На правах рукописи

АХМАТШИН ФАРИД ГАЛИУЛЛОВИЧ

МОДЕЛИ И АЛГОРИТМЫ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОБЪЕКТОВ ДЛЯ СИСТЕМ АНАЛИЗА И ХРАНЕНИЯ ДАННЫХ НА ОСНОВЕ МЕТОДОВ СЕМЕЙСТВА *k*-СРЕДНИХ

2.3.1 – Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2025

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева», г. Красноярск

Научный руководитель: доктор технических наук, профессор
Казаковцев Лев Александрович

Официальные оппоненты: **Кравец Олег Яковлевич,**
доктор технических наук, профессор,
ФГБОУ ВО «Воронежский государственный технический университет», г. Воронеж,
профессор кафедры «Автоматизированные и вычислительные системы»

Муравьёв Сергей Борисович,
кандидат технических наук,
ФГАОУ ВО «Национальный исследовательский университет информационных технологий, механики и оптики», г. Санкт-Петербург, доцент института прикладных компьютерных наук

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Кемеровский государственный университет»

Защита состоится «28» марта 2025 года в 15:30 часов на заседании диссертационного совета 24.2.403.01, созданного на базе ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» по адресу: 660037, г. Красноярск, пр. им. газеты «Красноярский рабочий» 31, зал заседаний диссертационного совета, ауд. Л-205

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» и на сайте https://www.sibsau.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью, просим отправлять по адресу: 660037, Россия, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» (СибГУ им. М.Ф. Решетнева), Диссертационный совет E-mail: dissovet@sibsau.ru

Автореферат разослан «___» _____ 2025 г.

Ученый секретарь
диссертационного совета Панфилов Илья Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Прогнозируемый объем данных, накапливаемых человечеством, с каждым годом увеличивается в геометрической прогрессии. Современное и будущее развитие технологий анализа и хранения информации связано, в том числе, с использованием методов автоматической группировки (кластеризации) данных и приближенного поиска ближайших соседей (ANN – Approximate Nearest Neighbors) как ценного источника новых знаний о структуре хранимой информации. Автоматическое получение информации о сходстве и компактности (упорядоченности и однородности) данных стимулирует процесс получения новой информации в области анализа и хранения данных.

Задачи автоматической группировки (кластеризации) применяются в быстроразвивающихся сферах деятельности, например при анализе данных неразрушающих тестовых испытаний промышленной продукции, в системах архивации данных в системах хранения данных, в векторных базах данных.

Традиционные методы приближенного поиска ближайших соседей, в том числе с применением локально-чувствительного хэширования (Locality Sensitive Hashing, LSH), снижают вычислительную сложность решения NP-трудных задач автоматической группировки. В таких задачах кластерная структура упорядоченных однородных наборов данных определяется путем оптимизации целевой функции с учетом меры подобия (различия), которую выражают, в том числе, с использованием функции расстояния. При этом способы предварительной подготовки данных, в частности нормализация числовых показателей, иногда имеют решающее значение.

Способы обработки информации (например, кластеризации) исходных данных не всегда применимы напрямую. Для унификации шкал величин исходные данные стандартизируют, применяя затем расстояние Евклида или квадрат расстояния. Выявление шума в исходных данных и обнаружение «выбросов» – одно из требований к методам нормализации исходных данных; при этом специфические задачи требуют особых способов предобработки данных. Многообразные методы кластерного анализа используют различные способы предобработки информации, комбинируя применение различных метрик, методов вычисления координат и расстояний для поиска оптимальных решений, дающих приемлемое решение для большинства практических задач.

С одной стороны, развитие систем анализа данных требует разработки адекватных методов обработки и предобработки данных (в том числе, нормализации), новых моделей и алгоритмов автоматической группировки для той или иной предметной области. С другой стороны, вследствие увеличения объемов обрабатываемых и анализируемых данных усложняется внутренняя структура систем анализа данных, включая подсистемы хранения данных, требуя применения алгоритмов машинного обучения, в том числе алгоритмов кластерного анализа, для повышения эффективности управления в системах хранения данных (дисковые массивы, векторные базы данных). При этом увеличение объема обрабатываемых данных выявляет низкую вычислительную

эффективность существующих алгоритмов. Решению задач в этих областях и посвящена данная работа.

Степень разработанности темы. С предложенной Г. Штейнгаузом реализацией алгоритма С. Ллойда для задачи *k*-средних связано наиболее популярное направление развития методов автоматической группировки данных. Первоначальное название предиктора пакетирования для описания подхода к автоматической кластеризации, открытое Г. Штейнгаузом, приобрело в дальнейшем название *k*-средних, которое ввел Дж. Маккуин. Позже независимо Эдвард У. Форги опубликовал тот же метод. Дальнейшее развитие популярной алгоритмической реализации алгоритма *k*-средних связано с усовершенствованием простых шагов инициализации, классификации и разбиения на кластеры до конвергенции, из которых состоит данный алгоритм. Методы инициализации начальных центров на основе случайного выбора впервые были предложены одновременно с началом использования алгоритма *k*-средних в методах Форги, Спата, и Маккуина. Развитие данных идей продолжено в методе инициализации *k*-means++ Д. Артуром и С. Василвицким. Дальнейшее развитие популярной алгоритмической реализации связано с модернизацией алгоритма *k*-means++.

Настоящая диссертация направлена на разработку новых алгоритмов автоматической группировки, используемых в системах анализа и хранения данных, на повышение эффективности алгоритмов кластеризации при обработке больших данных в системах автоматической группировки объектов, в том числе в составе векторной СУБД и подсистем компрессии данных в составе систем хранения данных.

Объектом диссертационного исследования являются задачи автоматической группировки объектов для систем анализа и хранения данных, **предметом исследования** – являются алгоритмы для решения данных задач.

Целью исследования является повышение эффективности (вычислительной производительности, а также повышения качества результата по внешним и внутренним критериям качества) алгоритмов автоматической группировки объектов для систем анализа и хранения данных.

Задачи, решаемые в процессе достижения поставленной цели:

1. Разработать подход к нормализации данных для предобработки входных данных, используемых в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции, комбинирующий нормализацию по допустимым значениям параметров оцениваемых значений продукции и оценке Джеймса-Штейна.
2. Разработать алгоритм кластеризации для системы анализа данных электрорадиоизделий на основе данных тестовых испытаний с использованием жадной эвристической процедуры выбора радиуса локальных концентраций по размеченным данным.
3. Разработать алгоритм кластеризации для создания индекса векторной базы данных предназначенного для построения индекса приближенного поиска ближайших соседей, как компромисс между точностью и временем

вычислений, значительно улучшающий метрику полноты в задачах приближенного поиска ближайших соседей.

4. Разработать алгоритм автоматической группировки повторяющихся фрагментов блоков данных для использования в системах хранения данных на основе алгоритма *k*-средних совместно с применением локально-чувствительного хэширования LSH.

5. Разработать процедуру инициализации центров кластеров для алгоритмов кластеризации, способную быстро находить приемлемое начальное решение при большом объеме данных.

Новые научные результаты, выносимые на защиту:

1. Предложен новый подход к нормализации данных для предобработки входных данных, используемых в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции, комбинирующий нормализацию по допустимым значениям параметров оцениваемых характеристик продукции и оценки Джеймса–Штейна.

2. Предложен новый алгоритм кластеризации для системы анализа данных электрорадиоизделий на основе данных тестовых испытаний с использованием жадной эвристической процедуры выбора радиуса локальных концентраций по размеченным данным.

3. Предложен новый алгоритм кластеризации для построения индекса векторной базы данных для приближенного поиска ближайших соседей, обеспечивающий компромисс между точностью и временем вычислений, существенно улучшает метрику полноты в задачах приближенного поиска ближайших соседей.

4. Предложен новый алгоритм автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма *k*-средних совместно с локально-чувствительным хэшированием (LSH), обеспечивающий увеличение эффективности сжатия данных в системах хранения данных.

5. Предложена новая процедура инициализации центров кластеров для алгоритмов автоматической группировки больших объемов данных, использующая вспомогательную структуру данных – массив слагаемых для вычисления суммы квадратов расстояния.

Значение для теории. Теоретическая значимость состоит в дополнении эффективных алгоритмов решения задач автоматической группировки, а также алгоритмов предобработки данных для таких задач.

Практическая ценность состоит в дополнении модельно-алгоритмического инструментария, используемого в системах анализа данных результатов тестирования образцов промышленной продукции с повышенными требованиями качества, в частности – электронной компонентной базы космического применения, и могут использоваться в соответствующих испытательных технических центрах. Кроме того, новая процедура инициализации центров кластеров для алгоритмов кластеризации, имеет универсальный характер и может применяться при обработке больших данных в любых системах автоматической группировки объектов. Новые алгоритмы кластеризации применяются для построения индекса для векторной базы

данных и для разработки модели оптимального использования дискового пространства с учетом компрессии данных.

Методы исследования. Результаты прикладных и теоретических задач получены с применением методов системного анализа, исследования операций, теории размещения, теории оптимизации.

Положения, выносимые на защиту:

1. Подход к нормализации данных для предобработки входных данных, используемых в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции, комбинирующий нормализацию по допустимым значениям параметров оцениваемых характеристик продукции и оценки Джеймса–Штейна, обеспечивает повышение точности решения задачи кластеризации на 10% по индексу Рэнда, на примере задачи автоматической группировки электрорадиоизделий.
2. Алгоритм кластеризации для системы анализа данных электрорадиоизделий на основе данных тестовых испытаний с использованием жадной эвристической процедуры выбора радиуса локальных концентраций по размеченным данным обеспечивает повышение точности (по индексу Рэнда) и скорости получения результатов автоматической группировки по сравнению с алгоритмом *k*-средних.
3. Алгоритм кластеризации для построения индекса векторной базы данных для приближенного поиска ближайших соседей обеспечивает компромисс между точностью и временем вычислений, существенно улучшает метрику полноты в задачах приближенного поиска ближайших соседей.
4. Алгоритм автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма *k*-средних совместно с локально-чувствительным хэшированием (LSH), обеспечивает увеличение эффективности сжатия данных в системах хранения данных.
5. Процедура инициализации центров кластеров для алгоритмов автоматической группировки больших объемов данных, использует вспомогательную структуру данных – массив слагаемых для вычисления суммы квадратов расстояния, обеспечивает снижает вычислительные затраты в сравнении с алгоритмом k-means++ без снижения качества получаемого начального решения.

Практическая реализация результатов: программная разработка алгоритма автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма *k*-means, совместно с локально-чувствительным хэшированием (LSH) предназначена для использования в системах хранения данных. Она реализована в разработанной модели оптимального использования дискового пространства с учетом компрессии данных в рамках работ по договору №20769 от 05.10.2023 г. Исследование по разработке нового алгоритма кластеризации, основанного на жадной агломеративной процедуре, для построения индекса для векторной базы данных выполнено по договору № ТС2024051430 от 14.06.2024г. Исследование по разработке новой процедуры инициализации центров кластеров для алгоритмов кластеризации,

применяемых к большим данных выполнено при поддержке Министерства науки и высшего образования Российской Федерации (проект FEFE-2023-0004).

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на международных семинарах и конференциях: «Решетневские чтения» (2020 г., г. Красноярск), Mathematical Optimization Theory and Operations Research (MOTOR, 2023 – 2024 гг., г. Екатеринбург, г. Омск), семинар «Математические модели принятия решений» Институт математики имени С.Л.Соболева, (2024 г., г.Новосибирск).

Публикации. Основные теоретические и практические результаты диссертации содержатся в 14 публикациях, среди которых 5 работ в ведущих рецензируемых журналах, рекомендуемых в действующем перечне ВАК, 5 – в международных изданиях, индексируемых в системах цитирования Web of Science и Scopus. Имеется свидетельство о государственной регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, 5 разделов и заключения и приложения. Она изложена на 131 листах машинописного текста, содержит список литературы из 245 наименований.

Основное содержание работы

Во введении обоснована актуальность, поставлена цель и указаны задачи исследования, научная новизна и практическая значимость работы, изложены методы исследования, сформулированы основные положения, выносимые на защиту.

Первый раздел посвящен вопросам нормализации данных в задаче автоматической группировки промышленной продукции на примере электрорадиоизделий (ЭРИ) по однородным производственным партиям, основанной на модели *k*-средних, а также разработке нового подхода по нормализации данных промышленной продукции, комбинирующего нормализацию по допустимым значениям параметров оцениваемых характеристик продукции и оценку Джеймса-Штейна.

Контроль качества промышленной продукции требует получения наиболее точного и стабильного результата в отрасли с высокими требованиями к качеству продукции. Точность относится к снижению доли ошибок автоматической группировки, а стабильность – к повторяемости результатов при многократном запуске алгоритма.

Промышленная продукция должна быть изготовлена из одной партии сырья (пластин), что не гарантируется для устройств, используемых в космической промышленности. Различные алгоритмы кластеризации реализуются на основе многомерных результатов тестирования для решения задачи обнаружения однородных партий продукции.

В ранних исследованиях сравнивались различные способы нормализации для решения задачи автоматической группировки ЭРИ по однородным производственным партиям, основанной на модели *k*-средних. Производилось сравнение различных способов нормализации данных в решении задачи *k*-средних, изучались результаты точности кластеризации, и был предложен

новый способ нормализации по допустимым значениям параметра многомерных данных.

Использование многомерных данных в модели автоматической группировки влияет на точность решения задачи. Между некоторыми характеристиками часто существуют явные корреляции. Задача кластеризации нормализованных данных эффективно решается с помощью алгоритма *k*-средних со специальными мерами расстояния, в которых учитываются эти зависимости.

В исходных многомерных данных могут присутствовать выбросы, которые влияют на интерпретацию полученных результатов. Для снижения влияния выбросов применяется нормализация значений исходных данных.

Методы нормализации оказывают большое влияние на решение задачи автоматической группировки изделий промышленной продукции по однородным производственным партиям. При этом каждый из продуктов имеет большое количество входных характеристик. Характеристики с такими диапазонами значений также оказывают значительное влияние на группировку продуктов. В связи с этим возникает задача выработки такого подхода к нормализации данных в результате тестовых испытаний образцов промышленной продукции, который бы обеспечивал повышение точности решения задачи выделения однородных партий промышленной продукции.

Пусть набор данных $D = \{x_1, \ldots, x_N\}$ состоит из N точек, обозначим центры, полученные после применения автоматической группировки *k*-средних $c_1, \ldots, c_j, \ldots, c_k$. Цель алгоритма автоматической группировки состоит в том, чтобы сумма квадратов расстояний от известных точек до ближайших центров достигла своего минимума:

$$argminF(c_1, \ldots, c_j, \ldots, c_k) = \sum_{i=1}^{n} min_{j \in \{1,k\}} \|x_i - c_j\|^2. \quad (1)$$

Предлагаемый в настоящей работе подход заключается в использовании улучшенной оценки усадки Джеймса–Штейна для уменьшения влияния неинформативных параметров, нормализованных по диапазону допустимых значений параметров исходных данных $x_1, \ldots, x_N$, с использованием диапазона допустимых значений A, зависящего от контролируемых характеристик продукции для соответствующих режимов испытаний $x` = \frac{(x - x_{min})}{A}$, где $x`$ – вектор нормализованных значений измеряемого параметра.

Для минимизации значения целевой функции (1) используем средние значения точек в кластере относительно ближайших центров. Для улучшения средних значений $\|x`_i - c_j\|$ в кластере мы используем оценку Джеймса-Штейна от x_i до среднего значения μ_j всего набора данных:

$$x_{JS} = \left(1 - \frac{(p-2)t\sigma^2}{\|x`_i - \mu\|^2}\right)^{+} (x`_i - \mu) + \mu, \quad (2)$$

где $\mu = (\mu_1, \ldots, \mu_p)$ – среднее значение параметра $x`_i$; $\mu = \frac{1}{N}\sum_{i=1}^{N} x`_i$ – среднее значение выборки; $(\quad)^+$ – оценка с положительной частью, равной нулю при отрицательных значениях; σ^2 –дисперсия значений $x`$; $t$ – коэффициент сокращения оценки усадки к некоторому значимому ненулевому значению. Улучшенная усадка оценки Джеймса-Штейна способствует уменьшению среднего значения $\|x`_i - c_j\|$ в кластере для минимизации значения целевой функции (1) за счет применения коэффициента *t* уменьшения оценки Джеймса-Штейна.

Сравнительный анализ эффективности нового подхода с использованием нормализации по стандартному отклонению (A), с использованием нормализации по значениям допустимого дрейфа (B), с использованием нормализации по допустимым значениям параметров (C), приведено в таблице 1 и на рисунке 1.

Таблица 1 – Сравнительные результаты кластеризации для ИС 140УД26А с использованием нормализации по допустимым значениям параметров с коэффициентом t уменьшения оценки Джеймса-Штейна

JS	Значение целевой функции							Индекс Рэнда				
t	Макси-маль-ное	Мини-маль-ное	Сред-нее	Меди-анное	Стан-дарт-ное откло-нение	Вари-ация	Раз-мах	Макси-маль-ное	Мини-маль-ное	Сред-нее	Меди-анное	Стан-дарт-ное откло-нение
0	34,203	26,052	26,844	26,058	1,546	0,058	8,152	0,611	0,584	0,603	0,586	0,006
30	0,210	0,126	0,149	0,126	0,020	0,136	0,084	0,957	0,788	0,879	0,793	0,049
100	0,552	0,430	0,479	0,430	0,041	0,085	0,122	0,901	0,647	0,733	0,664	0,056

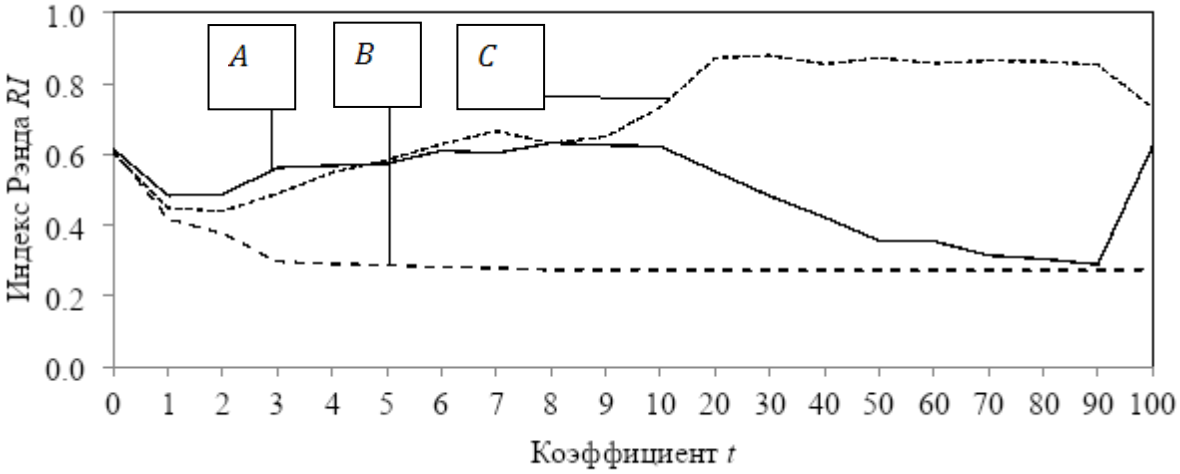

Рисунок 1 – Сравнение значений индекса Ренда *RI* в результате кластеризации наборов данных *A, B, C* для интегральной схемы 140УД26А с коэффициентом *t* сокращения оценки усадки Джеймса–Штейна

Таким образом, применение предложенного подхода к нормализации для предобработки входных данных, используемых в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции, кроме повышения точности кластеризации, позволяет выявлять наиболее

значимые контролируемые параметры продукции, что в перспективе позволит сократить затраты на испытания за счет уменьшения числа контролируемых параметров.

Второй раздел посвящен разработке алгоритма кластеризации электрорадиоизделий (решается задача разделения смешанной партии ЭРИ на однородные партии продукции) по результатам неразрушающих тестов с жадной эвристической процедурой выбора радиуса локальных концентраций по размеченным данным. Нами рассматривается задача поиска радиуса локальных концентраций (сгущений) в алгоритме кластеризации с заранее заданным числом кластеров. Результаты сравнивались с различными способами нормализации данных для решения задачи кластеризации методом *k*-средних.

Процесс кластеризации заключается в группировке *n* наблюдений в *k* групп. Репрезентативный центр задается с помощью *M*-мерного вектора. Оптимальная кластеризация достигается при минимальном значении целевой функции. Учитывая набор наблюдений $x_1, \dots, x_i, \dots, x_n$, которые разделены на *k* кластеров $C = \{C_1, \dots, C_j, \dots, C_k\}$, целевая функция достигает своего минимума:

$$argmin \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2, \quad (3)$$

где μ_j – среднее значение наблюдений в кластере C_j. Кластеры, полученные алгоритмом *k*-средних, стремятся к сферической форме.

Алгоритмы семейства *k*-средних основаны на мере похожести (сходства, близости) на центр оцениваемой между точками в многомерном пространстве через расстояние Евклида. В работах Загоруйко Н.Г. и др. сформулированы требования локальности, нормированности и инвариантности, которым должна удовлетворять мера сходства. Кроме того, эти алгоритмы основаны на использовании гипотезы компактности, по которой точки относятся к одному кластеру, если в пространстве они расположенные рядом. Подход, предложенный в работах Загоруйко Н.Г. основанный на использовании конкурентного сходства FRiS-функции, следует данным требованиям и гипотезе, и аналогичен в этом отношении алгоритму Ллойда. Использование FRiS-компактности в алгоритме GRAD используется для выбора информативных параметров (признаков).

Работы Федосова В.И., Шкабериной Г.Ш., Голованова С.М. и др. показывают, что, хотя в зависимости от условий производства параметры однотипных изделий различных партиях могут существенно различаться, изделия в составе одной партии имеют сходство параметров, благодаря чему партии можно разделить. В алгоритмах FRiS кластеризация осуществляется с использованием расстояния от центра кластера, который назовем радиусом локальных концентраций, определяет предельный совокупный разброс значений параметров элементов кластера, т.е. предельный разброс значений параметров изделий в партии, являясь дополнительной характеристикой получаемого результата автоматической группировки, важный для эксперта, принимающего заключение о составе исследуемой партии. При этом такой радиус в практических задачах заранее неизвестен.

В общей постановке задачи алгоритм автоматической группировки ищет такое разбиение n объектов на k кластеров, чтобы критерий похожести F на центр для всех кластеров был минимальным. Обозначим координаты центра j-го кластера символом C_j. Сумма расстояний $p(C_j, a_i)$ между центром сферы с радиусом R_0 и всеми n_j точками a_i этого кластера $p_j = \sum p(C_j, a_i)$, где $i = 1, \dots, n_j$, а сумма таких внутренних расстояний для всех k кластеров $F = \sum p_j$, $j = 1, \dots, k$.

Для нового алгоритма LEES (Localized Estimation Exploratory Search – локального оценочного исследовательского поиска) функция качества кластеризации F использует дополнительный критерий f, соответствующий единице при равенстве искомого числа кластеров k и найденного числа кластеров K

$$F = argmin\, f(K, p_j) \sum_{j=1}^{k} p_j \begin{cases} f(K) = 1, \text{ если } K = k, \\ f(K) = \infty, \text{ если } K \neq k, \\ i \epsilon C_j, \text{ если } p_j(x_j, a_i) \leq R_0 R. \end{cases} \quad (4)$$

где R – коэффициент радиуса поиска локальных концентраций, C_j – j-ый кластер.

Условие присоединения точки к кластеру a_i определяем по формуле

$$p_j(x_j, a_i) \leq R_0 R, \quad (5)$$

Новый алгоритм LEES состоит из следующих шагов:

Алгоритм 1. LEES

Дано: Задаем точное число кластеров k сферической формы с радиусом R_0. Для вычисления расстояний между объектами применяем метрику Евклида.

Шаг 1. Случайно выбираем первый центр C_1, а следующие центры на удалении R_0 заданного радиуса с коэффициентом R от каждого центра по формуле (5), для улучшения инициализации центров C_j.

Шаг 2. Присваиваем последовательно точки ближайшему центру C_j. Для минимизации целевой функции (4) используем метод жадного выбора ближайшего центра.

Шаг 3. Пересчитываем новые центры для каждого кластера.

Повторять Шаги 2 и 3 пока изменения целевой функции не стабилизируются.

Для алгоритма LEES число кластеров должно быть заранее задано. Во всех вариациях алгоритма LEES лежит базовая процедура, в основном совпадающая с алгоритмом k-средних. Шаг 1 отличается тем, что улучшается инициализация центров кластеров.

Средние результаты кластеризации по скорости работы нового алгоритма k-средних (1) и LEES (2) приведены в таблице 2 для индекса Рэнда (RI) и значения целевой функции (F). Для каждого алгоритма мы провели 30 экспериментов. На каждый запуск отводилось предельное фиксированное время - 1, 10 или 60 секунд.

Использование жадной эвристики для выбора радиуса поиска локальных концентраций в новом алгоритме кластеризации электрорадиоизделий на основе данных тестовых испытаний с точно заданным числом кластеров имеет преимущество по скорости точной кластеризации в сравнении с алгоритмом *k*-средних, использующего жадную эвристику для выбора центров, при использовании нормализованных значений тестовых испытаний по стандартному отклонению и по допустимым значениям параметра.

Таблица 2 – Точность кластеризации 1, 3, 8 и 9 партии данных ИС 140УД25А алгоритмами *k*-средних и LEES с ограничением по времени 1 секунда

Смешанная партия количеством n, с нормализацией по стандартному отклонению											
Алгоритм	Индекс Рэнда				Значения целевой функции						Кол-во векторов данных n
	максимальное	минимальное	среднее	стандартное отклонение	максимальное	минимальное	среднее	стандартное отклонение	вариация	размах	
1	0,605	0,550	0,596	0,010	85,1	66,9	69,4	3,5	0,051	18,2	201
2	**0,794**	**0,605**	**0,712**	0,052	85,0	76,1	81,1	2,5	0,030	9,0	201
1	0,600	**0,557**	0,584	0,013	326,5	273,0	296,5	14,6	0,049	53,5	807
2	**0,760**	0,539	**0,669**	0,063	411,6	325,0	365,0	21,1	0,058	86,5	807
1	0,646	**0,588**	0,613	0,016	64,7	57,8	59,6	1,5	0,026	6,8	132
2	**0,747**	0,538	**0,684**	0,046	68,4	63,7	66,0	1,4	0,021	4,7	132
1	0,654	**0,583**	0,621	0,020	251,3	235,8	242,7	4,2	0,017	15,5	532
2	**0,738**	0,559	**0,676**	0,047	308,7	269,0	286,9	11,1	0,039	39,7	532
Смешанная партия количеством n, с нормализацией по допустимым значениям параметра											
1	0,603	0,548	0,594	0,010	51,8	40,0	42,0	2,4	0,057	11,8	201
2	**0,809**	**0,556**	**0,680**	0,059	58,9	48,9	54,6	2,2	0,041	10,0	201
1	0,601	**0,532**	0,580	0,019	214,0	165,5	184,6	15,1	0,082	48,4	807
2	**0,784**	0,349	**0,620**	0,104	329,4	223,6	260,2	24,4	0,094	105,8	807
1	0,613	0,576	0,606	0,010	34,4	26,1	27,7	1,8	0,063	8,3	132
2	**0,752**	**0,623**	**0,703**	0,031	37,1	32,8	34,4	1,1	0,031	4,2	132
1	0,615	**0,556**	0,604	0,013	137,4	106,1	113,1	6,4	0,057	31,3	532
2	**0,748**	0,545	**0,686**	0,042	175,4	138,3	153,0	8,6	0,056	37,2	532

Согласно полученным значениям индекса Рэнда, подход с использованием алгоритма LEES с жадной эвристикой выбора радиуса поиска локальных концентраций продемонстрировал наилучшую точность при большем значении целевой функции в сравнении с алгоритмом *k*-средних.

Точность и скорость работы программной реализации алгоритма вполне приемлемы для решения задачи кластеризации системы анализа данных электрорадиоизделий на основе данных тестовых испытаний. Использование жадной эвристики в алгоритме LEES показало большую эффективность по сравнению с классической реализацией алгоритма *k*-средних.

Третий раздел посвящен разработке нового алгоритма кластеризации, основанного на жадной агломеративной процедуре для построения индекса для векторной базы данных. В вычислительных экспериментах показано, что индекс (набор центров), сгенерированный этим алгоритмом, достигает более

высокой оценки полноты по сравнению с алгоритмами *k*-средних и *k*-means++.

Современные инструменты сбора данных позволяют накапливать большие объемы многомерной информации об объекте исследования. Обработка данной информации с использованием соответствующих методов и алгоритмов интеллектуального анализа становится ценным источником знаний об объекте исследования. Поиск ближайших соседей (NNS) — это задача оптимизации, которая заключается в поиске вектора, наиболее похожего на заданный вектор (вектор запроса) в исходном наборе данных.

Проблема поиска центральных точек является классической проблемой размещения, которая может быть сформулирована как задача *p*-медианы или как задача *k*-средних, предложенная Г. Штейнгаузом.

Для решения обеих задач используется классический алгоритм — алгоритм *k*-средних, также известный как алгоритм Ллойда (Lloyd) или алгоритм альтернативного размещения-распределения (ALA). Этот метод начинается с начального решения, состоящего из набора центров или центроидов $c_1, \ldots, c_k$, которые итеративно обновляются с помощью двух чередующихся шагов:

Шаг 1: Назначить каждую точку данных ближайшему центру кластера (центроиду), создавая новые кластеры C_j .

Шаг 2: Пересчитать центры кластеров на основе вновь сформированных кластеров.

Шаги 1 и 2 повторяются до тех пор, пока изменения внутри каждого кластера не стабилизируются, и не произойдет никаких дальнейших существенных изменений.

Мы предлагаем алгоритм, основанный на жадной агломеративной процедуре для построения индекса для векторной базы данных, с использованием модели *k*-средних или *p*-медианной. Этот алгоритм несколько проще, чем ранее используемый эволюционный алгоритм с жадной агломеративной процедурой BasicAggl Алгоритм 5 с операторами кроссовера.

Алгоритм 2. BasicAggl(C)

Требуется:

Набор начальных центров $C = \{X_1, \ldots, X_K\}$,

где k – требуемое число центров, $K > k$,.

Шаг 1. Применить алгоритм Ллойда к набору $C \leftarrow Lloyd(C)$;

Шаг 2. Пока $|C| > k$ выполнять

Шаг 2a. Для $i = 1, \ldots, K$ выполнять $F_i \leftarrow F(C \backslash \{X_i\})$.

Шаг 2b. Выбрать подмножество $C' \subset C$ центров r_{elim} с минимальными значениями соответствующих переменных F_i; $r_{elim} = 1 + \lfloor 0.25\ (K - k) \rfloor$.

Шаг 2c.Удалить выбранные центры C' и снова применить алгоритм Ллойда

$$C \leftarrow Lloyd(C \backslash C').$$

Алгоритм 3. Greedy (Упрощенная жадная агломеративная процедура для построения индекса векторной базы данных)

Шаг 1. Инициализируем два набора центров: C_1, C_2. Можно использовать два способа инициализации: случайным образом выбрать k векторов данных в

качестве начальных центров или запустить алгоритм *k*-means++.

Шаг 2. Применяем алгоритм Ллойда для улучшения каждого набора

$$C_1 \leftarrow Lloyd(C_1); S_2 \leftarrow Lloyd(C_2).$$

Шаг 3. Вычислить $BasicAggl(C_1 \cup C_2)$.

Для оценки эффективности поиска данных с использованием индексов, построенных с использованием различных моделей и алгоритмов кластеризации, мы использовали два классических набора данных для задачи приближенного поиска ближайших соседей (ANN): SIFT10K (10 000 векторов данных размерностью 128) и SIFT1M (1 000 000 векторов данных размерностью 128). Индексы были построены с использованием различного количества центров, причем для кластеризации использовались как *k*-средние, так и *p*-медианные формулировки. Для генерации начальных решений в классическом алгоритме *k*-средних и предлагаемом агломеративном алгоритме использовались два метода инициализации: случайный выбор векторов данных в качестве начальных центров и алгоритм *k*-means++.

В таблице 4 представлены сравнительные результаты для различных алгоритмов. Каждый алгоритм был запущен 30 раз. В таблицах показаны достигнутые значения целевой функции (сумма расстояний до ближайших центров в случае задачи *p*-медианы или сумма квадратов расстояний до ближайших центров в случае задачи *k*-средних), а также достигнутые значения полноты. Тестовые наборы запросов для наборов данных были определены для расчета метрики полноты, и мы использовали конкретные тестовые наборы запросов, предоставленные разработчиками наборов данных.

Таблица 4 – Сравнительные результаты алгоритмов кластеризации для задачи *k*-средних и использования индекса на их основе для приближенного поиска ближайших соседей

Алгоритм кластеризации	Метод инициализации	Достигнутое значение целевой функции, 30 запусков			Достигнутая полнота 100@100, 30 запусков алгоритма		
		максимальное (худшее)	среднее	минимальное (лучшее)	максимальное (лучшее)	среднее	максимальное (худшее)
Набор данных SIFT1M, 256 центров							
Lloyd	Random	23508792	23516733.6	23525960	95.0547	94.95983	94.8563
Lloyd	*k*-means++	23500268	23513930	23521506	95.0643	94.96042	94.8627
Greedy	Random	23475608	23481650	23520708	95.3264	95.22806	94.8898
Greedy	*k*-means++	23475712	23478027.71	23481328	95.2568	95.21683	95.1711

Как видно из таблицы, даже при очень незначительных различиях в достигнутых значениях целевой функции кластеризации значения полноты могут существенно различаться. Тем не менее, определенная корреляция между значением целевой функции и полнотой все же наблюдается (рисунок 4). Алгоритм инициализации *k*-means++ не всегда улучшает производительность агломеративного алгоритма и не является достаточно требовательным к вычислительным ресурсам. С ростом объема данных и числа выделяемых

кластеров вычислительные затраты алгоритма инициализации *k*-means++ начинают многократно превышать затраты основного итеративного алгоритма Ллойда.

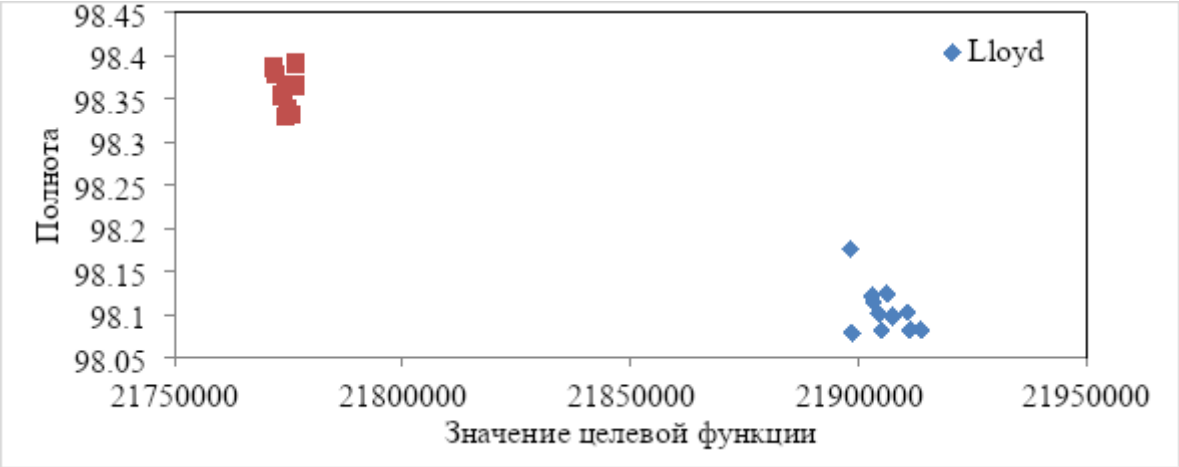

Рисунок 4 – Зависимость полноты поиска ближайших соседей на основе индекса, построенного с помощью алгоритма Ллойда и жадного агломеративного алгоритма для задачи *k*-средних, от достигнутого значения целевой функции *k*-средних для набора данных SIFT1M с 1024 центрами

Предложенный алгоритм, обеспечивающий компромисс, демонстрирует высокую эффективность кластеризации данных, но его текущая реализация требует значительных вычислительных ресурсов по сравнению с алгоритмом Ллойда. Дальнейшие исследования будут направлены на повышения эффективности самой жадной агломеративной процедуры, т.е. уменьшения разницы во времени вычислений при работе с большими наборами данных и увеличении количества центров/центроидов в результирующем индексе.

Четвертый раздел посвящен разработке нового алгоритма автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма *k*-средних совместно с локально-чувствительным хэшированием (LSH) для использования в системах хранения данных.

Предложен алгоритм автоматической группировки повторяющихся фрагментов блоков данных на основе упрощенной жадной агломеративной процедуры Greedy (с применением алгоритма *k*-средних), совместно с локально чувствительным хэшированием (LSH) для использования в системах хранения данных. Показано, что новый алгоритм повышает производительность сжатия.

Растущее количество данных на серверах создает проблему их эффективного хранения, частично решаемую сжатием данных. Ключевым фактором эффективности сжатия является способность обнаруживать повторяющиеся или похожие фрагменты в данных. Однако данные на сервере часто представлены разнообразными файлами, которые зачастую не объединены по сходству. Большинство существующих методов сжатия ограничены словарем или длиной сжатия. Обеспечение эффективного сжатия для такого рода данных чрезвычайно сложно.

Предлагается решение с поиском блоков данных по сходству с локально-чувствительным хэшированием (LSH), основанным на создании функций хеширования, которое с большей вероятностью присвоит одинаковый хеш

повторяющимся фрагментам блоков данных. Из предварительно обработанных блоков данных на Шаге 1 алгоритма LSH *K*-MEANS создадим массив признаков (хешей), с помощью алгоритма LSH SimHash для генерации 64-битного отпечатка (признака вектора данных) каждого блока данных, из которых сформируем фрагменты данных (вектор признаков данных). Затем с помощью алгоритма кластеризации *k*-средних с использованием расстояния Хэмминга (также известного как манхэттенское расстояние) сгруппируем повторяющиеся фрагменты данных. Алгоритм LSH *K*-MEANS отличается от алгоритма *k*-средних дополнительной операцией Шаг 1. Назовем фрагментом данных несколько отпечатков блоков данных.

Алгоритм 4.2 LSH *K*-MEANS

Шаг 1. Подготовить данные.

Шаг 1.1. Разделить данные на блоки из *n* строк, каждая из которых разбита на *m* блоков фиксированного размера.

Шаг 1.2. Для каждого блока вычислить 64-битный отпечаток с использованием алгоритма SimHash.

Шаг 1.3. Объединить хэш-значения блоков каждой строки в векторы длины *m*, для получения *n* векторов, каждый из которых состоит из *m* 64-битных значений.

Шаг 2. Инициализировать центры кластеров.

Шаг 2.1. Задать количество кластеров *k*.

Шаг 2.2. Случайным образом выбрать *k* векторов из подготовленных данных в качестве начальных центров кластеров.

Шаг 3. Кластеризовать данные.

Шаг 3.1. Для каждого вектора вычислить расстояние Хэмминга до каждого из k центров кластеров. Каждому вектору кластера назначить ближайший центр.

Шаг 3.2. Пересчитать центры кластеров алгоритмом Greedy, минимизируя сумму расстояний Хэмминга до векторов, принадлежащих кластеру.

Шаг 4. Сгруппировать повторяющиеся фрагменты блоков одного кластера.

Для сравнения эффективности LSH SimHash для группировки схожих рядов блоков используем функцию хеширования md5, не зависящую от локальности.

В экспериментах использовалась известная программа сжатия данных gzip. Алгоритм 4 (LSH *K*-MEANS) был реализован в одном потоке на центральном процессоре. Тестовая система состояла из процессора Intel Core i3-3220 с 8 ГБ оперативной памяти.

Для тестирования производилось архивирование и сжатие начальных рядов блоков жесткого диска хранилища данных. Наивное сжатие с помощью XZ_OPT=9 tar Jcvf sdd.iso.tar.xz sdd.iso. Команда tar формирует архив с именем sdd.iso.tar.xz, включающий файл sdd.iso. Опция «J» указывает, что архив создан в формате xz, а опция «c» указывает на создание архива.

Предварительная сортировка файлов выполнялась с помощью хеш-функции md5. Архив данных sdd.iso разбит на блоки данных по 1Мб - командой “split

../sdd.iso b 1M d", каждый файл пронумерован по порядку, содержит фактические данные архива.

Во-первых, добавим все файлы в архив tar с помощью команды "tar cvf - sdd.tar *", где каждый блок данных будет добавлен в архив последовательно. Повторяющиеся блоки данных алгоритм gzip ищет на расстоянии менее 32 КБ друг от друга. При уличении параметра тщательности сжатия до 9 xz ищет повторяющиеся блоки данных на расстоянии менее 64 МБ друг от друга.

Во-вторых, сортируем все похожие блоки данных по хэшу md5 и архивируем в одном файле. Для сравнения эффективности сжатия были применены следующие способы:

1. В первом способе проводилось архивирование без сжатия на блоки данных, каждый объемом 1Мб, и сжатие компрессором gzip.
2. Во втором методе проводилось архивирование без сжатия на блоки, каждый объемом 1Мб, с сортировкой блоков данных с помощью хеш-функции md5 для группировки повторяющихся фрагментов блоков данных и последующим сжатием компрессором gzip.
3. В третьем методе проводилось архивирование без сжатия на блоки, каждый объемом 1Мб, с группировкой Алгоритмом 4 LSH *K*-MEANS и последующим сжатием компрессором gzip.

Для сравнения различных способов сжатия данных компрессором gzip используем файлы виртуальных операционных систем linux, файл VDI_1904 размером 10739515392 байт, файл VDI_2004 размером 7702839296 байт, файл VDI_2204 размером 5309988864 байт.

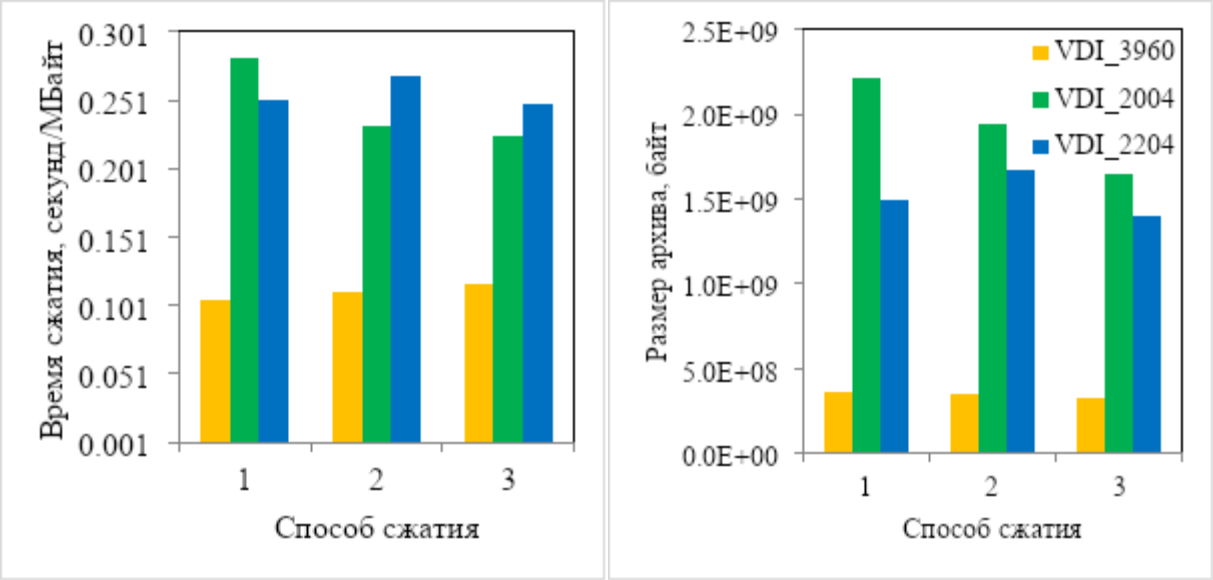

Рисунок 3 – Сравнительные результаты различных способов сжатия данных компрессором gzip: 1 – gzip, 2 – md5 и gzip, 3– LSH *K*-MEANS и gzip

Применение *k*-средних совместно с LSH для сортировки блоков данных для лучшей сжимаемости файлов архивов виртуальных операционных систем linux на рисунке 3 является сравнительно более эффективным относительно результатов стандартного компрессора gzip.

Преимущество применения нового алгоритма сжатия основанного на

архивации с помощью упрощенной жадной агломеративной процедура Greedy, основанной на алгоритме *k*-средних, совместно с LSH состоит в адаптации к базовому распределению данных, что повышает производительность сжатия.

Пятый раздел посвящен разработке новой процедуры инициализации центров кластеров для алгоритмов кластеризации, представляющей собой модификацию алгоритма инициализации *k*-means++. Полученная процедура инициализации применяется к большим данным. Учитывая меньшие вычислительные затраты, связанные с алгоритмом *k*-средних по сравнению с агломеративным алгоритмом, их применение в сочетании с модифицированным алгоритмом инициализации *k*-means++ к большим данным позволит уменьшить вычислительную сложность агломеративного алгоритма и улучшить его производительность.

С ростом объема данных и числа выделяемых кластеров вычислительные затраты алгоритма инициализации *k*-means++ (предложенного Артуром и Васильвицким) начинают многократно превышать затраты основных итеративных алгоритмов, таких как алгоритм Ллойда или PAM. Для больших наборов векторов данных (порядка миллиарда и более) нужны алгоритмы, в которых вычислительная сложность уменьшается относительно числа векторов данных.

Рассмотрим постановку задачи *k*-means++ и решение нового алгоритма инициализации центров кластеров. Пусть имеется набор векторов данных $a_i,\ i \in \{1, \dots, N\}$ в *M*-мерном пространстве измерений. Целью алгоритма *k*-means++ является нахождение центров $x_j,\ j \in \{1, \dots, k\}$ кластеров. Первый центр выбирается случайно из набора векторов данных. Алгоритм *k*-means++ для нахождения центра x_{j+1} вычисляет сумму квадратов расстояний от центра x_j до вектора данных a_i по формуле

$$\Sigma_{i=1}^{N}\left(d(x_j, a_i)\right)^2 = \Sigma_{i=1}^{N}(x_j - a_i)^2, \tag{6}$$

Алгоритм *k*-means++ для наборов данных из N векторов данных в *M*-мерном пространстве при однократном проходе по входным данным выполняет $O(k)$ итераций; каждая итерация характеризуется сложностью $O(NMk)$, где k – число центров. Таким образом, общее время выполнения равно $O(NMk^2)$. Поскольку для инициализации *k*-means++ требуется k проходов через данные, он плохо масштабируется для больших наборов данных.

Предлагаем в новой процедуре инициализации центров кластеров определить сумму квадратов расстояний от k центров x_j до N векторов данных a_i по формуле

$$\Sigma_{j=1}^{k}\Sigma_{i=1}^{N}\left(d(x_j, a_i)\right)^2 = NX^2 - 2\Sigma_{M`=1}^{M}(A_M X_M) + kA^2, \tag{7}$$

где X^2 – сумма сумм квадратов координат вектора в *M*-мерном пространстве для k центров x_j; A^2 – сумма сумм квадратов векторных координат в *M*-мерном пространстве для N векторов данных a_i. За счет сокращения количества операций суммирования по i и j для вычисления суммы квадратов расстояний снижается сложность вычислений. Процедура Mini-Batch

K-means++ последовательно вычисляет промежуточные значения суммы квадратов расстояний от центров x_j до векторов данных a_i в пакетах B.

Процедура Mini-Batch *K*-means++ выполняет k итераций, $j \in \{1, \dots, k`, \dots, k\}$.

На первом этапе находим сумму расстояний от N векторов данных до первого центра x_1 для выбора следующего центра x_j .

Поиском в упорядоченном массиве предварительно вычисленных слагаемых A_{MB}, A_B^2 мы определяем пакет $B`$, в которой находится искомый центр $x_j$. Далее мы итеративно уточняем желаемый центр $x_j$, используя векторы данных $a_i$ выбранного пакета $B`$.

На второй и последующих итерациях суммируются суммы расстояний от n векторов данных a_i до найденных центров x_j.

Алгоритм 5 Mini-Batch *K*-means ++:

Шаг 1. Выбрать равномерно случайным образом один центр $\boldsymbol{x_1}$ среди векторов данных $\boldsymbol{a_i}$. Пусть $\boldsymbol{B}$ количество пакетов, на которые разделен набор данных размера $\boldsymbol{N}$, $\boldsymbol{b}$ количество векторов данных в одном пакете, $\boldsymbol{N \leq Bb}$ и $\boldsymbol{N_{B`}}$ количество векторов данных из первого из всего набора данных до последнего в пакете $\boldsymbol{B`}$ включительно, где $\boldsymbol{i`} \in \{\boldsymbol{1 \dots N_{B`}}\}$. Пусть слагаемые $A_{MB`}$, $A_{B`}^2$ предварительно рассчитаны для всех пакетов $\boldsymbol{B}$.

Шаг 2. Вычислить слагаемые X_M и X^2 для центров x_j. Мы вычисляем сумму квадратов расстояний от k центров до векторов данных a_i, используя A_{Mi} чтобы найти второй центр, и X_{Mj} следующий (случайно выбранный) *j*-й центр x_j.

Шаг 3. Выбрать новый центр с вероятностью, пропорциональной сумме квадратов расстояний от выбранных центров x_j до векторов данных a_i. Для этого выбираем случайное расстояние

$$R \in \left\{0 \dots \sum_{j=1}^{k`} \sum_{i=1}^{N} \left(d(x_j, a_i)\right)^2\right\},$$

где $k`$ – количество найденных центров. В алгоритме 6 находим пакет $B`$, который содержит самый дальний вектор данных на расстоянии не больше R. Шаги 2 и 3 повторяются пока не выбраны все *k* центров.

После выполнения Алгоритма 5 мы получаем центры, которые используются для получения исходного результата кластеризации и сравниваем с результатом начальной кластеризации, полученным с использованием классического *k*-means++. Качество решения не уступает классическому *k*-means++, поскольку последовательность шагов в новой процедуре Mini-Batch *K*-means++ аналогична классическому *k*-means++ за исключением того, что новая процедура увеличивает скорость вычисления следующего центра (из всех векторов данных) так, чтобы вероятность выбора вектора была пропорциональна вычисленному квадрату расстояния до всех выбранных центров.

Алгоритм 6 Поиск в упорядоченном массиве предварительно вычисленных слагаемых A_{MB}, A_B^2 для определения ближайшего вектора данных на расстоянии не более $\boldsymbol{R}$.

Шаг 1. Установить максимальный и минимальный номер пакета $\boldsymbol{B`}$, $\boldsymbol{i_{min} = 0}$ и $\boldsymbol{i_{max} = B}$.

Шаг 2. Повторять до тех пор, пока $\boldsymbol{|i_{min} - i_{max}| > 1}$:

Шаг 3. Определить текущий номер пакета как

$$t = \frac{(i_{min} + i_{max})}{2}.$$

Шаг 4. Вычислить сумму квадратов расстояний до $\boldsymbol{k`}$ найденных центров от $\boldsymbol{n_t}$ векторов данных:

$$\boldsymbol{\sum_{j=1}^{k`} \sum_{i=1}^{n_t} \left(d(x_j, a_i)\right)^2 = n_t X^2 - 2 \sum_{M`=1}^{M} (A_{M`t} X_{M`}) + k` A^2.}$$

Шаг 5. Если $R > \sum_{j=1}^{k} \sum_{i=1}^{n_t} \left(d(x_j, a_i)\right)^2$.

Затем установить минимальный номер пакета $\boldsymbol{i_{min} = t}$, а сумма квадратов расстояний от $\boldsymbol{k`}$ центров до их векторов данных больше, чем

$$S_{min} = \sum_{j=1}^{k} \sum_{i=1}^{n_t} \left(d(x_j, a_i)\right)^2 < R.$$

В противном случае установить максимальный номер пакета $i_{max} = t$.

Шаг 6. Конец цикла (Шаг 2).

Шаг 7. Повторять до тех пор, пока $\boldsymbol{S_{min} < R}$, где $\boldsymbol{S}$ — сумма квадратов расстояний до вектора данных $\boldsymbol{b` \in \{1 \ldots b\}}$ из пакета $\boldsymbol{i_{min}}$:

Шаг 8. Рассчитать $\boldsymbol{S_{min} = S_{min} + \sum_{j=1}^{k} \left(d(x_j, a_{(t+b`)})\right)^2}$.

Шаг 9. Конец цикла (шаг 7). Выбираем следующий центр набора данных $\boldsymbol{x_j}$ из векторов данных a_i с индексом $i = t + b`$.

Шаг 10. Вычислить слагаемые $X_{M`j}$, X_j^2, X^2 для набора данных $\boldsymbol{x_j}$.

Таблица 3 – Статистическая значимость результатов вычислительных экспериментов

Алгоритм	Значение целевой функции / Время, секунд.					*p*-значения и статистическая значимость разницы при p=0.05
	Мини-мальное	Макси-мальное	Медиан-ное	Среднее	Стандар-тное отклоне-ние	
Набор данных BIRCH3, количество векторов n=10^5, центров k=100, запусков 30						
k-means++	1.072E+14	2.381E+14	1.490E+14	1.464E+14	3.654E+13	0.41794
	1.1088	1.2259	1.1558	1.1561	0.0273	незначима
Mini-Batch	1.125E+14	2.083E+14	1.648E+14	1.581E+14	2.982E+13	
K-means++	0.0005	0.0034	0.0005	0.0005	0.0005	
Набор данных ANN_SIFT1M, количество векторов n=10^6, центров k=2, запусков 30						
k-means++	1.616E+11	2.886E+11	2.235E+11	2.193E+11	0.355E+11	p=0.82588.
	1.2479	1.5288	1.4267	1.4077	0.0875	незначима
Mini-Batch	1.705E+11	2.850E+11	2.291E+11	2.213E+11	0.372E+11	
K-means++	0.00002	0.000026	0.000023	0.000023	0.000002	
Набор данных SUSY, количество векторов n=5000000, центров k=25, запусков 30						
k-means++	1345876	1537414	1417747	1419697	43747	p=0.19706.
	51.5	59.8	56.4	56.4	2.3	незначима
Mini-Batch	1301506	1503959	1399338	1402135	54719	
K-means++	0.182	0.199	0.185	0.185	0.0032	

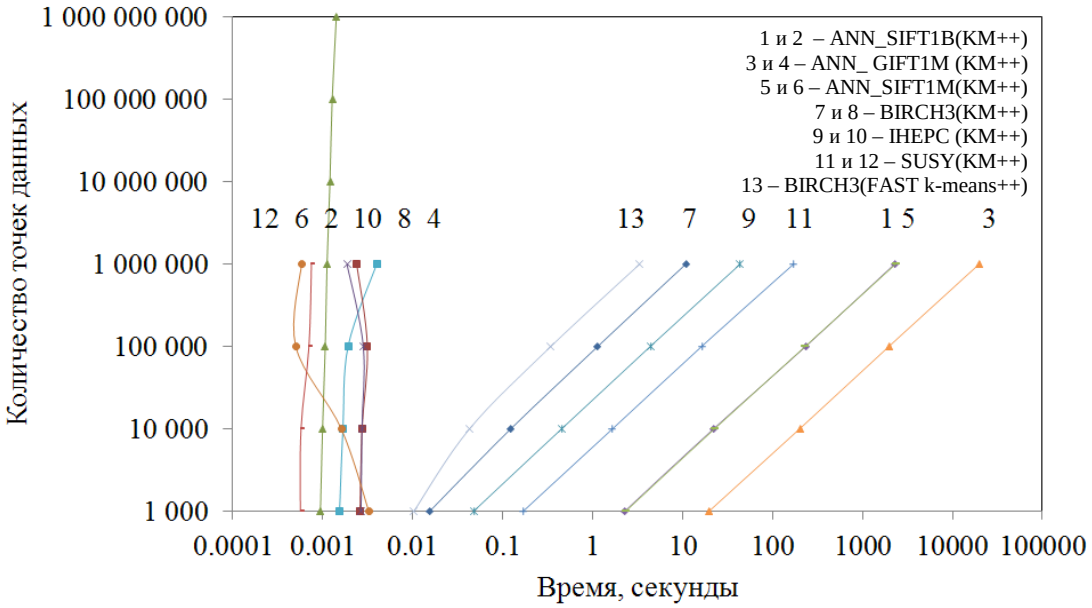

Рисунок 2 – Сравнение вычислительной сложности алгоритма FAST *k*-means++, *k*-means++(KM++) и Mini-Batch *K*-means++(MBKM++),

Основное преимущество в скорости инициализации центров, после вычисления вспомогательной структуры данных представленой упорядоченным массивом предварительно вычисленных слагаемых (с вычислительной сложностью $O(MkN)$), связано с уменьшением сложности вычислений (до $O(Mklog_2N)$, где N – количество векторов данных, M – количество измерений координат векторов данных, k – количество кластеров) с помощью формулы пакетного вычисления суммы квадратов расстояний путем поиска в упорядоченном массиве предварительно вычисленных слагаемых.

ЗАКЛЮЧЕНИЕ

В диссертации предложены новые решения, повышающие эффективность (точность и стабильность результатов, т.е. улучшение достигаемых значений целевой функции за заданное время) алгоритмов автоматической группировки объектов при большом объеме входных данных.

Цель диссертации достигается путем решения поставленных задач, а именно:

Разработанный новый подход к нормализации данных для предобработки входных данных, который может использоваться в системах анализа данных результатов неразрушающих испытаний образцов промышленной продукции. Он комбинирует нормализацию по допустимым значениям параметров оцениваемых значений продукции и оценку Джеймса-Штейна. Кроме того, он позволяет выявлять наиболее значимые контролируемые параметры продукции, что в перспективе позволит сократить затраты на испытания за счет уменьшения числа контролируемых параметров.

Разработанный новый алгоритм кластеризации для системы анализа данных электрорадиоизделий на основе данных тестовых испытаний характеризуется вполне приемлемой точностью и скоростью работы программной реализации. Он расширяет набор (ансамбль) алгоритмов для применения в системах анализа данных электрорадиоизделий (разделения смешанной партии ЭРИ на

однородные партии продукции) с применением задачи k-средних для нормализованных данных.

Разработанный новый алгоритм кластеризации для построения индекса векторной базы данных для построения приближенного индекса поиска ближайших соседей, реализован в рамках хоз договора в составе векторной. Дальнейшие исследования будут направлены на уменьшения разницы во времени вычислений при работе с большими наборами данных и увеличении количества центров/центроидов в результирующем индексе.

Разработанный новый алгоритм автоматической группировки повторяющихся фрагментов блоков данных на основе алгоритма k-средних, совместно с LSH для использования в системах хранения данных, реализован в рамках работ по договору СФУ для разработки модели оптимального использования дискового пространства с учетом компрессии данных.

Разработанная новая процедура инициализации центров кластеров для быстрых алгоритмов кластеризации больших и сверхбольших объемов данных, при поддержке Министерства науки и высшего образования Российской Федерации, по результатам проведенных нами экспериментов, востребована в системах дискового хранения данных и в векторных базах данных Российской технической компанией.

Результаты, изложенные в диссертации, имеют непосредственное практическое применение, в настоящее время новые разработанные алгоритмы был успешно применены в составе системы управления дисковыми хранилищами в ООО «Центр вычислительных технологий», а также в составе векторной СУБД Российской технологической компанией.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК России:

1. Ахматшин Ф. Г. О нормализации данных в задаче автоматической группировки промышленной продукции по однородным производственным партиям / Ф. Г. Ахматшин, И. Р. Насыров, В. Л. Казаковцев, Л. А. Казаковцев // Системы управления и информационные технологии. – 2020. – № 2(80). – С. 86-89. (ВАК К2).

2. Ахматшин Ф. Г. Подбор свободного параметра алгоритма FOREL-2 в задаче автоматической группировки промышленной продукции по однородным производственным партиям / Ф. Г. Ахматшин // Системы управления и информационные технологии. – 2021. – № 4(86). – С. 28-31. – DOI 10.36622/VSTU.2021.86.4.006. (ВАК К2).

3. Ахматшин Ф. Г. Алгоритм FOREL-2 с жадной эвристикой выбора радиуса поиска локальных сгущений / Ф. Г. Ахматшин, Л. А. Казаковцев // Системы управления и информационные технологии. – 2022. – № 3(89). – С. 39-42. – DOI 10.36622/VSTU.2022.89.3.009. (ВАК К2).

4. Ахматшин Ф. Г. О методе инициализации для алгоритмов кластеризации / Ф. Г. Ахматшин // Системы управления и информационные технологии. – 2024. – № 1(95). – С. 4-10. (ВАК К2).

5. Ахматшин Ф. Г. О сжатии блоков данных с использованием алгоритма кластеризации *k*-средних / Ф. Г. Ахматшин // Системы управления и информационные технологии. – 2024. – № 3(97). – С. 68-72. (ВАК К2)

Публикации в изданиях, входящих в системы цитирования WoS и Scopus:

6. Akhmatshin F. G., Reducing the James–Stein Shrinkage Estimator for Automatically Grouping Heterogeneous Production Batches / F. G. Akhmatshin, I. A. Petrova, L. A. Kazakovtsev, I. N. Kravchenko // Journal of Machinery Manufacture and Reliability. – 2024. – Vol. 53, No. 3. – P. 254-262. – DOI 10.1134/S1052618824700043.

7. Ahmatshin F. G. Impact of data normalization methods and clustering model in the problem of automatic grouping of industrial products / F. G. Ahmatshin, L. A. Kazakotsev // Journal of Physics: Conference Series, Krasnoyarsk, Russian Federation, 25 сентября – 04 2020 года. Vol. 1679. – Krasnoyarsk, Russian Federation: Institute of Physics and IOP Publishing Limited, 2020. – P. 32085. – DOI 10.1088/1742-6596/1679/3/032085.

8. Ahmatshin F. G. Greedy Heuristics for the Choice of the Radius of Local Concentrations in Forel-2 Algorithm / F. G. Ahmatshin, L. A. Kazakovtsev // Hybrid methods of modeling and optimization in complex systems : Proceedings of the International Workshop “Hybrid methods of modeling and optimization in complex systems” (HMMOCS 2022), Krasnoyarsk, 22–24 ноября 2022 года. – London, United Kingdom: European Proceedings, 2023. – P. 366-371. – DOI 10.15405/epct.23021.45.

9. Ahmatshin F. G., Kazakovtsev L. A. Mini-batch *K*-means++ clustering initialization. //XXIII International Conference Mathematical Optimization Theory and Operations Research MOTOR-2024 Omsk Russia, June 30 - July 06, 2024. – pp. 293-307.

10. Akhmatshin F. Clustering of *k*-means based on Euclidean distance metric and Mahalanobis metric / F. Akhmatshin, P. Egarmin, M. Gerasimova [et al.] // E3S Web of Conferences. – 2024. – Vol. 531. – P. 03002. – DOI 10.1051/e3sconf/202453103002.

В других изданиях:

11. Ахматшин Ф. Г. Подбор информативных параметров в задаче автоматической группировки промышленной продукции по однородным производственным партиям / Ф. Г. Ахматшин, Л. А. Казаковцев // Решетневские чтения : Материалы XXIV Международной научно-практической конференции, посвященной памяти генерального конструктора ракетно-космических систем академика М. Ф. Решетнева. В 2 ч., Красноярск, 10–13 ноября 2020 года / под общ. ред. Ю. Ю. Логинова. Том 1. – Красноярск: Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева", 2020. – С. 246-247.

12. Ahmatshin F. G. Selection of free parameters in the problem of automatic grouping of industrial products by homogeneous production batches / F. G.

Ahmatshin // Молодежь. Общество. Современная наука, техника и инновации. – 2021. – No. 20. – P. 315-317.

13. Ахматшин Ф. Г., Ершов, А. А. Применение алгоритма кластеризации *k*-means в задаче автоматической группировки данных / А. А. Ершов, Ф. Г. Ахматшин // Молодые ученые в решении актуальных проблем науки : сборник материалов Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, Красноярск, 22–23 апреля 2021 года. – Красноярск: Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева", 2021. – С. 568-569.

14. Ахматшин Ф. Г. О нормализации данных в задаче автоматической группировки промышленной продукции по однородным производственным партиям / М. И. Цепкова, Ф. Г. Ахматшин, И. А. Петрова, Л. А. Казаковцев // Молодые профессионалы : II Всероссийская конференция : сборник научных трудов, Санкт-Петербург, 10–12 октября 2023 года. – Санкт-Петербург: Национальный исследовательский университет ИТМО, 2024. – С. 24-26

Свидетельство о государственной регистрации программы для ЭВМ:

15. Свидетельство о государственной регистрации программы для ЭВМ № 2022615155 Российская Федерация. Система интеллектуального анализа данных для задачи разделения сборной партии промышленной продукции на однородные группы на основе алгоритма FOREL : № 2022613996 : заявл. 21.03.2022 : опубл. 30.03.2022 / Ф. Г. Ахматшин, Л. А. Казаковцев, И. П. Рожнов, М. И. Цепкова ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева».