах маркетинга и планирования продаж, обосновывается возрастающую роль методов интеллектуального анализа данных, позволяющих извлекать ценные для бизнеса знания и закономерности[12].

На примере использования алгоритма K-средних показана эффективность кластеризации в группировке клиентов для последующей настройки таргетированных маркетинговых стратегий. Несомненно, применение подобных подходов способно существенно оптимизировать работу отделов маркетинга и продаж в компаниях. (Oyewole, Thopil) обобщают применение кластерного анализа в выбранных отраслях промышленности, важных для достижения целей устойчивого развития [11]. Рассматриваются компоненты кластеризации, проводится классификация алгоритмов. Обсуждаются традиционные методы и новые варианты, меры сходства, оптимизация и валидация кластеризации, особенности разных типов данных.

Wei с соавторами посвящают свою работу анализу относительно нового метода кластеризации – алгоритму плотностных пиков (Density Peaks Clustering, DPC) [13].

Данный подход использует информацию о локальной плотности точек и расстояний между точками для определения центров кластеров и последующего отнесения остальных точек к этим кластерам.

В работе подробно анализируются теоретические основы алгоритма DPC, даются его преимущества и недостатки. Затем рассматриваются различные модификации этого алгоритма, предложенные в последние годы для улучшения его эффективности.

Таким образом, в результате проведенного исследования, сформирована следующая таблица основы методов кластеризации данных в интеллектуальном анализе.

Таблица 1 - Основы методов кластеризации данных в интеллектуальном анализе

Алгоритм	Математическая основа	Принцип работы	Преимущества	Недостатки
Иерархический	Теория графов	Пошаговая оптимизация критерия связи кластеров	Визуализация, автоопределение числа кластеров	Экспоненциальная вычислительная сложность
K-средних	Минимизация дисперсии внутри кластеров	Итеративный пересчёт центроидов	Асимптотическая сходимость при выпуклых кластерах	Зависимость от инициализации
Плотностной	Оценки плотности вероятности	Поиск областей максимальной плотности	Устойчивость к выбросам	Квадратичная вычислительная сложность
Нечёткий C-средних	Теория нечётких множеств	Минимизация функционала нечёткой принадлежности	Робастность к шумам	Проблема определения числа и формы кластеров
Спектральный	Дискретная математика	Спектральные свойства матриц смежности	Не зависит от формы кластеров	Кубическая вычислительная сложность
SOM Кохонена	Нейронные сети	Самоорганизация сети по критерию схожести	Визуализация, устойчивость	Проблема интерпретации размерности
Модельный	Теория вероятностей	Параметрическая оптимизация правдоподобия моделей	Интерпретируемость	Зависит от выбора класса моделей
Графовый	Теория графов	Спектральные свойства графа смежности	Эффективность для связных данных	Требует построения графа

Проведенный в представленных публикациях анализ убедительно свидетельствует о высокой актуальности рассматриваемой проблематики в контексте современных задач интеллектуальной обработки данных. За прошедшие десятилетия накоплен обширный арсенал разнообразных подходов к решению задач кластеризации – как классических (иерархический анализ, методы частиционирования), так и относительно новых (спектральные методы, алгоритмы на основе графов, вероятностное моделирование).

Определены их теоретические основы, выявлены сильные и слабые стороны, области эффективного использования. Установлено, что выбор того или иного алгоритма кластеризации во многом зависит от природы и особенностей анализируемых данных.

Хотя в настоящее время нельзя выделить универсальный метод, одинаково хорошо решающий любые задачи, активно развиваются комплексные гибридные и ансамблевые подходы. Их применение позволяет получать более точные и устойчивые результаты за счёт сочетания преимуществ базовых алгоритмов.

Таким образом, несмотря на достигнутый прогресс, вопросы адекватного моделирования сложных распределений реальных данных, оптимального выбора числа кластеров, обеспечения масштабируемости алгоритмов остаются по-прежнему актуальными и требуют дальнейшего изучения с привлечением аппарата современной прикладной математики, статистики и искусственного интеллекта.

Результаты

Кластеризация позволяет выявлять внутреннюю структуру в кажущемся хаотичном наборе многомерных наблюдений, группируя похожие объекты и отделяя аномалии и выбросы. Без этого последующий анализ и принятие обоснованных решений крайне затруднены.

К настоящему времени разработан и апробирован весьма широкий спектр математических методов, позволяющих эффективно решать задачи кластеризации для данных различной природы. В их числе – древовидные иерархические алгоритмы, метод К-средних, плотностная кластеризация, спектральные методы группировки данных, нейросетевые технологии самоорганизующихся карт признаков и другие.

Как показывают приведенные в настоящей статье результаты исследований, каждый из этих методов имеет собственную область эффективного применения, недостатки и ограничения. Так, иерархические алгоритмы хорошо визуализируют скрытую структуру кластеров, но требуют значительных вычислительных ресурсов при больших объемах данных. Метод К-средних прост и быстр, однако нуждается в априорном задании числа групп.

Несмотря на достигнутые успехи в разработке алгоритмов интеллектуальной аналитики, ряд фундаментальных задач в части масштабирования, верификации и интерпретируемости результатов кластеризации больших данных остается нерешенным и требует пристального внимания исследователей.

Проблему представляет интерпретация и верификация результатов автоматической кластеризации экспертами предметной области. Как правило, получаемые группы данных весьма многомерны, что существенно усложняет их анализ, осмысление и принятие решений на их основе. Требуются новые методы визуализации, сравнения кластерных структур, оценки значимости и стабильности кластеров.

Вторая проблема заключается в обработке огромных и быстрорастущих объемов разнотипных данных, генерируемых на таких платформах. Например, известно, что лидирующие маркетплейсы накапливают десятки петабайт информации о товарах, заказах и покупателях. Применение традиционных алгоритмов кластеризации к таким данным крайне затруднительно или неэффективно ввиду ресурсоемкости обработки и масштабируемости.

Большинство популярных алгоритмов, таких как кластеризация методом k-средних, требуют ручной настройки этого ключевого параметра. При крупномасштабном анализе профилей клиентов, историй взаимодействия, поведения на сайте выбор неверного значения приводит к неадекватным, трудно интерпретируемым результатам.

В этой связи актуально развитие гибридных технологий, интегрирующих разные математические подходы с целью нейтрализации их недостатков и усиления достоинств. Перспективно также использование ансамблевых алгоритмов, комплексирующих результаты от нескольких базовых методов кластеризации.

Тем не менее ряд фундаментальных вопросов в рассматриваемой предметной области остается открытым и требует дальнейших исследований. В их числе – разработка эффективных критериев для автоматического определения оптимального числа кластеров в произвольной выборке объектов, создание высокопроизводительных распределенных алгоритмов кластеризации, способных обрабатывать постоянно увеличивающиеся массивы данных большой размерности.

Таким образом, проблематика интеллектуального анализа сложно структурированных данных методами кластеризации, несомненно, сохранит свою исключительную актуальность в обозримой перспективе, оставаясь предметом активных междисциплинарных исследований на стыке статистики, вычислительной математики и компьютерных наук.

Обсуждение результатов

Кластерный анализ является популярным методом интеллектуального анализа данных и машинного обучения, позволяющим группировать множество объектов таким образом, чтобы объекты в одном кластере были более схожи между собой, чем с объектами в других кластерах. Целью кластеризации является упорядочение и структурирование данных, выявление скрытых в них закономерностей.

Существует множество алгоритмов кластеризации, которые условно можно разделить на иерархические, методы частиционирования, плотностные, графовые и др. Практические примеры проведем на примере функционирования маркетплейсов. Во-первых, группировка ассортимента на однородные категории товаров с учётом различных параметров (цена, характеристики, отзывы, продажи) позволяет оптимизировать структуру каталога, облегчая поиск нужных товаров для покупателя.

Во-вторых, выявление групп похожих товаров открывает возможности для анализа конкурентной среды, определения перспективных ниш и выработки стратегии продвижения товарных групп.

В-третьих, исследование профилей и предпочтений онлайн-покупателей на основе поведенческих и транзакционных данных методами кластерного анализа является основой для разработки персонализированных маркетинговых стратегий.

Рассмотрим задачу разбиения 100 товаров каких-либо категорий (например, смартфонов) на 3 кластера по сходным характеристикам с использованием алгоритма k-средних.

Пусть каждый товар описан 4 параметрами: X1 - цена в условных единицах, X2 - средний рейтинг покупателей (целое от 1 до 5), X3 - число отзывов (целое), X4 - вес устройства в граммах.

На первом этапе произвольным образом выбираются 3 объекта в качестве начальных центров кластеров (μ1, μ2, μ3). Далее на каждой итерации пересчитываются центры как средние по объектам в кластере.

Например, после 3-й итерации получены следующие центры:

μ1 = (16000, 4.2, 124, 150)

μ2 = (50000, 4.7, 552, 180)

μ3 = (23000, 3.9, 46, 110)

Расстояние объектов до центров вычисляется по евклидовой метрике. Для объекта X = (22000, 4.1, 77, 130) оно составит:

d1 = √(22000 - 16000)2 + (4.1 - 4.2)2 + (77 - 124)2 + (130 - 150)2 = 9483

Аналогично вычисляются d2 и d3. Объект относится к ближайшему центру μ1.

Критерий останова итераций - стабилизация центров и состава кластеров. После останова оценивается качество разбиения, например, по индексу Дэвиса-Болдина:

DBI = 1/3 ∑i=1k∑x∈Si maxj≠i d(μi, μj)/σi

где σi - среднеквадратичное отклонение радиусов в i-м кластере. Меньшее DBI соответствует лучшей кластеризации.

Рассмотрим иной пример расчёта кластеризации методом k-средних для небольшой выборки данных об 8 товарах некого интернет-магазина. Каждый товар описан тремя характеристиками:

Цена в условных денежных единицах;

Средняя оценка в баллах (от 1 до 5);

Число отзывов покупателей.

Исходные данные выглядят следующим образом:

Таблица 2 - Исходные данные

Товар	Цена	Оценка	Отзывы
A	13 750	3	122
B	21 300	5	63
C	15 900	4	310
D	52 100	4	455
E	80 300	3	37
F	94 250	5	781
G	36 750	2	88
H	42 100	4	227

Зададим число кластеров равным 3. В качестве начальных центров кластеров произвольно выберем объекты A, D и F:

μ1 = (13 750, 3, 122)

μ2 = (52 100, 4, 455)

μ3 = (94 250, 5, 781)

На первой итерации посчитаем евклидовы расстояния от каждого объекта до текущих центроидов кластеров. К наиболее близкому центру и будет отнесён товар.

Например, для объекта E:

d1 = √(80 300 - 13 750)2 + (3 - 5)2 + (37 - 122)2 = 80 527

d2 = 34 196

d3 = 19 131

Таким образом, объект E попадает в кластер 3. Аналогично распределяются остальные товары.

После 1-й итерации состав кластеров:

A, G

B, C, H

D, E, F

Далее пересчитываем центры кластеров и повторяем распределение товаров. Процесс продолжается до стабилизации членства в кластерах.

Таблица 3 – Итог стабилизации членства товаров в кластерах

Товар	Цена	Оценка	Отзывы	Расстояние до центра 1	Расстояние до центра 2	Расстояние до центра 3	Кластер после итерации
A	13750	3	122	0	81048	2E+05	1
B	21300	5	63	60525	32261	1E+05	2
C	15900	4	310	44900	21904	1E+05	2
D	52100	4	455	1E+06	0	59553	2
E	80300	3	37	80527	34196	19131	3
F	94250	5	781	1E+05	50428	0	3
G	36750	2	88	58956	1E+05	2E+05	1
H	42100	4	227	1E+05	21176	1E+05	2

Построенная диаграмма наглядно демонстрирует расчёт расстояний от каждого объекта до текущих центров кластеров, определение принадлежности объектов к ближайшему центру, а также перераспределение объектов после пересчёта центроидов кластеров (на примере одной итерации). Результаты кластеризации методом k-средних приведены на рисунке. Диаграмма наглядно демонстрирует расчёт расстояний от каждого объекта до текущих центров кластеров, определение принадлежности объектов к ближайшему центру, а также перераспределение объектов после пересчёта центроидов кластеров (на примере одной итерации).

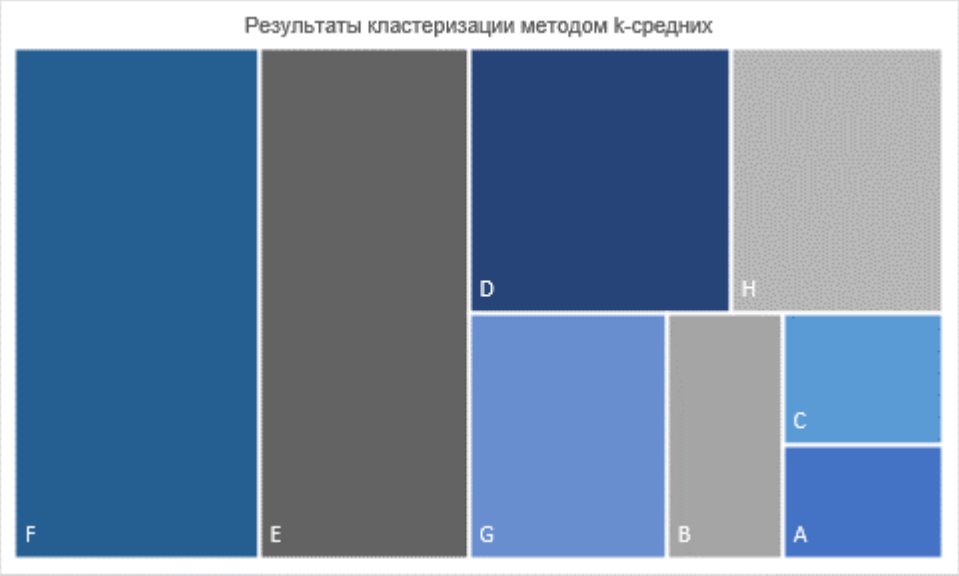

Рисунок 1 - Результаты кластеризации методом k-средних

Таким образом, технологии кластеризации данных предоставляют маркетплейсам в 2023 году актуальные инструменты для оптимизации работы с товарными предложениями. Эффективность работы достигается благодаря более компактным индексам после сжатия идентичных предложений.

Выводы

Применение процедур кластеризации позволяет выявлять скрытые в хаотичных массивах данных закономерности и знания, крайне востребованные в науке и бизнесе.

Как показано в работе, в арсенале исследователей уже находится обширнейший спектр самых разнообразных методов кластер-анализа, базирующихся на математическом аппарате статистики, теории искусственных нейронных сетей, дискретной математики, вероятностного моделирования и иных фундаментальных областях.

Детально проанализированы сильные и слабые стороны каждого из современных научно-методических подходов, а также определены области их наиболее эффективного применения в зависимости от природы целевых данных и характера решаемых аналитических задач. Очевидно, в условиях цифровой трансформации экономики и общества первостепенное значение приобретает создание высокопроизводительных распределённых алгоритмов кластеризации, позволяющих масштабно обрабатывать постоянно прирастающие объёмы разнотипной информации с использованием возможностей современных суперкомпьютерных технологий и облачной инфраструктуры.

Итак, несмотря на впечатляющие успехи в рассматриваемой области научного знания, остаётся открытым ряд фундаментальных проблем, требующих приоритетного внимания исследовательского сообщества. К их числу прежде всего следует отнести задачи разработки универсальных критериев автоматического определения оптимального числа кластеров в произвольном наборе многомерных наблюдений, повышения интерпретируемости и доверия к результатам интеллектуальной аналитики путём их адекватной визуализации и статистической верификации.

Резюмируя изложенное, стоит особо подчеркнуть, что методология кластерного анализа структурно сложных данных будет и впредь стремительно совершенствоваться, отвечая на нарастающий поток новой научной информации, знаний и соответствующих технологических вызовов, дальнейшие интенсивные усилия исследователей в этом междисциплинарном направлении уже в ближайшие годы принесут немало интереснейших результатов на стыке компьютерных и информационных наук, статистики, прикладной математики.

Литература

1. Исаков, С.С. Кластеризация и многоступенчатый анализ научных текстов / С.С. Исаков // Моделирование и анализ данных. – 2022. – Т. 12, No 4. – С. 105-109. URL: https://psyjournals.ru/journals/mda/archive/2022_n4/mda_2022_n4_Isakov.pdf

2. Махрусе Насма. Современные тенденции методов интеллектуального анализа данных: метод кластеризации // Московский экономический журнал. – 2019. – No 4. – С. 243-249. URL: https://cyberleninka.ru/article/n/sovremennye-tendentsii-metodov-intellektualnogo-analiza-dannyh-metod-klasterizatsii

3. Пуговкина, Е.Д. Использование методов кластеризации текстов на естественном языке в рекомендательных системах / Е.Д. Пуговкина, А.А. Белоусов // Информационные технологии и нанотехнологии. – 2022. – Т. 4. – С. 1022-1031. URL: http://repo.ssau.ru/bitstream/Informacionnye-tehnologii-i-nanotehnologii/Ispolzovanie-metodov-klasterizacii-tekstov-na-estestvennom-yazyke-v-rekomendatelnyh-sistemah-100180/1/ИТНТ-2022.%20Том%204.%20Искусственный%20интеллект/978-5-7883-1792-2_2022-041022.pdf

4. Рожнов, И.П. Повышение эффективности отбора однородных партий с использованием гибридных алгоритмов кластерного анализа / И.П. Рожнов, С.Н. Ежеманская, Л.А. Казаковцев, Е.Б. Козловская // Международный научно-

исследовательский журнал. – 2022. – No 10(124). – С. 95-100. URL: https://research-journal.org/archive/10-124-2022-october/10.23670/IRJ.2022.124.35

5. Харахинов, В.А. Нейросетевые технологии решения задач кластеризации и классификации данных в технических системах: дис. ... канд. техн. наук. – Иркутск, 2023. – 212 с. URL: https://www.irgups.ru/sites/default/files/oo/science/dissert%20sovet/dissertazii%20predsavlennyu%20k%20zashite/Харахинов%20Владимир%20Александрович/Полный%20текст%20диссертации%20Харахинов%20В.А..pdf

6. Хечми Шили. Кластеризация в аналитике больших данных: системный обзор и сравнительный анализ (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, No 5. С. 967–979. URL: https://ntv.ifmo.ru/ru/article/22369/%09klasterizaciya_v_analitike_bolshih_dannyh:_sistemnyy_obzor_i_sravnitelnyy_analiz_(obzornaya_statya).htm

7. Ширинкина, Е.В. Методы интеллектуального анализа данных и образовательной аналитики / Е.В. Ширинкина // Современное образование. – 2022. – No 1. – С. 51-67. URL: https://nbpublish.com/library_read_article.php?id=37582

8. Casanova, C. Hierarchical clustering-based framework for a posteriori exploration of Pareto fronts: application on the bi-objective next release problem / C. Casanova, E. Schab, L. Prado [et al.]. – 2023. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1179059/full

9. Gao, C.X. An overview of clustering methods with guidelines for application in mental health research / C.X. Gao, D. Dwyer, Y. Zhu [et al.] // Psychiatry Research. – 2023. – Vol. 327. – P. 115265. URL: https://www.sciencedirect.com/science/article/pii/S0165178123002159

10. Omran, M. An overview of clustering methods / M. Omran, A.P. Engelbrecht, A.A. Salman // Intelligent Data Analysis. – 2007. – Vol. 11, No 6. – P. 583–605. URL: https://www.researchgate.net/publication/220571682_An_overview_of_clustering_methods

11. Oyewole, G.J. Data clustering: application and trends / G.J. Oyewole, G.A. Thopil // Artificial Intelligence Review. – 2023. – Vol. 56, No 9. – P. 6439–6475. URL: https://link.springer.com/article/10.1007/s10462-022-10325-y

12. Pitafi, S. A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms / S. Pitafi, T. Anwar, Z. Sharif // Applied Sciences. – 2023. – Vol. 13, No 6. – P. 3529. URL: https://www.mdpi.com/2076-3417/13/6/3529

13. Wei, X. An overview on density peaks clustering / X. Wei, M. Peng, H. Huang [et al.] // Neurocomputing. – 2023. – Vol. 554. – P. 126633. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231223007567

14. Yang, Y. A generalized fuzzy clustering framework for incomplete data by integrating feature weighted and kernel learning / Y. Yang, H. Chen, H. Wu // PeerJ Computer Science. – 2023. – Vol. 9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10588703/

15. Zhang, M. Review of Clustering Methods for Functional Data / M. Zhang, A. Parnell // ACM Transactions on Knowledge Discovery from Data. – 2023. – Vol. 17, No 7. – P. 91. URL: https://dl.acm.org/doi/10.1145/3581789

References in Cyrillics

1. Isakov, S.S. Klasterizaciya i mnogostupenchatyj analiz nauchnyh tekstov / S.S. Isakov // Modelirovanie i analiz dannyh. – 2022. – T. 12, No 4. – S. 105-109. URL: https://psyjournals.ru/journals/mda/archive/2022_n4/mda_2022_n4_Isakov.pdf

2. Mahruse Nasma. Sovremennye tendencii metodov intellektual'nogo analiza dannyh: metod klasterizacii // Moskovskij ekonomicheskij zhurnal. – 2019. – No 4. – S. 243-249. URL: https://cyberleninka.ru/article/n/sovremennye-tendentsii-metodov-intellektualnogo-analiza-dannyh-metod-klasterizatsii

3. Pugovkina, E.D. Ispol'zovanie metodov klasterizacii tekstov na estestvennom yazyke v rekomendatel'nyh sistemah / E.D. Pugovkina, A.A. Belousov // Informacionnye tekhnologii i nanotekhnologii. – 2022. – T. 4. – S. 1022-1031. URL: http://repo.ssau.ru/bitstream/Informacionnye-tehnologii-i-nanotehnologii/Ispolzovanie-metodov-klasterizacii-tekstov-na-estestvennom-yazyke-v-rekomendatelnyh-sistemah-100180/1/ITNT-2022.%20Tom%204.%20Iskusstvennyj%20intellekt/978-5-7883-1792-2_2022-041022.pdf

4. Rozhnov, I.P. Povyshenie effektivnosti otbora odnorodnyh partij s ispol'zovaniem gibridnyh algoritmov klasternogo analiza / I.P. Rozhnov, S.N. Ezhemanskaya, L.A. Kazakovcev, E.B. Kozlovskaya // Mezhdunarodnyj nauchno-issledovatel'skij zhurnal. – 2022. – No 10(124). – S. 95-100. URL: https://research-journal.org/archive/10-124-2022-october/10.23670/IRJ.2022.124.35

5. Harahinov, V.A. Nejrosetevye tekhnologii resheniya zadach klasterizacii i klassifikacii dannyh v tekhnicheskih sistemah: dis. ... kand. tekhn. nauk. – Irkutsk, 2023. – 212 s. URL: https://www.irgups.ru/sites/default/files/oo/science/dissert%20sovet/dissertazii%20predsavlennyu%20k%20zashite/Harahinov%20Vladimir%20Aleksandrovich/Polnyj%20tekst%20dissertacii%20Harahinov%20V.A..pdf

6. Hechmi SHili. Klasterizaciya v analitike bol'shih dannyh: sistemnyj obzor i sravnitel'nyj analiz (obzornaya stat'ya) // Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki. 2023. T. 23, No 5. S. 967–979. URL: https://ntv.ifmo.ru/ru/article/22369/%09klasterizaciya_v_analitike_bolshih_dannyh:_sistemnyy_obzor_i_sravnitelnyy_analiz_(obzornaya_statya).htm

7. SHirinkina, E.V. Metody intellektual'nogo analiza dannyh i obrazovatel'noj analitiki / E.V. SHirinkina // Sovremennoe obrazovanie. – 2022. – No 1. – S. 51-67. URL: https://nbpublish.com/library_read_article.php?id=37582

8. Casanova, C. Hierarchical clustering-based framework for a posteriori exploration of Pareto fronts: application on the bi-objective next release problem / C. Casanova, E. Schab, L. Prado [et al.]. – 2023. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1179059/full

9. Gao, C.X. An overview of clustering methods with guidelines for application in mental health research / C.X. Gao, D. Dwyer, Y. Zhu [et al.] // Psychiatry Research. – 2023. – Vol. 327. – P. 115265. URL: https://www.sciencedirect.com/science/article/pii/S0165178123002159

10. Omran, M. An overview of clustering methods / M. Omran, A.P. Engelbrecht, A.A. Salman // Intelligent Data Analysis. – 2007. – Vol. 11, No 6. – P. 583–605. URL: https://www.researchgate.net/publication/220571682_An_overview_of_clustering_methods

11. Oyewole, G.J. Data clustering: application and trends / G.J. Oyewole, G.A. Thopil // Artificial Intelligence Review. – 2023. – Vol. 56, No 9. – P. 6439–6475. URL: https://link.springer.com/article/10.1007/s10462-022-10325-y

12. Pitafi, S. A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms / S. Pitafi, T. Anwar, Z. Sharif // Applied Sciences. – 2023. – Vol. 13, No 6. – P. 3529. URL: https://www.mdpi.com/2076-3417/13/6/3529

13. Wei, X. An overview on density peaks clustering / X. Wei, M. Peng, H. Huang [et al.] // Neurocomputing. – 2023. – Vol. 554. – P. 126633. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231223007567

14. Yang, Y. A generalized fuzzy clustering framework for incomplete data by integrating feature weighted and kernel learning / Y. Yang, H. Chen, H. Wu // PeerJ Computer Science. – 2023. – Vol. 9. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10588703/

15. Zhang, M. Review of Clustering Methods for Functional Data / M. Zhang, A. Parnell // ACM Transactions on Knowledge Discovery from Data. – 2023. – Vol. 17, No 7. – P. 91. URL: https://dl.acm.org/doi/10.1145/3581789

Кучумов Илья Вадимович - Руководитель отдела разработки
(Head of development department) в ООО «Яндекс»
Kuchumov.ilya@gmail.com

Ключевые слова:
Интеллектуальный анализ данных, кластеризация данных, интерпретация данных, плотностная кластеризация, иерархическая кластеризация, Big Data, распределенные вычисления, алгоритм k-средних, нейронные сети.

Kuchumov Ilya Vadimovich, Theoretical foundations of data clustering methods in intelligent analysis

Annotation

This article examines the theoretical foundations and principles of the main data clustering methods used in intelligent data analytics tasks. The paper provides an in-depth analysis of the research works of several scholars in the field of cluster analysis; it describes the progress in developing and applying classical and advanced approaches to grouping structurally complex, heterogeneous data using statistical techniques, neural networks, and mathematical modeling. The mathematical basics of hierarchical, probabilistic, density-based, graph-based and other clustering methods are considered, the effectiveness of their application to different data types is theoretically proven, depending on the analytical goals. Particular attention is paid to the clustering of large volumes of heterogeneous information in the context of increasing data flow rates and demands for the timeliness of their processing. The potential of hybrid neural network and distributed clustering methods for efficient scalable Big Data analysis using high-performance computing systems is demonstrated. It is shown that, despite significant advances, a number of fundamental issues in this area remain open and require further interdisciplinary research at the intersection of statistics, mathematics and computer science.

Keywords: Intelligent data analysis, data clustering, data interpretation, density-based clustering, hierarchical clustering, Big Data, distributed computing, k-means algorithm, neural networks.