ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧЕРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ (РТУ МИРЭА)»

На правах рукописи

ФИЛАТОВ АЛЕКСАНДР СЕРГЕЕВИЧ

ОБРАБОТКА И КЛАСТЕРИЗАЦИЯ СПЕКТРАЛЬНЫХ ДАННЫХ ЖИДКИХ СРЕД

Научная специальность 2.3.1. Системный анализ, управление и обработка информации, статистика

Диссертация на соискание учёной степени кандидата технических наук

Научный руководитель
Доктор технических наук, профессор
Николаева С.В.

Москва 2025

ОГЛАВЛЕНИЕ

Введение

Основное направление и актуальность исследования.

Актуальность исследования обусловлена возрастающей потребностью в автоматизации и стандартизации процессов контроля качества в химической промышленности, что является одним их ключевых условий развития экономики страны и повышения конкурентоспособности соответствующих отраслей промышленности на мировом рынке. В условиях модернизации производственных процессов и внедрения инновационных технологий необходимо обеспечивать стабильное соблюдение стандартов на всех этапах производства. Современные высокоразрешающие методы сбора информации для анализа генерируют огромные объемы данных, обработка которых требует быстродействующих и точных алгоритмов для своевременной интерпретации результатов. В этой связи разработка программных средств, способных интегрировать различные этапы анализа – от сбора данных до их визуализации и кластеризации – является необходимым условием для повышения эффективности контроля и оптимизации технологических процессов.

Проведенное исследование направлено на разработку и оптимизацию методов обработки спектральных данных, что является ключевым направлением в современной аналитической химии и машинном обучении. Актуальность данной темы определяется необходимостью повышения точности диагностики химического состава химических веществ на основе спектральных измерений, а также стремлением к автоматизации процессов анализа в условиях растущего объема экспериментальных данных. С одной стороны, современные приборы спектроскопии генерируют огромные массивы данных, обладающих высокой размерностью и сложной структурой, что существенно затрудняет их интерпретацию с

использованием традиционных методов анализа. С другой стороны, точное выявление и классификация спектральных особенностей различных веществ имеют важное практическое значение в химической промышленности, где своевременное и корректное определение состава материала позволяет принимать обоснованные решения, минимизировать риски и оптимизировать технологические процессы.

Одной из фундаментальных проблем является так называемое «проклятие размерности», когда с ростом числа измеряемых признаков ухудшается различимость объектов, а расстояния между точками становятся менее информативными. Это приводит к тому, что алгоритмы, основанные на стандартных метрических расстояниях, не справляются с выделением устойчивых кластеров, что в свою очередь снижает точность классификации и ухудшает интерпретацию полученных результатов. Кроме того, высокая вычислительная сложность применяемых методов ограничивает возможности использования таких алгоритмов в реальном времени и при обработке больших объемов информации.

Еще одной проблемой является необходимость предварительной обработки спектральных данных, позволяющая выделять ключевые характеристики сигнала. Отсутствие эффективной стратегии предобработки может привести к потере информации или, наоборот, к усилению шума, что негативно сказывается на качестве анализа. Таким образом, выбор оптимальных методов обработки и адаптивных алгоритмов является ключевым фактором, влияющим на конечные результаты исследования.

В настоящее время в области контроля качества продукции большое внимание уделяется методам спектроскопии. Для решения проблем анализа спектральных данных широко используют высокоразрешающие методы числовой обработки, в том числе нейросетевые методы. Исследованию в этих областях посвящены работы многих учёных: Алаторцева Е.И., Апяри В.В., Битюкова В.К., Большакова О.В.,

Вагина В.А., Голяка И.С., Дмитриенко С.Г., Жижина Г.Н., Краснова А.Е. Морозова А.Н., Костогрызова А.И., Красникова С.А., Николаевой С.В., Хаустова И.А., Хвостова А.А., Christian W. Huck, Justyna Grabska, Krzysztof B. Beć, Thomas Bocklitz и др.

Таким образом, создание интегрированной системы, объединяющей методы обработки спектральных данных и автоматизированного контроля качества химической продукции является актуальной задачей, решение которой позволяет повысить качество и конкурентоспособность отечественной продукции, укрепить позиции российских компаний на мировом рынке и обеспечить безопасность использования химической продукции. Применение современных алгоритмов машинного обучения позволит ускорить процесс анализа и повысить его надежность, что является важным шагом в развитии технологий оценки качества продукции. Разработка таких систем имеет большое практическое значение, поскольку позволяет создавать продукцию, отвечающую самым строгим международным стандартам, и способствует стабильному развитию экономики страны за счет повышения эффективности производственных процессов и оптимизации контроля качества на всех этапах производства.

Диссертация соответствует паспорту специальности 2.3.1, а именно пунктам 4, 5 и 17.

Целью диссертационной работы повышение достоверности оценки качества жидких сред путем разработки методики формирования и распознавания их цифровых образов.

Для достижения поставленной цели необходимо решить следующие задачи.

1. Исследовать методы распознавания образцов жидких сред на основе их спектральных характеристик, провести системный анализ полученных результатов и определить наиболее эффективные подходы.

2. Разработать модификации алгоритмов снижения размерности данных для формирования цифровых образов спектральных данных жидких сред и метод кластеризации данных по этим образам.
3. Провести проверку и обосновать разработанные решения для обеспечения достоверности и применимости полученных результатов.
4. Разработать программный комплекс для обработки, анализа и подготовки спектральных данных к кластеризации и последующему использованию в методах машинного обучения, включающий базу данных цифровых образов и интегрировать разработку в существующую экспертно-нейросетевую систему для практического применения.

Объектом исследования являются системы оценки качества жидких сред по их спектральным характеристикам.

Предметом исследования являются методы формирования и распознавания цифровых образов жидких сред на основе их спектральных характеристик.

Методы и средства исследования. В исследовании применялись современные методы математического анализа и машинного обучения для обработки и интерпретации спектральных данных химических продуктов. Ключевыми методами стали алгоритмы предварительной обработки, такие как дискретная свёртка, автокорреляция и дискретная производная, которые позволили устранить шум и выделить существенные признаки сигналов. Для снижения размерности высокоразмерных данных использовались как классические методы (PCA, MDS), так и современные нелинейные алгоритмы (t-SNE, UMAP, Isomap, LLE), обеспечивающие сохранение локальной и глобальной структуры данных. Для кластеризации спектральных данных были среди прочих задействованы алгоритмы DBSCAN, HDBSCAN, Spectral Clustering и K-means, что позволило группировать данные в

соответствии с их физико-химическими характеристиками. Оценка качества кластеризации проводилась с использованием таких метрик, как коэффициент силуэта и индекс Дэвиса-Боулдина, что обеспечило объективное сравнение результатов. Помимо этого, интеграция вычислительных методов с системами хранения данных на основе PostgreSQL и использование инструментов сериализации моделей через протоколы pickle способствовали автоматизации и масштабируемости анализа.

Научная новизна работы. В диссертационной работе впервые получены следующие научные результаты.

1. Разработана методика формирования цифровых образов жидких сред для их последующего анализа с применением методов машинного обучения, в том числе кластерного анализа, отличающаяся дополнительным предварительным этапом обработки данных путем нахождения первой производной исходных спектров, позволяющим эффективно выделять информативные признаки из спектральных сигналов, и адаптивном подбором метрик расстояния что обеспечивает оптимальное сравнение объектов в высокоразмерном пространстве.
2. Разработан метод кластеризации цифровых образов жидких сред, обеспечивающий высокую точность их разделения за счёт применения адаптивных алгоритмов, позволяющих автоматически настраивать параметры кластеризации в зависимости от специфики данных. Данный подход базируется на использовании критериев оценки внутрикластерной компактности и межкластерной разделимости, что позволяет достичь стабильного и интерпретируемого разбиения спектральных данных.
3. Проведен системный анализ результатов кластеризации с применением методов многокритериальной оценки точности кластеризации, выявивший закономерности в распределении спектральных характеристик жидких сред. Этот анализ позволил выявить взаимосвязь между качеством предварительной обработки, выбором

метрики расстояния и эффективностью кластеризации, что в итоге привело к адаптации разработанного подхода с целью повышения точности и надёжности аналитической модели.

Теоретическая значимость результатов работы состоит в формировании и обосновании представления о системном подходе к обработке спектральных данных с применением методов машинного обучения. Результаты работы уточняют теоретические основы построения аналитических моделей, интегрирующих предобработку, снижение размерности и кластеризацию в единую систему.

Практическая значимость диссертационной работы заключается в следующем.

Разработанное методическое обеспечение определяет простой способ формирования цифровых образов веществ на основе их спектральных характеристик, что позволяет повысить точность кластеризации и качество анализа данных в целом. Такой подход может применяться для контроля качества продукции, мониторинга технологических процессов, а также для быстрого анализа сырья и готовых продуктов без необходимости обращения в специализированные лаборатории. Это существенно ускоряет процесс принятия решений на всех этапах производства и транспортировки, обеспечивая оперативное реагирование на возможные отклонения от заданных параметров.

Сформированная база данных и внедренные методы позволяют экспертно-нейросетевым системам выполнять свои задачи, обеспечивая непрерывный сбор, хранение и обработку спектральных данных. Благодаря гибкой архитектуре базы данных происходит систематизированное хранение информации о веществе и результатах её последующей обработки. Это позволяет не только сохранять исторические данные для ретроспективного анализа, но и осуществлять динамический мониторинг, что особенно важно для систем, требующих регулярной переоценки состояния анализируемых объектов. Система поддерживает интеграцию с

различными алгоритмами машинного обучения и нейросетевыми моделями, что способствует автоматизации анализа и повышению его достоверности за счет использования адаптивных методов оптимизации параметров.

Кроме того, разработанный программный комплекс предоставляет широкие возможности для анализа данных и их последующего использования в экспертно-нейросетевых системах. Встроенные модули по снижению размерности, кластеризации и визуализации спектральных данных позволяют создавать интерпретируемые модели, способные выявлять скрытые закономерности в сложных многомерных данных. Такая система может применяться для разработки интеллектуальных диагностических средств, в которых на основе спектрального анализа осуществляется раннее обнаружение аномалий и прогнозирование отклонений в технологических процессах. Возможность интеграции с нейросетевыми архитектурами, использующими как обученные модели для распознавания цифровых образов, так и адаптивные алгоритмы для корректировки параметров анализа, обеспечивает автоматизацию и повышение точности экспертных оценок. В совокупности, разработанные методические подходы и реализация программной системы дают возможность реализовать комплексное решение для мониторинга и анализа процессов в промышленности, что является важным шагом на пути к созданию автономных экспертно-нейросетевых систем, способных работать в режиме реального времени.

Научные и практические результаты, полученные в диссертации, внедрены в:

– учебном процессе ФГБОУ ВО «НИУ «МЭИ»;

– учебном процессе ФГБОУ ВО «МГУТУ им. К.Г. Разумовского (ПКУ)»

– ФГБУН Научно-технологическом центре уникального приборостроения Российской академии наук (НТЦ УП РАН);

– АО Всероссийский научно-исследовательский институт по переработке нефти (АО «ВНИИ НП»);

– ООО «КВС Электро»;

– АО Можайское экспериментально-механическое предприятие (АО МЭМП).

Разработанное программное обеспечение и база данных ЭНС защищены свидетельствами о государственной регистрации программ для ЭВМ № 2024665128 от 27.06.2024, № 2025619420 от 16.04.2025 и № 2025622177 от 23.05.2025.

Основные положения, выносимые на защиту.

1. Предложенная методика формирования цифровых образов веществ на основе предварительно обработанных спектров различных химических веществ позволяет создавать компактные и структурированные представления данных, пригодные для последующего применения алгоритмов машинного обучения и кластеризации.
2. Разработанная база данных позволяет хранить и получать доступ к спектральной информации, результатам обработки и метаданным, описывающим образцы и спроектирована с учетом требований к масштабируемости, целостности данных и совместимости с аналитическими программами, а также обеспечивает воспроизводимость экспериментов и накопление знаний.
3. Специализированное программное обеспечение, включающее модули для идентификации спектров, визуализации, построения моделей и проведения обучения, позволяет проводить полный цикл анализа – от загрузки исходных спектров до интерпретации результатов кластерного анализа и принятия практических решений, связанных с идентификацией веществ, контролем качества и выбором дальнейших аналитических процедур.

Личный вклад автора.

Представленные результаты диссертации являются итогом исследования, проводимого лично автором в рамках поставленных цели и задач данной работы в период с 2022 по 2025 гг.

Содержание диссертации и основные положения, выносимые на защиту, являются персональным вкладом автора в опубликованные работы. Значительная часть опубликованных работ выполнена самостоятельно. Лично автором предложен комплексный подход, объединяющий этапы предварительной обработки, интеллектуальной обработки с применением современных алгоритмов машинного обучения. Автором разработана база данных цифровых образов жидких сред для их хранения и последующей обработки и написан программный код модуля для экспертно-нейронной системы для анализа спектральных данных. Также автором разработана обучающая программа спектрального анализа.

Степень достоверности результатов исследования.

Достоверность научных положений обеспечивается за счет комплексного подхода, включающего всесторонний анализ теоретических моделей, экспериментальное подтверждение гипотез с использованием современных методов машинного обучения и статистической обработки данных, а также проведение сравнительного анализа алгоритмов с применением объективных метрик качества. Обоснованность результатов подтверждается репликацией экспериментов, а также практической проверкой разработанных методик на реальных данных, что свидетельствует о высокой надежности и применимости предложенных решений в условиях промышленного контроля и анализа спектральных характеристик.

Апробация работы.

Основные результаты диссертационной работы докладывались на следующих научных форумах:

Современные информационные технологии в образовании, науке и промышленности. Москва, 10–11 ноября 2022 года.

Информационно-аналитические и интеллектуальные системы для производства и социальной сферы. Москва, 24 ноября 2022 года. Российский биотехнологический университет.

Современные информационные технологии в образовании, науке и промышленности. Москва, 09–10 ноября 2023 года. Региональное отделение "Информационные технологии и процессы" общественной организации "Международная академия информатизации", ФГАОУ ВО "Государственный университет просвещения", факультет изобразительного искусства и народных ремёсел, ФГБОУ ВО "МГУТУ им. К.Г. Разумовского (ПКУ)".

Современные информационные технологии в образовании, науке и промышленности. Москва, 25–26 апреля 2024 года. Общественная организация «Международная академия информатизации», Региональное отделение «Информационные технологии и процессы»; АО «Нейросети»; ФГАОУ ВО «Государственный университет просвещения», Факультет изобразительного искусства и народных ремёсел; ФГБОУ ВО «МГУТУ им. К.Г. Разумовского (ПКУ)»; ФГБОУ ВО «Финансовый университет при Правительстве РФ», Кафедра бизнес-информатики факультета информационных технологий и анализа больших данных.

Международная научно-практическая конференция "Информационные технологии, искусственный интеллект, большие данные: актуальные тенденции, перспективные исследования" (ITAIB 2024). 28-29 ноября 2024 г. Консорциум содействия развитию науки и технологий (AST Consortium).

Современные информационные технологии в образовании, науке и промышленности. Москва, 07–08 ноября 2024 года. Общественная организация «Международная академия информатизации ФГБОУ ВО «Финансовый университет при Правительстве РФ».

Пленарные доклады XIII национальной научно-практической конференции «Моделирование энергоинформационных процессов», г. Воронеж. 24 декабря 2024 г.

Всероссийская научно-практическая конференция, посвященная памяти советского математика, доктора физико-математических наук, профессора П.П. Коровкина «Математика в современном мире». Калуга, 23-24 мая 2025 года. Калужский государственный университет им. К.Э. Циолковского.

Результаты исследования представлялись и получили одобрение на 4 расширенных заседаниях кафедры математического обеспечения и стандартизации информационных технологий института информационных технологий ФГБОУ ВО «МИРЭА – Российский технологический университет».

Публикации.

По теме диссертации опубликовано 12 научных работ, в том числе 4 работы в рецензируемых научных периодических изданиях, рекомендованных ВАК РФ, и 8 тезисов в сборниках трудов конференций и получено 2 свидетельства о государственной регистрации программ для ЭВМ и 1 свидетельство о регистрации базы данных.

Структура и объем диссертации.

Диссертация состоит из введения, четырёх глав, заключения, списка сокращений и условных обозначений, списка литературы и десяти приложений. Работа изложена на 254 страницах основного текста; она содержит 27 таблиц, 51 рисунок; список литературы включает 167 наименований, из которых 67 отечественных и 100 зарубежных авторов.

Глава 1. Анализ методов обработки спектральных данных

1.1. Представление вещества в виде спектра

Вещество, формально представленное в виде спектра, может быть изучено как объект, обладающий уникальным набором характеристик, которые выражаются через взаимодействие с электромагнитным излучением. Спектр является совокупностью зависимостей интенсивности или других параметров сигнала от частоты, длины волны или энергии фотонов. Этот подход основывается на том, что любое вещество взаимодействует с излучением в соответствии со своей структурой, составом и состоянием, оставляя в спектре характерные «отпечатки». Такое представление позволяет применять к веществу строгие математические методы анализа, выявлять его физические и химические свойства и идентифицировать его с высокой точностью.

Каждое вещество обладает уникальным спектром, будь то спектр поглощения, излучения или рассеяния. Например, молекулы вещества поглощают свет на определенных длинах волн, соответствующих энергетическим переходам между уровнями в электронных, вибрационных или вращательных состояниях. Этот процесс отражается в спектре поглощения, который служит своеобразной «подписью» вещества. Спектры излучения также дают ценную информацию, поскольку возбужденные атомы или молекулы испускают свет с длинами волн, которые зависят от их электронной структуры [23].

Представление вещества в виде спектра открывает широкие возможности для его анализа. Одна из ключевых областей применения – это идентификация вещества [28; 38; 53; 61]. Например, в спектроскопии ИК-диапазона молекулы анализируются на основе их вибрационных характеристик, что позволяет выявить функциональные

группы и химическую структуру соединений [25]. В атомно-абсорбционной спектроскопии можно определять присутствие определенных элементов, так как каждый атом поглощает свет на фиксированных длинах волн, соответствующих переходам электронов [24]. С помощью таких методов можно не только идентифицировать вещества, но и количественно оценить их концентрацию.

Спектры предоставляют возможность изучать физико-химические свойства вещества в динамике. Например, регистрация спектров в реальном времени позволяет исследовать химические реакции, изменения состояния вещества (например, плавление и испарение) или взаимодействие молекул с окружающей средой. При этом изменения в спектрах сигнализируют о перестройке молекулярной структуры, что помогает отслеживать промежуточные продукты реакции или определять механизмы взаимодействия.

Спектральное представление особенно ценно для анализа сложных многокомпонентных систем, таких как биологические растворы или смеси. Используя методы спектрального разложения и машинного обучения, можно выделять индивидуальные спектры компонентов, что позволяет проводить качественный и количественный анализ даже при высокой степени перекрытия спектров [1]. Например, спектроскопия в ближнем ИК-диапазоне (NIR) активно используется для анализа состава пищевых продуктов или фармацевтических препаратов без необходимости разрушать образец [18].

Еще одной важной областью применения является экологический мониторинг. Благодаря спектральным методам можно определять содержание загрязнителей в воздухе, воде или почве с высокой чувствительностью. Такие технологии особенно актуальны для обнаружения токсичных газов, тяжелых металлов или органических загрязнителей, которые могут быть представлены в микроконцентрациях [49]. Современные методы спектроскопии, такие как лазерная спектрометрия или

комбинационная спектроскопия, обеспечивают точность и оперативность измерений [21; 22]. В нанотехнологиях спектральные методы помогают изучать оптические эффекты, такие как плазмонный резонанс, который характерен для наночастиц [65]. Эти данные используются при создании новых материалов с заданными свойствами.

Формальное представление вещества в виде спектра позволяет интегрировать данные в вычислительные системы для автоматизации анализа. Применяя методы обработки больших данных, можно создавать базы спектров, которые используются для автоматического распознавания веществ [42]. Это особенно актуально в таких областях, как биомедицина и контроль производства.

В аналитической химии и физике принято различать несколько основных классов спектральных данных. Наиболее распространенным видом спектров являются оптические спектры. К данному классу относятся спектры поглощения, пропускания, отражения и люминесценции. Они основаны на взаимодействии электромагнитного излучения в диапазоне ультрафиолет-видимого и ближнего инфракрасного света с электронами и молекулярными орбиталями. Ультрафиолетовые и видимые спектры характеризуют электронные переходы, что делает их информативными при исследовании окрашенных соединений, комплексов металлов и органических красителей. Инфракрасные спектры отражают колебательные и вращательные переходы, позволяя анализировать функциональные группы в органических и неорганических соединениях.

Спектры комбинационного рассеяния (рамановские) фиксируют неупругое рассеяние фотонов на молекулярных колебаниях и дают возможность получать информацию о симметрии и структуре молекул. Такие спектры могут являться взаимодополняющими по отношению к инфракрасным.

Масс-спектральные данные описывают распределение ионов по отношению массы к заряду. Хотя данный тип спектров не относится к оптическим, он формируется

по тем же принципам численного отображения сигнала и используется для молекулярной идентификации, определения элементного состава и исследования сложных смесей.

Спектры ядерного магнитного резонанса фиксируют отклик ядер с ненулевым вращением на воздействие внешнего магнитного поля и радиочастотного излучения. Они содержат сведения о химическом окружении атомов, конформации молекул и межъядерных взаимодействиях. Этот вид спектров широко применяется в органической химии и биохимии как один из наиболее информативных.

Для высокочувствительного количественного анализа элементов применяются эмиссионные и атомно-абсорбционные спектры. В эмиссионных спектрах фиксируется излучение возбужденных атомов или ионов, а в атомно-абсорбционных спектрах измеряется степень поглощения света атомами в газовой фазе.

Таким образом, спектральное представление вещества – это универсальный и мощный подход, который раскрывает фундаментальные свойства материалов, позволяет идентифицировать вещества, изучать их поведение и взаимодействие, а также разрабатывать новые аналитические технологии. Его применение охватывает широкий спектр задач, от фундаментальных исследований до прикладных областей, создавая основу для научных и технических достижений.

1.2. Методы системного анализа, применяемые в области кластеризации спектральных данных

В этом подразделе рассматривается набор методов системного анализа, которые могут быть эффективно использованы при исследовании и построении алгоритмов кластеризации многопараметрических спектральных данных. Под системным

анализом понимается комплексный подход к изучению объектов как взаимосвязанной системы компонентов, включающий моделирование структуры и поведения, оценку чувствительности и устойчивости, формализацию потока информации и оптимизацию параметров. Для спектральных данных такой подход позволяет не только повысить качество разбиения на кластеры, но и обеспечить интерпретируемость, устойчивость к шуму и воспроизводимость результатов.

Первое направление – формальное моделирование процесса получения спектров и выделение функциональных подсистем. Полезно рассматривать наблюдаемый спектр как результат прохождения физических сигналов через измерительную систему с собственным откликом и шумовыми источниками. В терминах системной идентификации измерение представимо как некий оператор, действующий на истинный спектр с добавлением шума. Анализ структуры этого оператора (например, свёртка с импульсной характеристикой, фильтрация, искажения детектора) позволяет формализовать предобработку (включая свёртку, фильтрацию, нормализацию) как корректировку модели системы и, таким образом, повысить однородность данных перед кластеризацией. Практически это приводит к алгоритмическим блокам: оценка и удаление систематического фона, выравнивание по масштабу и смещению, аппроксимация шума (например, моделирование шума как аддитивного нормально распределённого процесса с параметрами, оценёнными из реплик).

Второе направление – многоуровневый разбор системы признаков и многомасштабный анализ. Системный подход рекомендует декомпозировать спектр на несколько уровней представления: смещение базовой линии и глобальные тренды, локальные пиковые структуры и высокочастотные флуктуации. Для каждого уровня целесообразно выбирать собственный класс признаков и методов кластеризации: для глобальных профилей – методы уменьшения размерности, для локальных пиков – признаки формы и положения, для шумовых компонент – вейвлет-анализ и

автокорреляционные характеристики. Такой модульный подход позволяет строить ансамблевые схемы, где объединение результатов разных уровней производится по правилам системной интеграции.

Третье направление – применение теорий информации и устойчивости для выбора и валидации кластеров. Информационные меры, такие как энтропия распределения меток, взаимная информация между признаковыми подпространствами и стабильность разбиения, являются инструментами системной оценки качества кластеров. Конкретно, для оценки устойчивости можно измерять согласованность кластеров при случайных подвыборках или при добавлении шума. В системном анализе важен анализ чувствительности: исследование, как небольшие вариации входных данных и параметров алгоритма влияют на структуру разбиения – это формализуется как локальная и глобальная чувствительность модели.

Четвертое – графовые и сетевые представления как средство системной абстракции. Спектральные данные сводятся к графу сходства: узлы – спектры, веса ребер задаются функцией ядра, например, гауссовым ядром. В этом представлении методы спектральной кластеризации используют спектральную теорию графов: строится матрица смежности, степенная матрица и лапласиан, затем решение сводится к анализу собственных векторов лапласиана и последующему применению алгоритма разбиения, например, k-means в пространстве собственных векторов. В системном контексте такой подход позволяет учитывать не только попарные расстояния, но и глобальные структурные свойства множества спектров.

Пятое направление – динамическая и адаптивная кластеризация как задача управления. Если спектральные измерения собираются во времени или в условиях изменяющейся среды, полезно применять подходы из теории управления и адаптивной фильтрации: поточные алгоритмы кластеризации трактуются как регуляторы, поддерживающие актуальную модель данных при поступлении новых

измерений. Формализация в виде системы обратной связи позволяет задать критерии стабильности и быстродействия при изменениях.

Шестое направление – оптимизационные и операционные методы для подбора структурных гиперпараметров. Системный анализ рекомендует рассматривать выбор гиперпараметров как задачу многокритериальной оптимизации: минимизация внутрикластерной дисперсии и максимизация межкластерного расстояния при ограничениях на сложность модели и вычислительные ресурсы. Здесь применимы как классические методы как градиентный спуск, так и эвристические глобальные оптимизаторы, например, эволюционные стратегии или байесовская оптимизация с учётом ограничений системы измерения и требований воспроизводимости.

Наконец, важна методология верификации и документирования как часть системного построения эксперимента. Это включает построение сценариев валидации, моделирование искусственных смесей спектров для проверки разрешающей способности алгоритмов, а также формализацию метрик интерпретируемости. Такой подход обеспечивает научную обоснованность выводов и повышает доверие к результатам в диссертационной работе.

1.3. Методы кластерного анализа и их связь со структурой данных

Кластерный анализ как область системного исследования тесно связан с тем, каким образом исходные данные представлены и организованы. Структура данных играет не только техническую роль в вычислительной реализации, но и определяет, какие свойства объектов будут учтены при формировании кластеров и какие методы окажутся наиболее адекватными. В системном анализе принято рассматривать три уровня структурирования: объектно-признаковый, матричный и реляционно-сетевой.

Каждый из них задаёт собственный тип отношений между объектами, что, в свою очередь, обусловливает выбор алгоритмических решений.

Объектно-признаковая структура является наиболее распространённым представлением данных в виде множества объектов, каждый из которых описывается набором признаков. Такая структура подразумевает, что объект может быть интерпретирован как точка в многомерном пространстве, а задача кластеризации сводится к поиску областей с высокой плотностью или компактностью. В этом случае эффективно применяются методы, опирающиеся на геометрическую интерпретацию расстояний: алгоритмы k-means и его модификации [149], методы, основанные на моделях смесей [40], а также вероятностные и нечеткие подходы. Существенным преимуществом данной структуры является простота вычислений и возможность использования традиционных мер сходства. Однако именно здесь наиболее ярко проявляется «проклятие размерности», когда различия между объектами нивелируются при увеличении числа признаков, что требует предварительного снижения размерности или выбора более устойчивых мер сходства.

Другой подход к представлению данных заключается в том, что внимание переносится с отдельных объектов на попарные отношения между ними. В таком случае исходным материалом для анализа становится матрица расстояний или матрица сходства, которая отражает относительные различия и близости в системе. Иерархическая кластеризация, спектральные методы и алгоритмы [94], использующие графовую матрицу Лапласа, в значительной мере опираются именно на такую форму данных. Она позволяет выявлять глобальные зависимости, строить древовидные схемы и исследовать вложенные структуры. Важным следствием матричного представления является независимость от размерности исходного признакового пространства: анализ осуществляется на уровне отношений, что облегчает применение к разнородным данным. Вместе с тем матрица расстояний имеет

квадратичную размерность и становится вычислительно затратной при больших объёмах информации, что ограничивает масштабируемость методов.

Третий уровень представления данных – реляционно-сетевая структура – связан с ситуациями, когда система описывается не только свойствами отдельных объектов, но и сложной сетью связей между ними. Здесь исходными данными выступает граф или гиперграф, где вершины – это объекты, а ребра определяют наличие и силу отношений. Кластеризация в таком контексте приобретает смысл выявления множеств, обладающих высокой внутренней связанностью и относительно слабой связанностью с остальной частью системы. Методы на основе максимизации модульности, спектральные алгоритмы, а также плотностные подходы [100] демонстрируют свою эффективность именно в этой структуре. Преимущество этого описания заключается в том, что оно позволяет учитывать неявные топологические свойства системы и выявлять кластеры произвольной формы.

Сравнительный анализ показывает, что каждая форма представления данных задаёт собственную трактовку схожести объектов:

- в объектно-признаковой структуре схожесть измеряется через расстояние в пространстве признаков;
- в парной матричной структуре она определяется относительными отношениями между всеми объектами;
- в реляционно-сетевой структуре – топологией связей и локальными плотностями.

Таким образом, выбор метода кластерного анализа неразрывно связан с характером структурирования данных. Ошибочное приведение информации к неподходящей форме способно исказить результат анализа и привести к ложной интерпретации структуры системы. В системном анализе именно согласование формата данных и алгоритмического подхода выступает необходимым условием получения содержательно обоснованных результатов.

1.4. Методы спектрального анализа

Методы оптической молекулярной абсорбционной спектроскопии, такие как спектрофотометрия, спектроскопия диффузного отражения и цветометрия, занимают ключевое место среди современных инструментов аналитической химии. Их востребованность обусловлена сочетанием низкой стоимости и энергозатратности, что особенно важно в условиях необходимости проведения экспресс-анализа. Преимущества этих методов позволяют эффективно применять их как для рутинных исследований, так и для предварительного скрининга в ситуациях, где требуется оперативность и минимальное техническое обеспечение. Это делает их неотъемлемой частью аналитических процедур в различных областях науки и практики, что стимулирует развитие новых технологий и подходов.

Однако потенциал методов оптической спектроскопии все еще остается недостаточно реализованным. Для его раскрытия необходимо сосредоточить внимание на разработке новых спектрофотометрических реагентов и аналитических систем, а также оптимизировать методы разделения и концентрирования проб. Это особенно актуально в связи с расширением области применения этих методов и ростом требований к точности, чувствительности и селективности анализов.

Особую ценность представляет внедрение в практику анализа гетерогенных аналитических систем, в которых используются сорбенты, твердофазные реагенты, наночастицы, мицеллы и эмульсии. Среди них перспективными являются твердофазные хромогенные реагенты. Эти материалы позволяют объединить этапы аналитической реакции и концентрирования, исключая необходимость десорбции продуктов. Такой подход не только упрощает процедуру анализа, но и повышает его эффективность и точность [3].

Развитие методов оптической молекулярной абсорбционной спектроскопии связано с интеграцией новых материалов и подходов, таких как нанотехнологии, а также с совершенствованием методологии анализа. Это позволяет не только повысить точность и эффективность существующих методов, но и расширить их применение в науке и практике, включая биомедицину, экологию и промышленность.

Также известны исследования по разработке нового подхода к созданию твердофазных аналитических реагентов на основе химически модифицированного пенополиуретана (ППУ) [4]. Особое внимание уделено реакциям диазотирования концевых толуидиновых групп ППУ, что делает возможным его использование в качестве матрицы для аналитических целей [17]. Этот процесс позволяет вводить диазониевые группы в структуру ППУ, которые затем взаимодействуют с различными органическими соединениями. Такой подход значительно упрощает разделение продуктов реакции благодаря монолитной природе полимерной матрицы и низким концентрациям реагирующих веществ в системе.

Диазотированный ППУ демонстрирует способность вступать в реакции азосочетания с органическими соединениями различной природы, такими как ароматические амины, фенолы, аминофенолы и кетоны. Эти реакции происходят в щелочной среде (pH 8–12), за исключением случаев с гидроксибензойными кислотами, где оптимальный выход достигается в слабокислой среде (pH 3–5). Продукты реакций, как правило, обладают интенсивной окраской, что обусловлено изменением спектров диффузного отражения модифицированного ППУ. Например, взаимодействие с соединениями, имеющими разветвленные сопряженные π-системы, приводит к смещению полос поглощения в область больших длин волн, что демонстрирует улучшение спектральных характеристик.

Эти изменения спектральных характеристик подчеркивают важность природы и структуры органических соединений, участвующих в реакции. В частности, продукты взаимодействия с изомерными аминофенолами имеют разные спектры в зависимости от их химической структуры. Например, взаимодействие с 4-аминофенолом сопровождается окислительной конденсацией, что приводит к образованию хининиминового красителя с увеличением длины сопряженной системы связей. Эти особенности делают диазотированный ППУ универсальным инструментом, подходящим для использования в аналитической химии.

Дополнительно установлено, что продукты азосочетания, полученные на основе ППУ, участвуют в таутомерных равновесиях, меняя свои спектральные характеристики в зависимости от pH среды. Например, изменение pH приводит к батохромному смещению полос поглощения и увеличению их интенсивности. Это свойство может быть использовано для повышения чувствительности и селективности методов анализа, позволяя улучшать распознавание органических соединений [5].

Практическая ценность метода была подтверждена в определении органических соединений и нитрит-ионов. Для определения органических веществ, таких как 1-нафтиламин, 4-аминофенол, фенолы и их производные, были разработаны методики, которые позволяют достичь высоких чувствительностей. Пределы обнаружения составляют единицы микрограммов на миллилитр, что сравнимо с классическими спектрофотометрическими методами, но с преимуществами удобства и простоты в применении.

Определение нитрит-ионов основано на реакции диазотирования концевых групп ППУ нитритами в кислой среде, с последующим азосочетанием с органическими соединениями. Данный метод демонстрирует высокую чувствительность, достигая пределов обнаружения на уровне единиц нанограммов на миллилитр [2]. Это

позволяет использовать разработанный подход для анализа воды, в том числе питьевой и технической, а также для диагностики заболеваний. В частности, исследование конденсата выдыхаемого воздуха показало, что содержание нитритов может быть индикатором различных легочных патологий, таких как хроническая обструктивная болезнь легких, бронхиальная астма и пневмония.

Таким образом, диазотированный ППУ представляет собой универсальный аналитический инструмент, способный заменить традиционные методы химического анализа. Его преимущества включают простоту выделения продуктов реакции, возможность варьирования условий синтеза для получения желаемых характеристик, а также широкий спектр применений в области медицинских исследований, экологического мониторинга и фармацевтического анализа [4].

Среди инновационных форм реагентов важное место занимают наночастицы благородных металлов, таких как золото и серебро. Эти наноструктуры обладают уникальными оптическими свойствами, основанными на явлении локального поверхностного плазмонного резонанса. Возможность формирования устойчивых коллоидных растворов, а также чувствительность к изменению спектральных характеристик при взаимодействии с малыми количествами веществ открывают перед наночастицами широкие перспективы. Например, их способность изменять положение полосы поверхностного плазмонного резонанса и ее амплитуду при агрегации делает наночастицы золота и серебра мощным инструментом для детектирования в спектрофотометрии.

В других исследованиях предлагается метод, использующий наночастицы золота и серебра в спектрофотометрическом анализе [14; 46]. Этот подход базируется на уникальных оптических свойствах наночастиц, обусловленных явлением поверхностного плазмонного резонанса (ППР). Оно проявляется в формировании

интенсивных полос поглощения в видимой области спектра (520 нм для золота и 400 нм для серебра). Эти свойства делают наночастицы удобными инструментами для определения веществ, вызывающих их образование, окисление или агрегацию. Подход особенно применим для аналитических задач, где требуется высокая чувствительность и избирательность.

Первый из исследованных процессов включает формирование наночастиц при взаимодействии с восстановителями. Например, в присутствии флавоноидов или аскорбиновой кислоты в растворе происходит образование наночастиц серебра, что приводит к возникновению полосы ППР в спектре. Это можно использовать для определения низких концентраций восстановителей. На основе разработанных методов удалось оценить влияние природы восстанавливающих соединений, их концентрации и условий раствора на образование наночастиц. Например, формирование наночастиц золота и серебра на пенополиуретановой матрице (ППУ), модифицированной предшественниками наночастиц, также оказалось эффективным методом. При взаимодействии с растворами аскорбиновой кислоты наблюдалось изменение цвета матрицы и появление полос ППР, что свидетельствует об образовании наночастиц на поверхности.

Методы были протестированы на практике для анализа аскорбиновой кислоты в фармацевтических препаратах. Полученные результаты продемонстрировали точность и согласованность с заявленными производителями данными. Аналогично были исследованы возможности применения таких наночастиц для определения окислителей. Например, наночастицы серебра на ППУ теряли цвет и оптические свойства в присутствии солянокислых растворов окислителей. Это свойство позволяет рекомендовать такие системы для анализа содержания железа (III), дихромат-ионов и других окислителей [3].

Особое внимание уделяется процессам агрегации наночастиц. В результате агрегации изменяются оптические свойства наночастиц, что проявляется в изменении их окраски. Этот эффект нашел применение для определения соединений, не имеющих хромофорных групп, таких как насыщенные тиолы или полиэлектролиты. Исследование продемонстрировало, что агрегативные свойства наночастиц можно регулировать за счет изменения pH, введения дополнительных реагентов, а также благодаря специфике поверхности модифицированных ППУ [48].

Синтез и функционализация наночастиц проводились с использованием различных методов, включая цитратный и боргидридный. Были разработаны новые способы стабилизации наночастиц, которые обеспечивают их устойчивость и оптимальные оптические свойства. Также исследованы нанокомпозиты на основе пенополиуретана, которые демонстрируют стабильную сорбцию наночастиц, высокую степень извлечения и возможность регулирования агрегативных процессов.

Таким образом, методика использования наночастиц золота и серебра позволяет эффективно решать задачи аналитической химии. Ее основные преимущества включают высокую чувствительность, широкий диапазон определяемых концентраций и возможность адаптации к различным условиям. Метод находит применение как в традиционных спектрофотометрических анализах, так и в новых областях, таких как создание тест-систем и визуальных индикаторов.

1.5. Общие проблемы спектрометрии

Методы оптической молекулярной абсорбционной спектроскопии, несмотря на их популярность и широкую применимость, имеют определенные недостатки, которые могут ограничивать их использование в ряде случаев. Одной из главных

проблем является ограниченная чувствительность к очень низким концентрациям анализируемых веществ. Хотя эти методы подходят для многих рутинных анализов, их способность обнаруживать минимальные количества вещества уступает другим высокочувствительным методам, таким как масс-спектрометрия [36] или атомно-эмиссионная спектроскопия [16]. Это может быть критичным при анализе сверхчистых материалов или следовых концентраций токсичных загрязнителей.

Другой сложностью является влияние матричных эффектов, когда компоненты окружающей среды, присутствующие в образце, искажают результаты измерений. Например, поглощение или рассеяние света примесями или сопутствующими соединениями может приводить к значительным ошибкам в определении концентрации целевого вещества [37]. Это особенно важно в сложных многокомпонентных системах, таких как биологические жидкости или промышленные смеси, где требуется учитывать взаимодействие множества факторов.

Методы оптической спектроскопии также чувствительны к физическому состоянию образца и его оптическим свойствам. Неоднородность или высокая мутность материала, например, может затруднить прохождение света через образец, вызывая погрешности при регистрации спектра. Твердые или полутвердые образцы зачастую требуют специальной подготовки, что увеличивает время и сложность анализа [50].

Еще одним недостатком является ограниченная селективность. Спектры различных веществ могут частично перекрываться, особенно в сложных смесях, что затрудняет их точное распознавание. Хотя современные алгоритмы обработки данных способны справляться с такими ситуациями, они требуют мощных вычислительных ресурсов и хорошо обученных моделей, что может быть недоступно в стандартных лабораториях или полевых условиях. Также стоит отметить, что методы оптической спектроскопии в ряде случаев требуют строгого соблюдения условий измерений.

Например, стабильность температуры, давления и других параметров среды может существенно влиять на точность анализа. Отклонения от идеальных условий могут привести к вариациям в спектрах, что усложняет интерпретацию результатов [12].

Кроме того, оборудование для оптической молекулярной спектроскопии, хотя и считается относительно доступным, все же имеет свои ограничения. Высокоточные приборы для работы в ультрафиолетовом, видимом или инфракрасном диапазонах зачастую требуют значительных инвестиций, регулярного обслуживания и калибровки. Это может стать препятствием для внедрения метода в небольших лабораториях или при работе в удаленных регионах [51].

Наконец, методы оптической спектроскопии могут быть малоэффективны для веществ, которые не обладают ярко выраженными оптическими свойствами, такими как низкая поглощательная способность или слабое взаимодействие с излучением в изучаемом диапазоне. Для таких случаев требуются дополнительные модификации образцов, например использование реагентов для формирования окрашенных соединений, что усложняет и удорожает анализ.

Таким образом, несмотря на свои многочисленные преимущества, методы оптической молекулярной абсорбционной спектроскопии имеют ряд недостатков, связанных с ограничениями чувствительности, селективности и зависимости от условий анализа. Эти проблемы требуют дополнительных исследований и разработки усовершенствованных подходов для их преодоления.

1.6. Проблематика обработки информации для спектрального анализа

Разработка и применение эффективных методов идентификации спектров способствуют быстрому и достоверному анализу состава веществ в лабораторных и

полевых условиях. Современные методы спектральной идентификации развиваются на пересечении физических, химических и информационных наук. Применение вычислительных алгоритмов, таких как корреляционный анализ и методы машинного обучения, значительно расширяет возможности по обработке и интерпретации спектральных данных. Важным аспектом является оптимизация отношения сигнал/шум, что позволяет извлекать полезную информацию даже из данных, полученных при ограниченной чувствительности приборов, а также учитывать вариации спектральных характеристик, вызванные внешними факторами.

Особое внимание в исследовательской практике уделяется разработке новых алгоритмов, способных преобразовывать и интерпретировать спектральные данные для повышения надежности идентификации. Это включает создание эталонных баз данных, применение методов интегрального преобразования и использование численных моделей для предсказания спектральных характеристик. Комплексный подход, сочетающий совершенствование аппаратных средств и внедрение передовых программных алгоритмов, открывает перспективы для создания портативных устройств, способных выполнять экспресс-анализ без предварительной подготовки образцов.

Автоматизированная идентификация и кластеризация веществ по спектральным данным – это не новая, но быстро развивающаяся область на стыке хемометрики, спектроскопии и машинного обучения. Классические хемометрические конвейеры (PCA/PLS, SIMCA/PLS-DA) остаются основными инструментами для надежной, интерпретируемой классификации, в то время как современные методы машинного обучения и глубокого обучения обещают более высокую точность, сквозную автоматизацию и улучшенную производительность при решении очень больших или сильно нелинейных задач [71; 113; 129; 136; 139; 150; 161; 162].

Спектроскопические измерения (рамановские; инфракрасные – ИК-Фурье, ближний ИК, средний ИК; УФ-видимые; масс-спектрометрия и другие) дают насыщенные, многомерные сигналы, кодирующие молекулярную структуру, состав и микроокружение образца. Преобразование этих сигналов в автоматизированные идентификаторы или кластеры химических веществ критически важно для различных приложений – от дистанционного зондирования и горнодобывающей промышленности до безопасности пищевых продуктов, фармацевтики и клинической диагностики. Эту задачу можно разделить на две взаимосвязанные задачи: идентификация (отнесение наблюдаемого спектра к известному веществу или классу) и кластеризация (группировка спектров в когерентные подмножества, которые могут соответствовать химическим семействам, смесям или немаркированным классам). Обе задачи сталкиваются с общими техническими трудностями: изменчивая базовая линия и флуоресценция, эффекты рассеяния, дрейф прибора, перекрывающиеся полосы, ограниченное количество маркированных данных и необходимость надежной, обобщающей валидации.

1.6.1. Предварительная обработка

В литературе единогласно утверждается, что предварительная обработка оказывает решающее влияние на производительность последующих этапов: такие варианты, как удаление базовой линии, сглаживание, нормализация, удаление космических лучей, пиков и коррекция рассеяния, часто сильнее изменяют рейтинги модели, чем выбор классификатора.

Типичные этапы предварительной обработки (часто применяемые в стандартизированных последовательностях) включают в себя:

- Выбор/кадрирование волнового числа (удаление шумных спектральных краев).
- Удаление пиков, космических лучей в спектрах комбинационного рассеяния – медианной фильтрацией или удалением выбросов на основе модели.
- Сглаживание и шумоподавление для снижения высокочастотного шума с сохранением формы пиков, например, фильтрация Савицкого-Голея, Режекторная и Гильберт-фильтрация.
- Коррекция базовой линии и флуоресценции для удаления медленно меняющегося фона, например, асимметричный метод наименьших квадратов, сглаживание Уиттекера, метод наименьших квадратов со штрафом.
- Коррекция рассеяния и мультипликативных эффектов, такая как стандартная нормальная дисперсия (SNV) и мультипликативная коррекция рассеяния (MSC).
- Производные спектров (первая/вторая производная) для выделения узких особенностей и снижения вклада широких базовых линий.
- Нормализация для уменьшения различий в шкале интенсивности.

Фундаментальный обзор 2015 года рассматривает широкий набор методов, адаптированных для анализа данных рамановской и инфракрасной спектроскопии. Среди них – коррекция базовой линии, многомерное разрешение кривых, многоканальный анализ изображений и применение хемометрических моделей. Этот обзор до сих пор служит концептуальной основой для понимания рабочих процессов многомерной спектральной обработки и показывает, каким образом базовые операции предобработки формируют основу последующего анализа [106].

Дальнейшее развитие этого направления отражено в систематическом обзоре [162], где проводится каталогизация алгоритмов предварительной обработки,

обсуждаются их теоретические основания и сравнительная эффективность на реальных наборах данных. Авторы делают акцент на том, что выбор методов предварительной обработки должен быть эмпирическим и сопровождаться перекрёстной проверкой, поскольку априорные решения часто оказываются неоптимальными.

Важным дополнением стало исследование 2023 года, где была опубликована открытая реализация алгоритма удаления базовой линии. Тестирование на множестве наборов данных показало, что этот подход значительно повышает устойчивость и воспроизводимость последующей классификации, предоставляя исследователям практический инструмент для интеграции в автоматизированные конвейеры обработки [150].

Существенный вклад в методологию предобработки внёс эмпирический эксперимент 2020 года, в рамках которого была предложена специализированная методика калибровки одномерных свёрточных нейронных сетей (1D-CNN) для данных ближней инфракрасной спектроскопии. Авторы показали, что благодаря обучению фильтров, улавливающих локальные спектральные паттерны, можно существенно снизить ошибку предсказания по сравнению с традиционными методами, такими как PLS. Однако одновременно подчёркивалось, что эффективность подхода напрямую зависит от объёма обучающих данных и необходимости применения строгих методов регуляризации [92].

Появляется всё больше проверенных открытых пакетов предварительной обработки. Для рамановских спектров несколько открытых инструментов реализуют надёжное удаление базовых линий, проверенное на наборах данных тканей/биологических жидкостей, а пакеты, разработанные сообществом (RamanSPy, PyFasma, RamPy и другие), предлагают интегрированные конвейеры для удаления пиков, определения базовых линий и нормализации. Важность общих проверенных

процедур подчёркивается исследованиями, которые показывают, что вариабельность предварительной обработки в разных группах может приводить к невоспроизводимой эффективности модели. Пакет с открытым исходным кодом для удаления базовых линий, проверенный на нескольких наборах данных человека, является ярким примером усилий сообщества по валидации и воспроизводимости [108; 150].

1.6.2. Снижение размерности и проектирование признаков

Спектры имеют высокую размерность, но строго структурированы: пики, формы полос и коррелированные признаки снижают внутреннюю размерность. Снижение размерности обеспечивает фильтрацию шума, сжатие, визуализацию и иногда улучшенную классификацию.

Среди линейных методов снижения размерности наиболее популярными являются методы PCA и PLS. Анализ главных компонент (PCA) остаётся стандартным исследовательским инструментом для спектральных наборов данных: визуализация дисперсии, обнаружение выбросов и предварительное вычисление ортогональных базисов для последующей классификации. Метод частичных наименьших квадратов (PLS) (и PLS-DA для задач с учителем) широко используется, когда целевой объект прогнозирования является непрерывным (PLS-регрессия) или категориальным (PLS-DA). PLS особенно популярен для БИК-спектроскопии благодаря своей способности использовать ковариацию между предикторами и откликами для формирования информативных латентных переменных. В нескольких систематических обзорах и методологических работах отмечается, что объединение латентных оценок PLS в качестве признаков в гибких классификаторах (SVM, случайный лес) часто повышает

производительность по сравнению с необработанными спектрами [72; 74; 107; 109; 113; 131; 143; 147; 156; 164].

Нелинейные методы и многообразные методы снижения размерности, такие как t-SNE и UMAP, обычно используются для визуализации и обнаружения кластеров; автокодировщики (включая вариационные автокодировщики) применяются для шумоподавления и извлечения признаков без учителя, особенно при наличии больших неразмеченных наборов данных. Глубокие автокодировщики также служат для предварительного обучения последующих моделей с учителем в режимах с малым количеством данных. В литературе подчеркивается необходимость осторожного подхода при интерпретации нелинейных визуализаций снижения размерности для принятия решений о кластеризации, так как существуют риски появления артефактов встраивания [27; 74; 106; 107; 124; 147; 151].

Обзор Ерохина С. Д. из МТУСИ и других авторов 2022 года посвящён методам снижения размерности и отбора признаков применительно к системам обнаружения вторжений, однако представляемая таксономия методов имеет прямую релевантность для спектроскопических задач. Авторы обосновывают необходимость снижения размерности, разделяя методы отбора на стандартные классы: методы-обёртки, методы фильтрации, встроенные (вложенные) методы и гибридные подходы. В статье подробно описывается математическая идея, лежащая в основе нескольких распространённых фильтров; объясняются стратегии поиска обёрток и их вычислительные преимущества; обосновывается целесообразность гибридных конвейеров при работе с большими зашумлёнными наборами данных, таких как журналы систем обнаружения вторжений. Приводимые примеры и ссылки на практические алгоритмы отбора иллюстрируют преимущества снижения размерности. Несмотря на предметную направленность обзора на кибербезопасность, авторы подчёркивают, что таксономия и конкретные алгоритмы, обсуждаемые в

работе, непосредственно связаны с распространёнными хемометрическими проблемами – высокой размерностью, сильной коллинеарностью и ограниченным числом маркированных образцов – что делает предложенные подходы применимыми и полезными в спектроскопии [19].

1.6.3. Классическая хемометрическая классификация

Классическая хемометрика предоставляет интерпретируемые, хорошо зарекомендовавшие себя методы, подходящие для рутинного контроля качества и областей, требующих прослеживаемости (анализ пищевых продуктов, фармацевтика).

PLS-DA: широко используется для контролируемой классификации спектров; выигрывает от снижения размерности и интерпретируемости благодаря нагрузкам.

Линейный/квадратичный дискриминантный анализ (LDA/QDA) и машины опорных векторов (SVM) являются стандартными базовыми методами; SVM часто хорошо работают при использовании подходящих ядер и тщательного масштабирования.

Подходы мягкого независимого моделирования по аналогии классов (SIMCA) и одноклассовая классификация (OCC) имеют отдельную литературу, посвященную их теории и применению в спектральной классификации. Одноклассовая классификация особенно важна в сценариях нецелевого обнаружения, например, для выявления фальсификации или обнаружения аномалий. В недавних обзорах обобщены современные методы моделирования классов [68; 70; 75; 83; 97; 98; 106; 116; 122; 130; 133; 138; 140–142; 146; 148; 152; 153].

Алгоритмы без учителя (иерархическая кластеризация, k-means, DBSCAN, спектральная кластеризация) используются для группировки схожих спектров,

выявления подкластеров и разведочного анализа данных. Во многих прикладных рабочих процессах кластеризации предшествует PCA или другой метод снижения размерности для удаления шума и сжатия входных данных. В литературе показано, что иерархическая кластеризация с подходящими метриками сходства и расстояния часто используется в хемометрике благодаря своей интерпретируемости, в то время как методы, основанные на плотности (DBSCAN), полезны для выявления выбросов и гетерогенного шума [106].

1.6.4. Подходы машинного и глубокого обучения к анализу данных

В последние годы наблюдается распространение методов машинного и глубокого обучения, а также сравнительных исследований, демонстрирующих как перспективы, так и недостатки. Обзоры показывают, что глубокое обучение позволяет извлекать сложные признаки из необработанных или слабо предварительно обработанных спектров, но при этом поверхностные модели остаются конкурентоспособными при ограниченном количестве данных или хорошо настроенной предобработке [69; 73; 86; 90; 91; 160].

Классические методы машинного обучения: градиентно-бустированные деревья (XGBoost, LightGBM), случайные леса и ядерные SVM по-прежнему широко используются благодаря надежности, интерпретируемости (важности признаков) и высокой производительности на табличных наборах спектральных признаков. Гибридные конвейеры, с предварительной обработкой и снижением размерности (PCA/PLS) сопровождаются классификаторами на основе деревьев, часто обеспечивают выгодный компромисс между точностью и интерпретируемостью. В частности, обзоры в области масс-спектрометрии и ближнего инфракрасного

излучения показывают, что методы древовидной структуры часто превосходят глубокие сети при работе с данными малого и среднего объёма [78; 88; 96; 117; 134; 155; 157; 159].

Методы глубокого обучения, свёрточные нейронные сети (CNN), адаптированные к одномерным спектрам или двумерным представлениям, получили широкое применение. Архитектуры варьируются от поверхностных одномерных свёрточных нейронных сетей до более глубоких многоветвевых и основанных на внимании моделей. Глубокие модели обеспечивают автоматическое извлечение признаков и иногда превосходную точность на больших гетерогенных наборах данных. Обзоры также сообщают о многообещающих результатах для биомедицинской рамановской классификации, но предупреждают о переобучении и необходимости тщательной валидации. В одном известном экспериментальном исследовании была разработана глубокая калибровочная CNN для данных ближнего инфракрасного излучения, которая показала значительное снижение ошибки калибровки при выполнении конкретных задач, что иллюстрирует важность выбора архитектуры и адаптации предметной области [92; 110; 124; 151; 163; 167].

Рекуррентные нейронные сети (RNN), ориентированные на анализ последовательностей, применяются для интерпретации временных зависимостей или многомерных векторов, где каждый элемент соответствует интенсивности сигнала на определённой частоте. Глубокие нейронные сети (DNN), благодаря многослойной структуре, обеспечивают высокую точность в задачах классификации и регрессии, особенно при работе с большими объёмами данных [42].

Особый интерес вызывает комбинирование архитектур, например, интеграция CNN и RNN. Подобный симбиоз позволяет объединить преимущества пространственного анализа и временной динамики. Как показывают исследования, CNN эффективно выделяют локальные паттерны в спектрах, а RNN обрабатывают их

как последовательности, выявляя долгосрочные зависимости. Такой подход демонстрирует повышенную эффективность в задачах прогнозирования свойств материалов или анализа химических соединений.

Методы моделирования классов (SIMCA, одноклассовый SVM, методы выпуклых оболочек) рассматриваются и позиционируются как необходимые, когда доступны только данные целевого класса или когда целью является обнаружение аномалий (например, фальсификации). В недавнем обзоре одноклассовых классификаций обобщены алгоритмические варианты и отмечено, что их эффективность зависит от надежного моделирования внутриклассовой изменчивости и консервативных пороговых значений для контроля ложных тревог. Обзор непосредственно актуален для случаев использования нецелевого обнаружения [152].

Применение нейронных сетей для спектрального анализа уже находит применение в идентификации материалов, обнаружении аномалий, классификации объектов и биомедицинских исследованиях. Например, в гиперспектральной визуализации нейросетевые модели позволяют точно определять химический состав веществ, что критически важно для экологического мониторинга и фармацевтики. Преимущество нейронных сетей перед традиционными методами заключается в их способности обучаться на неструктурированных данных и выявлять скрытые закономерности, недоступные для классических алгоритмов [119].

Перспективы данного направления связаны с дальнейшей оптимизацией архитектур, разработкой гибридных моделей и расширением областей применения. Интеграция методов глубокого обучения с физико-математическими моделями спектральных процессов может повысить интерпретируемость результатов. Кроме того, внедрение методов трансферного обучения и обработки данных в реальном времени открывает возможности для создания адаптивных систем анализа, востребованных в промышленности и науке. Таким образом, исследование нейронных

сетей для обработки спектральных данных остаётся одним из ключевых направлений в этой области. В обзоре 2021 года прослеживается переход от традиционной хемометрики к современным алгоритмам машинного и глубокого обучения. В нём обсуждаются лучшие практики работы с данными: правильная предварительная обработка, применение перекрёстной проверки, предотвращение переобучения. Особое внимание уделено гибридным рабочим процессам, в которых хемометрическая предобработка сочетается с возможностями машинного обучения, что позволяет использовать как знания предметной области, так и алгоритмическую гибкость [113].

Комплексный обзор 2022 года, посвящённый ближней инфракрасной спектроскопии, систематически анализирует стратегии проектирования признаков, сравнивает популярные классификаторы и документирует области применения. В обзоре подчёркивается, что несмотря на рост популярности ансамблей деревьев решений и нейронных сетей, методы PLS и SVM остаются актуальными благодаря своей устойчивости и проверенной эффективности. Авторы приводят сравнительные таблицы, где обобщаются показатели точности и рекомендуемые конвейеры обработки для различных прикладных сценариев [164].

Статья 2022 года по глубокому обучению для рамановской спектроскопии подробно рассматривает архитектуры свёрточных нейронных сетей, автокодировщики для шумоподавления и стратегии переноса обучения. Авторы акцентируют внимание на биомедицинских приложениях, где глубокие модели демонстрируют высокую чувствительность и специфичность, но одновременно предупреждают, что воспроизводимость результатов сильно зависит от качества исходных данных и применяемых процедур предобработки. Отдельно подчёркивается необходимость сопровождать публикации не только метриками точности, но и матрицами ошибок, чтобы обеспечить прозрачность оценки [126].

Клинически ориентированный обзор 2022 года по применению методов рамановского машинного обучения в онкологии суммирует результаты множества исследований, выделяет типичные проблемы – малые размеры выборок, нестабильная предобработка, ограниченная внешняя валидация – и показывает, что несмотря на перспективность глубоких моделей, их внедрение в клинику возможно только при условии строгой валидации и высокой интерпретируемости [80].

Серьёзное внимание уделяется интерпретируемости алгоритмов машинного обучения. В обзоре 2025 года по ближней инфракрасной спектроскопии анализируются современные методы интерпретируемости, включая VIP-оценки для PLS, SHAP для деревьев решений и объяснения на основе градиентов для нейронных сетей. Эти методы позволяют визуализировать, каким образом предсказательные модели соотносят свои решения с конкретными спектральными областями, что облегчает понимание химических процессов. Авторы подчёркивают, что интерпретируемость является не факультативной, а обязательной характеристикой моделей для таких прикладных областей, как контроль подлинности пищевых продуктов и фармацевтический анализ, где результаты должны быть прозрачными и воспроизводимыми. Другие более ранние исследования интерпретируемости также посвящены основному ограничению многих моделей машинного обучения, используемых в спектральном анализе: необходимости объяснять решения, принимаемые при выборе модели, в терминах спектральных характеристик или химических полос [76; 77; 127; 158].

1.6.5. Валидация, воспроизводимость и факторы ненадёжности

Валидация – наиболее часто упоминаемый методологический недостаток спектрального машинного обучения. Многочисленные недавние систематические обзоры и эмпирические исследования документируют распространённые факторы ненадёжности: утечка данных из-за предварительной обработки, некорректные стратегии кросс-валидации и настройка гиперпараметров без вложенной кросс-валидации, приводящая к оптимистичным ошибкам. Несколько недавних статей и обзоров содержат структурированные рекомендации и эмпирически сравнивают процессы валидации [85; 101; 118; 125; 135].

Выделяются следующие распространенные ошибки валидации и способы их устранения:

- Утечка данных из-за предварительной обработки: выполнение нормализации или PCA на полном наборе данных перед разделением создаёт оптимистичные результаты. Способ устранения: подгонка преобразователей предварительной обработки только к обучающим данным и применение к данным валидации и тестирования [162].
- Коррелированные измерения: повторные измерения из одного и того же сосуда, пациента или производственной партии должны быть сгруппированы таким образом, чтобы все реплики попадали в один и тот же срез кросс-валидации. Игнорирование корреляции завышает кажущуюся точность [101].
- Смещение выбора гиперпараметров: для получения объективных оценок производительности настройка гиперпараметров должна быть вложена во внешнюю кросс-валидацию. Многие эмпирические исследования, сравнивающие конвейеры, демонстрируют значительное падение

производительности при принудительном применении вложенной кросс-валидации [125].

Систематические обзоры рекомендуют использовать: воспроизводимые конвейеры, инкапсулирующие предварительную обработку и моделирование; вложенную кросс-валидацию или контрольные тестовые наборы для окончательных оценок; групповое разделение, когда это целесообразно; отчетность о балансе классов и доверительных интервалах; публикацию кода и начальных значений для воспроизводимости. Недавнее сравнительное исследование наборов рамановских данных наглядно демонстрирует, как различные варианты проверки приводят к существенно различающимся результатам, и предлагает надежный рабочий процесс, сочетающий вложенную кросс-валидацию и этапы спектрального контроля качества [125; 150].

1.6.6. Прикладные исследования в области спектроскопии.

Важным направлением остаются прикладные исследования, где спектроскопия используется в конкретных отраслях. В статье 2022 года показано, что современные хемометрические конвейеры позволяют возродить интерес к УФ-видимой спектроскопии для нецелевого анализа спектральных отпечатков. Благодаря корректной предварительной обработке и многомерному распознаванию образов данный метод оказывается применимым для задач обнаружения фальсификаций и классификации сложных смесей [141].

Отечественные исследования также демонстрируют значительное разнообразие подходов. В работе Кочикова И. В. из МГУ им. М. В. Ломоносова и других исследователей 2007 года была разработана методика идентификации химических

веществ по инфракрасным спектрам на основе гибридной базы данных ИСМОЛ, включающей как экспериментальные, так и квантово-химически рассчитанные спектры. Предлагался двухуровневый анализ: поиск спектрально сходных соединений и определение состава многокомпонентных смесей. Для идентификации смесей применялся итерационный алгоритм минимизации квадратичного функционала, что позволяло учитывать до 3–5 компонентов одновременно и повышало устойчивость решений [26].

В исследовании Битюкова В. К. и Хвостова А. А. 2015 года изучалось применение ультразвуковой спектроскопии для автоматизированного контроля качества гомогенизации молока. Авторы связали акустические характеристики среды, такие как скорость звука и коэффициент поглощения, с распределением жировых шариков по размерам. Экспериментальная часть включала микрофотографию и ультразвуковые измерения, которые показали высокую корреляцию (до 0.99) между параметрами распределения и релаксационными спектрами, что подтвердило эффективность метода для оперативного мониторинга качества [6].

Работа Васильева Н. С. и других исследователей из МГТУ им. Баумана 2015 года была направлена на разработку алгоритма повышения надёжности идентификации химических веществ по спектрам вторичного излучения в условиях низкого отношения сигнал/шум. Авторы использовали люминесцирующие вещества в качестве объектов и предложили метрику, основанную на косинусе угла между спектральными векторами. Сравнительный анализ различных методов показал, что комбинирование данных из разных спектральных диапазонов значительно повышает точность, а наилучшие результаты достигались при использовании гиперболических границ в пространстве признаков [10].

В исследовании 2019 года предложена методика идентификации химических соединений с использованием спектров рассеянного инфракрасного излучения,

регистрируемых с помощью перестраиваемого квантово-каскадного лазера. Авторы показали, что применение преобразования Крамерса-Кронига позволяет пересчитать спектры рассеяния в спектры поглощения, которые обладают большей селективностью. При идентификации веществ вероятность правильного обнаружения достигала 100% для ряда соединений, а на реальных подложках надёжность варьировала от 50% до 83% [13].

В прикладных медицинских исследованиях стоит отметить работу Павлова В.Н. и соавторов 2018 года, где рамановская спектроскопия в сочетании с алгоритмами машинного обучения использовалась для диагностики рака. Авторы детально описали процесс предварительной обработки данных (коррекция базовой линии, сглаживание, нормализация) и сравнили различные классификационные подходы. Представленные клинические примеры показали потенциал метода для автоматизированного скрининга [43].

В исследовании Носенко Т. Н. из ИТМО и других исследователей 2019 года ИК-Фурье спектроскопия сыворотки крови применялась для выявления биохимических различий у пациентов с эпилепсией. С использованием методов многомерной статистики (PCA, PLS-DA) были выявлены спектральные паттерны, отличающие пациентов от контрольной группы. Работа показала, что ИК-спектроскопия может быть использована для поиска биомаркеров и имеет потенциал в клинической практике [41].

В статье Таныковой Н. Г. и других от 2024 года рассматривается применение ИК-Фурье спектроскопии для анализа осадочных пород, описываются методологические процессы и приводятся конкретные геохимические примеры. В ней рассматриваются типичные этапы предварительной обработки и анализа данных ИК-Фурье, используемые в геологии (распределение полос, коррекция базовой линии, деконволюция, использование многомерного анализа для идентификации минералов

и оценки состава), а также приводятся примеры, демонстрирующие, как данные ИК-спектроскопии дополняют петрографические и геохимические измерения. Статья носит практический характер, ориентирована на специалистов в области геологии и демонстрирует, как автоматизированная спектральная обработка ускоряет скрининг образцов и минералогическую классификацию, что особенно важно при работе с большими сериями геологических проб. В тексте подробно обсуждаются этапы: распределение полос как средство выделения ключевых спектральных маркеров, коррекция базовой линии для удаления низкочастотных артефактов, деконволюция для разделения перекрывающихся компонент и применение многомерного анализа для интеграции результатов и уточнения состава образцов. Приводимые геохимические примеры иллюстрируют, как совокупность методов позволяет повысить точность минералогической идентификации и сопровождают практические рекомендации по постановке эксперимента и выбору процедур предобработки [47].

В статье Саакян А. В. и Левина А. Д. 2024 года представлен обзор российских программных инструментов и реализаций для обработки спектральных данных и хемометрики. В ней описываются функции, часто реализуемые в таких пакетах: сглаживание спектральной плотности (фильтр Савицкого-Голея), SNV/MSC, коррекция базовой линии, выбор пиков, PCA/PLS, модули классификации; приводятся примеры использования, а также обсуждаются пользовательские интерфейсы и возможности пакетной обработки, важные для автоматизации [45].

1.7. Экспертно-нейросетевые системы для анализа данных

Экспертно-нейросетевые системы (ЭНС) представляют собой интегрированный класс интеллектуальных систем, сочетающих методы экспертных систем и

искусственных нейронных сетей. Такая интеграция позволяет объединить преимущества формализованного представления знаний с возможностями адаптивного обучения и обработки больших данных [15].

Экспертная система является компонентом, обеспечивающим хранение и обработку знаний предметной области. Она включает базу данных, содержащую фактическую информацию, базу знаний, в которой формализованы правила и закономерности, а также машину логического вывода, предназначенную для получения новых знаний и принятия решений. В экспертной системе ключевым элементом является использование экспертных правил, созданных на основе опыта специалистов, что обеспечивает объяснимость и прозрачность принимаемых решений.

Нейросетевая компонента, напротив, ориентирована на обработку сложных, высокоразмерных и зашумлённых данных, которые трудно поддаются строгой формализации. Искусственные нейронные сети обладают свойством обучения на примерах, что позволяет им выявлять скрытые зависимости и закономерности, не всегда доступные для классических методов логического вывода. Благодаря этому они повышают устойчивость системы к ошибкам и неопределённости входных данных.

Интеграция экспертной системы и нейросетевых технологий обеспечивает синергетический эффект. Экспертная часть позволяет структурировать и интерпретировать поступающие данные, формируя предварительные гипотезы или сокращённые векторы признаков, которые затем подаются на вход нейронной сети. Нейросеть, в свою очередь, уточняет результаты классификации, выполняет аппроксимацию нелинейных зависимостей и повышает достоверность выводов. Такое взаимодействие делает систему более гибкой, надёжной и способной к адаптации в условиях изменяющихся параметров предметной области.

Основные задачи, решаемые экспертно-нейросетевыми системами, включают:

1. Автоматизированный контроль качества – идентификация и оценка характеристик объектов на основе многомерных данных.

2. Классификация и распознавание образов – определение принадлежности объектов к определённым классам в условиях шумных или неполных данных.

3. Прогнозирование параметров и состояний – моделирование и предсказание динамики технологических процессов или свойств материалов на основе накопленных данных.

4. Интеллектуальная поддержка принятия решений – формирование рекомендаций на основе сочетания экспертных правил и статистически обоснованных результатов нейросетевого анализа.

Несмотря на широкие возможности, ЭНС имеют ряд нерешённых проблем, ограничивающих их эффективность. Одной из основных является сложность формирования адекватного признакового пространства для высокоразмерных данных. Особенно остро эта проблема проявляется при работе со спектральными данными, которые могут содержать коррелированные и избыточные признаки, а также быть подвержены шумам, что не позволяет выполнять анализ с высокой точностью.

Другой проблемой выступает интерпретируемость данных и результатов. Экспертные правила могут обеспечивать объяснимость, но они часто оказываются недостаточными для описания сложных многомерных объектов. В результате возникает противоречие между требованием прозрачности экспертной части и адаптивностью нейросетевой.

Отдельно следует выделить проблему формирования цифровых представлений исследуемых объектов. Без качественного преобразования исходных данных экспертная система затрудняется в построении базы знаний, а нейросетевые алгоритмы – в достижении высокой надёжности работы.

Таким образом, основные проблемы ЭНС заключаются в обработке многомерных и зашумлённых данных, обеспечении интерпретируемости и формировании устойчивых цифровых образов. Эти ограничения существенно сдерживают развитие и практическое применение таких систем в сложных прикладных областях. Разработанные в данной работе решения устраняют эти проблемы при внедрении в ЭНС.

1.8. Выводы по первой главе

По итогам рассмотрения материалов первой главы можно сделать следующие выводы. Современные подходы к химическому анализу веществ демонстрируют существенный прогресс в части применения спектральных методов, однако остаются значимые проблемные зоны, снижающие эффективность их использования в задачах контроля качества и идентификации. Наиболее остро проявляется ограниченность традиционных методов обработки данных, которые зачастую не справляются с высокой размерностью и сложной структурой спектров. Это порождает трудности в извлечении информативных признаков и формировании адекватных моделей, отражающих реальную физико-химическую природу исследуемых объектов.

Рассмотрены методы системного анализа, применимые к задаче кластеризации спектральных данных, обеспечивающие комплексный и целостный подход к исследованию многомерных структур. Формальное моделирование измерительной системы и шумовых факторов позволяет корректировать исходные данные и повышать их сопоставимость. Многоуровневая декомпозиция спектров и использование информационных критериев обеспечивают выделение устойчивых и информативных признаков. Графовые представления и спектральные методы создают

возможности для выявления скрытых структур и сообществ, а адаптивные алгоритмы учитывают динамику изменения данных. Дополнение этих методов процедурами оптимизации и системной верификации обеспечивает воспроизводимость и интерпретируемость результатов. В совокупности это делает кластеризацию спектральных данных более надежной, информативной и применимой в широком круге научных и инженерных задач.

Показано, что кластерный анализ нельзя рассматривать вне контекста структуры данных, в рамках которой он применяется. Объектно-признаковая, парная и реляционно-сетевая формы задают различные основания для интерпретации близости, что влечёт за собой разные способы группировки. Таким образом, структурная организация информации выступает не пассивным элементом, а активным фактором, формирующим границы применимости методов и определяющим их аналитическую ценность.

С точки зрения системного анализа выявлено, что современная практика применения методов машинного обучения в аналитической химии сталкивается с рядом вызовов. Прежде всего, речь идёт о необходимости предварительного обучения моделей на репрезентативных наборах данных. В условиях ограниченных обучающих выборок и высокой вариативности состава смесей возникает угроза переобучения, когда модель демонстрирует хорошие результаты на обучающем наборе, но оказывается недостаточно устойчивой к новым данным. Это снижает достоверность прогнозов и ограничивает возможность использования таких моделей в реальном производственном процессе. Дополнительным ограничивающим фактором является отсутствие полноценной визуализации результатов анализа. В условиях работы с многомерными спектральными данными именно визуальное представление промежуточных и итоговых результатов играет ключевую роль для экспертной интерпретации и принятия решений. Недостаток удобных и наглядных инструментов

визуализации приводит к тому, что результаты машинного обучения остаются «чёрным ящиком», что снижает доверие специалистов к системе и ограничивает её интеграцию в практику контроля качества.

Таким образом, анализ показывает, что современное состояние области характеризуется несбалансированностью между высоким уровнем инструментальных методов получения спектральных данных и недостаточной зрелостью методов их обработки и интерпретации. Решение выявленных проблем требует комплексного подхода, включающего разработку методов формирования образов, устойчивых к переобучению алгоритмов машинного обучения и эффективных средств визуализации многомерных данных, что создаёт основу для построения автоматизированных систем контроля качества нового поколения.

Глава 2. Методы исследования

На вход системы подаётся набор экспериментально зарегистрированных спектров X, где каждый спектр x_i представляется вектором значений на m дискретных длинах волн. Основная цель – получить устойчивое и интерпретируемое разбиение множества X на однородные группы спектров, а также сформировать информативное маломерное представление данных Z элементов z_i с размерностью равной d при $d<<m$, пригодное для визуализации и дальнейшей обработки. Для достижения этой цели предлагается следующая последовательность операций: предварительная обработка P, снижение размерности R и алгоритм кластеризации C. Формально это записывается как $x_i \rightarrow P(x_i) \rightarrow R(P(x_i)) = z_i \rightarrow C(Z) = y$, где y – метки кластеров.

Предварительная математическая обработка спектра как числового вектора предназначена для устранения базовой линии, подавления шума и усиления информативных компонент спектра; в её состав могут входить методы сглаживания, корреляционные преобразования, дифференцирование и нормализация, подбираемые в зависимости от свойств конкретных измерений.

Снижение размерности рассматривается как задача сохранения локальной и/или глобальной геометрии данных с минимальными потерями информации, что формализуется через критерии реконструкции для обратимых методов и через меры доверительности и целостности для методов встраивания.

Задача кластеризации формулируется как оптимизация разбиения данных в пространстве, в качестве критериев качества используются внутренняя оценка структуры разбиения, например силуэт кластера, или внешние метрики при наличии меток данных.

В этой главе будут рассмотрены методы предварительной обработки спектральных данных, методы снижения размерности, методы кластеризации, а также метрики расстояния (сходства), используемые как методами снижение размерности, так и методами кластеризации.

2.1. Методы снижения размерности

Многообразное обучение и уменьшение размерности являются основополагающими методами машинного обучения и анализа данных. Эти методы решают проблемы, связанные с данными высокой размерности, включая шум, избыточность и вычислительную неэффективность. Преобразуя данные в пространства с меньшей размерностью, они упрощают анализ, обеспечивают визуализацию и повышают производительность в последующих задачах, таких как кластеризация и классификация.

Высокоразмерным данным присущи несколько проблем, которые в совокупности называются проклятием размерности. По мере увеличения размерности точки данных становятся разреженными, а расстояния между точками теряют возможность их различения. Эта разреженность подрывает традиционные статистические и модели машинного обучения, что приводит к плохим результатам классификации [34]. Уменьшение размерности смягчает эти проблемы, проецируя данные в пространство с меньшей размерностью, которое сохраняет их внутреннюю структуру. Этот процесс уменьшает шум, снижает требования к хранению и упрощает вычисления, сохраняя при этом значимую информацию.

Методы снижения размерности можно в целом разделить на линейные и нелинейные. Линейные методы предполагают, что данные лежат в плоском

евклидовом подпространстве, и используют линейные проекции для захвата дисперсии или корреляций. Примером являются анализ главных компонент. Нелинейные методы, также называемые многообразным обучением, предполагают, что данные лежат на искривленном многообразии меньшей размерности, встроенном в пространство большей размерности. Эти методы направлены на сохранение геометрических или топологических структур, а не линейных отношений.

Многообразное обучение работает в предположении, что высокоразмерные данные лежат на гладком низкоразмерном многообразии. Многообразие – это топологическое пространство, которое локально схоже с евклидовым пространством, но может иметь сложную глобальную структуру. Например, изображения вращающегося объекта образуют многообразие, где близлежащие точки соответствуют схожим позициям.

Основополагающая задача многообразного обучения – аппроксимировать эту внутреннюю геометрию с использованием локальной информации о соседстве точек. Алгоритмы строят графы или распределения вероятностей для моделирования взаимосвязей между точками данных, а затем оптимизируют вложения, которые учитывают эти взаимосвязи.

Ключевые математические принципы методов:

- Теория графов. Данные представляются в виде графов, где ребра означают сходства между точками.
- Разложение собственных значений. Низкоразмерные вложения соответствуют собственным векторам матриц графов, захватывая самые гладкие вариации.
- Геодезические расстояния. Алгоритмы аппроксимируют расстояния вдоль многообразия, сохраняя нелинейные структуры.

Снижение размерности можно сформулировать как задачу оптимизации. Цель состоит в том, чтобы минимизировать ошибку реконструкции, функцию напряжения

или расхождение между распределениями вероятностей в пространствах высокой и низкой размерности.

Примеры задач оптимизации:

- Минимизация потери дисперсии при проецировании данных на главные компоненты.
- Сохранение парных расстояний путем оптимизации напряжения при многомерном масштабировании.
- Снижение расхождения между распределениями в таких методах, как t-распределенное стохастическое встраивание соседей.

Эти формулировки позволяют алгоритмам сбалансировать сохранение локальной и глобальной структуры, гарантируя, что встраивания сохранят значимые отношения.

Современные методы многообразного обучения часто используют вероятностные и графовые представления. Распределения вероятностей определяют отношения между точками, фиксируя сходства в терминах правдоподобия или сродства. Графы описывают эти отношения в виде взвешенных ребер, позволяя алгоритмам распространять информацию по окрестностям.

Ключевые концепции включают вероятностные модели, использующие теорию нечетких множеств для кодирования неопределенностей в отношениях, матрицы Кирхгофа, которые представляют структурные ограничения и направляют оптимизацию, и ядерные методы – преобразования, отображающие данные в пространства признаков, где улучшается линейная разделимость. Эти инструменты формируют основу для алгоритмов, которые моделируют нелинейные структуры и сложные геометрические зависимости.

Многообразное обучение имеет ряд проблем:

- Чувствительность к параметрам. Алгоритмы часто требуют тонкой настройки параметров, таких как размер соседства и ширина ядра.
- Вычислительная сложность. Построение графа, расчеты кратчайшего пути и собственные разложения имеют высокий рост вычислительной сложности с ростом объема данных.
- Шум и выбросы. Методы на основе графов чувствительны к шуму, который может искажать отношения соседства и вложения.

Решение этих проблем требует гибридных подходов, масштабируемых алгоритмов и интеграции с фреймворками глубокого обучения.

Эффективность алгоритмов снижения размерности обычно оценивается несколькими способами:

1. Вычисление ошибки реконструкции, которая измеряет, насколько хорошо исходные данные могут быть реконструированы из вложений.
2. Расчет достоверности, оценивающей, сохраняют ли соседние области в низкоразмерном пространстве отношения в высокоразмерном пространстве.
3. Оценка целостности, определяющей, остаются ли соседние точки в исходных данных близкими после проецирования.
4. Визуальная оценка разделения и группировки кластеров на графиках.

Текущие исследования в области снижения размерности и многообразного обучения фокусируются на разработке алгоритмов, способных обрабатывать огромные наборы данных с помощью методов аппроксимации и распараллеливания, и разработке гибридных алгоритмов, объединяющих многообразное обучение с глубоким обучением для обработки сложных структурированных данных, таких как графики и последовательности.

Алгоритмы многообразного обучения и снижения размерности необходимы для анализа многомерных данных. Используя геометрические и вероятностные фреймворки, эти методы раскрывают низкоразмерные структуры, которые упрощают вычисления, улучшают визуализацию и улучшают обнаружение закономерностей. Хотя проблемы, связанные с масштабируемостью, интерпретируемостью и шумом, сохраняются, текущие исследования обещают решить эти проблемы, укрепляя многообразное обучение как важную составляющую современного машинного обучения. Ниже описаны исследованные алгоритмы [57].

Задача снижения размерности математически может быть сформулирована следующим образом.

Дано:

- Множество спектров $X = \{x_i\}_{i=1}^{N}$, где каждый спектр – это вектор признаков $x_i \in R^m, m \gg 1$.
- Число целевых признаков $d \ll m$.

Задача:

Найти отображение

$$T: R^m \to R^d, \tag{1}$$

такое, что для образцов $z_i = T(x_i)$ сохраняется внутренняя структура мн. X.

Критерии:

Минимизация несоответствия между матрицами расстояний в исходном пространстве и в низкоразмерном:

$$E = \sum_{i,j} w_{ij} \left(d_X(x_i, x_j) - d_z(z_i, z_j) \right)^2, \tag{2}$$

где d_x, d_z – метрики расстояния;

w_{ij} – весовые коэффициенты отношений между объектами.

Ограничения:

- $d \ll m$, что обеспечивает компактное представление.
- Сохранность локальной топологии, то есть сохранение соседств объектов при отображении.

2.1.1. Анализ главных компонент

Анализ главных компонент (Principal Component Analysis, PCA) – это метод линейного снижения размерности, широко используемый в анализе данных и машинном обучении. PCA определяет главные компоненты набора данных, которые являются ортогональными направлениями в пространстве признаков, которые максимизируют дисперсию. Эти направления получаются путем решения задачи разложения на собственные значения или выполнения сингулярные разложения.

Алгоритм начинается с центрирования данных, то есть вычитания среднего значения каждого признака, чтобы все признаки имеют среднее значение, равное нулю и единичную дисперсию. Затем вычисляется ковариационная матрица для измерения взаимосвязей между признаками. Затем вычисляются собственные векторы и собственные значения этой ковариационной матрицы. Собственные векторы представляют главные компоненты, а собственные значения указывают величину дисперсии, охватываемую каждым компонентом. Наконец, данные проецируются на главные компоненты для уменьшения их размерности [115].

Альтернативный метод использует разложение по сингулярным значениям, которое разлагает данные на ортогональные матрицы и сингулярные значения. Этот подход вычислительно более эффективен для больших наборов данных.

Вычислительная сложность анализа главных компонент зависит от используемого метода. Вычисление ковариационной матрицы и выполнение разложения по собственным значениям масштабируется квадратично с числом признаков и линейно с числом образцов. Сингулярная разложение, с другой стороны, снижает эту сложность, делая его более подходящим для наборов данных со многими признаками.

Анализ главных компонент эффективен с точки зрения вычислений и хорошо интерпретируем, что делает его универсальным инструментом для линейного анализа данных. Однако он менее эффективен для данных с нелинейными отношениями.

2.1.2. Многомерное масштабирование

Многомерное масштабирование (Multidimensional Scaling, MDS) – это метод снижения размерности, который стремится сохранить парные расстояния между точками данных при проецировании их в пространство с меньшей размерностью. Первоначально разработанный для визуализации данных о сходстве или несходстве, он стал универсальным инструментом для обучения многообразиям и визуализации данных.

Алгоритм начинается с вычисления начальной конфигурации точек, обычно используя классическое масштабирование или анализ главных компонент (PCA). Создается матрица различий, представляющая парные отношения между точками. Вычисляется начальное вложение, применяя разложение собственных значений к двуцентрированной матрице различий, классические методы масштабирования часто обеспечивают отправную точку. Затем оптимизируются положения точек в уменьшенном пространстве, путем минимизации функции напряжения. На этом этапе

используются итерационные методы, такие как градиентный спуск. Положения обновляются, пока функция напряжения не достигнет предопределенного порогового значения или не будет завершено максимальное количество итераций. В конце оценивается окончательное снижение размерности на устойчивость и точность [82].

Алгоритм минимизирует функцию напряжения, которая вычисляет сумму квадратов разностей между расстояниями в сокращенном пространстве и исходными данными:

$$\sqrt{\sum_{i \neq j = 1,\ldots,n} \left(d_{ij} - |x_i - x_j|\right)^2}, \quad (3)$$

где d_{ij} – элемент двуцентрированной матрицы различий;

x_i – входные векторы.

Процесс оптимизации минимизирует это значение напряжения, гарантируя, что отношения в сокращенном пространстве будут точно соответствовать отношениям в исходном пространстве.

Вычислительная сложность многомерного масштабирования зависит от размера набора данных. Расчет парных расстояний масштабируется квадратично с числом точек, а итеративная оптимизация добавляет дополнительную сложность пропорционально размерности уменьшенного пространства.

Многомерное масштабирование дает интерпретируемые результаты благодаря своему геометрическому представлению. Однако оно может быть вычислительно затратным для очень больших наборов данных и может быть чувствительным к инициализации.

2.1.3. Изометрическое отображение

Изометрическое отображение (Isometric Mapping, Isomap) – это нелинейный метод снижения размерности данных. Метод является одним из самых ранних методов снижения размерности. Он сохраняет расстояния вдоль многообразия, фиксируя базовую геометрию многомерных наборов данных. В отличие от линейных методов, таких как PCA, Isomap особенно эффективен для данных, которые лежат на криволинейных или нелинейных поверхностях. Он широко используется для визуализации и анализа данных, которые демонстрируют сложные геометрические структуры.

Isomap предполагает, что точки многомерных данных распределены вдоль многообразия меньшей размерности, встроенного в пространство большей размерности. Алгоритм оценивает эту маломерную структуру, сохраняя геодезические расстояния, которые аппроксимируют кратчайшие пути вдоль многообразия. Геодезические расстояния лучше отражают истинную геометрию данных, чем прямолинейные расстояния.

Для вычисления этих расстояний Isomap сначала строит граф, соединяющий близлежащие точки на основе локальных расстояний. Затем он оценивает глобальные расстояния, используя алгоритмы нахождения кратчайшего пути, чтобы охватить как локальные, так и глобальные связи в данных [154].

Алгоритм Isomap состоит из трех основных шагов:

1. Построение графа окрестностей путем соединения каждой точки с ее ближайшими соседями с использованием евклидовых расстояний.

2. Расчет геодезических расстояний между всеми парами точек с использованием алгоритмов кратчайшего пути, таких как алгоритм Дейкстры или Флойда-Уоршелла.
3. Применение классического многомерного масштабирования к матрице геодезических расстояний для создания низкоразмерного вложения.

Эти шаги гарантируют, что вложение сохранит внутреннюю геометрию данных.

Вычислительная сложность Isomap складывается из трех этапов. Построение графа окрестностей и поиск ближайших соседей масштабируется квадратично с числом точек данных. Вычисление кратчайших путей может масштабироваться квадратично или кубично в зависимости от используемого алгоритма. Окончательный шаг разложения собственных значений масштабируется квадратично с числом точек и линейно с целевыми размерами. Эта сложность делает Isomap подходящим для наборов данных среднего размера, но может ограничить его применение очень большими наборами данных без дополнительной оптимизации.

Isomap эффективно фиксирует нелинейные структуры и сохраняет как локальные, так и глобальные связи в данных. Он интерпретируем и полезен для выявления сложных геометрических моделей. Однако он может быть требовательным к вычислениям, особенно для больших наборов данных. Он также чувствителен к шуму и выбросам, которые могут исказить граф окрестностей. Выбор количества соседей требует тщательной настройки для балансировки локальных и глобальных связей.

2.1.4. Локальное линейное вложение

Локально-линейное вложение (Locally Linear Embedding, LLE) – это нелинейный метод снижения размерности, который сохраняет локальную геометрию многомерных данных, моделируя каждую точку данных как линейную комбинацию ее соседей. В отличие от линейных методов, таких как анализ главных компонент (PCA), LLE эффективно фиксирует нелинейные структуры и особенно подходит для данных, которые лежат на криволинейном многообразии.

LLE предполагает, что высокоразмерные данные лежат на многообразии меньшей размерности, а локальные линейные отношения между соседними точками примерно отражают глобальную структуру этого многообразия. Алгоритм сохраняет эти отношения во время снижения размерности, минимизируя ошибки реконструкции.

Основная идея заключается в вычислении весов, которые описывают каждую точку с точки зрения ее соседей, а затем нахождении низкоразмерного представления, которое сохраняет эти веса. Сосредоточившись на локальных отношениях, метод фиксирует базовую многообразную структуру [145].

Алгоритм локально-линейного вложения включает три основных шага.

1. Определение ближайших соседей для каждой точки данных. Эти соседи определяют локальное соседство для каждой точки.
2. Вычисление весов, которые реконструируют каждую точку как линейную комбинацию ее соседей. Этот шаг решает ограниченную задачу наименьших квадратов, которая минимизирует ошибку реконструкции, гарантируя, что вес для каждой точки в сумме равен единице.
3. Вычисление вложения меньшей размерности, минимизируя функцию стоимости, которая сохраняет ранее вычисленные веса. Этот шаг использует разложение

собственных значений для генерации окончательного представления данных меньшей размерности.

Вычислительная сложность локально-линейного вложения состоит из трех частей. Поиск ближайших соседей масштабируется квадратично с числом точек данных, что делает его осуществимым для умеренно больших наборов данных. Расчет весов включает решение набора линейных уравнений, который масштабируется кубически с числом соседей. Наконец, разложение собственных значений масштабируется квадратично с числом точек данных и линейно с искомым измерением. В совокупности эти шаги приводят к сложности, которая ограничивает применение локально-линейного вложения к наборам данных среднего размера.

Модифицированное локально-линейное вложение (MLLE) улучшается путем введения нескольких линейно независимых локальных весовых векторов для каждой окрестности. Этот подход повышает надежность, особенно для окрестностей с дефицитом ранга [165].

Локально-линейное вложение Гессе включает ограничения на основе Гессе для лучшего учета сложных структур в данных [99].

Локальное выравнивание касательного пространства (LTSA) делает акцент на выравнивании касательных пространств, а не на сохранении расстояний, что делает его подходящим для учета геометрии многообразия [166].

Локально линейное вложение эффективно учитывает нелинейные структуры и сохраняет локальные отношения, что делает его пригодным для анализа данных с искривленной или сложной геометрией. Он обеспечивает интерпретируемые вложение и требует меньше параметров, чем некоторые другие нелинейные методы. Однако алгоритм чувствителен к шуму, который может искажать локальные окрестности. Алгоритм также борется с несвязными многообразиями и требует

тщательной настройки параметров, таких как количество соседей, для балансировки локальных и глобальных отношений.

2.1.5. Спектральное вложение

Спектральное вложение (Spectral Embedding) – это нелинейный метод снижения размерности данных с сохранением локальных взаимосвязей. Он основан на теории спектральных графов и использует математические методы для раскрытия внутренней геометрии данных. Спектральное вложение особенно эффективно для наборов данных со сложными шаблонами, которые не очень хорошо описываются линейными методами, такими как анализ главных компонент.

Спектральное вложение представляет данные в виде графа, где узлы соответствуют точкам данных, а ребра фиксируют отношения, основанные на мерах сходства. Алгоритм предполагает, что точки данных лежат на гладкой низкоразмерной структуре, и моделирует эту структуру посредством графа [79].

Алгоритм использует матрицу Кирхгофа, математическое представление графа, для кодирования отношений между точками. Тем самым минимизирует функцию стоимости, которая гарантирует, что точки, близкие друг к другу в исходных данных, останутся близкими в сокращенном представлении.

Алгоритм спектрального вложение состоит из четырех основных шагов:

1. Построение графа, вычисляя парные сходства между точками. Это можно сделать с помощью гауссовых ядер или графов ближайших соседей для построения матрицы смежности.
2. Вычисление матрицы степеней, которая записывает сумму весов ребер для каждого узла. Она используется для вычисления матрицы Кирхгофа, которая

кодирует структуру графа. Матрица может быть ненормализованной или нормализованной в зависимости от применения.

3. Выполнение разложения собственных значений на матрице Кирхгофа. Нахождение собственных векторов, связанных с наименьшими ненулевыми собственными значениями, которые представляют самые гладкие изменения вдоль многообразия.
4. Расположение собственных векторов в матрице, где каждая строка соответствует точке в уменьшенном пространстве, обеспечивая окончательное низкоразмерное вложение.

Матрица смежности может быть вычислена следующими способами:

1. Косинусное сходство (cosine) – это мера сходства между двумя ненулевыми векторами, основанная на косинусе угла между ними. Она фокусируется на направленном выравнивании векторов, а не на их величинах. Для векторов X и Y, косинусное расстояние (k) вычисляется как:

$$k(x, y) = \frac{XY^{\top}}{|X||Y|}. \quad (4)$$

2. Ядро оператора Лапласа (laplacian) выводится из матрицы Кирхгофа, математического представления структуры графа. Ядро (k) для векторов X и Y вычисляется как:

$$k(X, Y) = \exp(-\gamma\|X - Y\|_1). \quad (5)$$

3. Линейное ядро (linear) – одна из простейших функций сходства, вычисляемая как скалярное произведение двух векторов X и Y:

$$k(X, Y) = X^{\top}Y. \quad (6)$$

4. Полиномиальное ядро (polynomial) расширяет линейное ядро, вводя нелинейные взаимодействия между признаками, таким образом учитывая сходство между

векторами между измерениями или между разными признаками объектов. Ядро (k) степенью d для векторов X и Y определяется как:

$$k(X, Y) = (\gamma X^\top Y + c_0)^d, \quad (7)$$

где γ – коэффициент скалярного произведения векторов, равный обратному значению длины вектора;

c_0 – смещение, обычно равное единице.

5. Ядро радиальной базисной функции (rbf), также известное как ядро Гаусса, вычисляет сходство на основе расстояния между двумя точками в пространстве признаков. Ядро (k) для векторов X и Y определяется как:

$$k(X, Y) = \exp(-\gamma \|X - Y\|^2), \quad (8)$$

где γ – коэффициент скалярного произведения векторов, равный обратному значению длины вектора.

6. Сигмоидальное ядро (sigmoid), также известное как гиперболический тангенс или многослойный перцептрон, используемый как функция активации в нейронных сетях. Ядро (k) для векторов X и Y определяется как:

$$k(X, Y) = \tanh(\gamma X^\top Y + c_0), \quad (9)$$

где γ – коэффициент скалярного произведения векторов, равный обратному значению длины вектора;

c_0 – смещение, обычно равное единице.

Вычислительная стоимость спектрального вложения зависит от трех этапов. Построение графа требует вычисления парных сходств, которые масштабируются квадратично с числом точек данных. Вычисление матрицы Кирхгофа для разреженных матриц масштабируется линейно с числом соседей. Решение задачи собственных значений масштабируется квадратично с числом точек данных и линейно с целевым измерением. Разреженные представления матрицы Лапласа могут

использоваться для повышения эффективности, что делает спектральное вложение осуществимым для умеренно больших наборов данных.

Спектральное вложение эффективно учитывает нелинейные структуры и сохраняет локальные окрестности. Оно хорошо масштабируется с разреженными матрицами и обеспечивает математически обоснованный подход, основанный на теории графов. Однако оно может быть вычислительно затратным для больших графов. Оно также чувствительно к выбору параметров, таких как количество соседей и ширина ядра, и требует тщательной предварительной обработки, включая нормализацию признаков.

2.1.6. T-распределенное стохастическое вложение соседей

T-распределенное стохастическое вложение соседей (T-distributed Stochastic Neighbor Embedding, t-SNE) – это нелинейный метод снижения размерности, широко используемый для визуализации многомерных данных. Метод сохраняет локальные связи между точками данных, что делает его особенно эффективным для идентификации кластеров и закономерностей в сложных наборах данных.

Основная идея t-SNE заключается в минимизации разницы между распределениями вероятностей, которые представляют парные сходства в пространствах высокой и низкой размерности. Алгоритм моделирует высокоразмерные отношения с использованием гауссовых распределений, а низкоразмерные отношения – с использованием t-распределений Стьюдента, которые лучше справляются с изменениями плотности и эффектами скученности в ограниченных пространствах [128].

Алгоритм t-SNE работает следующим образом:

1. Сначала вычисляются парные сходства в многомерном пространстве. Используется гауссово ядро для определения вероятностей, которые измеряют сходство между точками. Каждая точка имеет свой собственный параметр дисперсии, который корректирует масштабирование на основе локальной плотности. Для заданной точки условная вероятность того, что другая точка является ее соседом, вычисляется на основе расстояния между ними.
2. Затем симметризуются вероятности, усредняя условные вероятности в обоих направлениях, гарантируя, что отношения между парами точек будут согласованными.
3. Вычисляются парные сходства в маломерном пространстве. Эти сходства моделируются с помощью распределения Стьюдента с одной степенью свободы. Это распределение допускает более тяжелые хвосты по сравнению с гауссовыми распределениями, гарантируя, что удаленные точки оказывают меньшее влияние, сохраняя при этом кластеризацию близлежащих точек.
4. Минимизируется расхождение Кульбака-Лейблера, которое измеряет разницу между распределениями вероятностей в многомерном и маломерном пространствах. Функция стоимости минимизируется с помощью градиентного спуска. Градиенты вычисляются для каждой точки на основе сил притяжения от соседей и сил отталкивания от удаленных точек.
5. Итеративно выполняются обновления градиента. На ранних этапах оптимизации фактор преувеличения усиливает силы притяжения для разделения кластеров и формирования более четких структур. На более поздних итерациях оптимизация стабилизируется по мере удаления этого фактора преувеличения.

Для обработки больших наборов данных часто используется приближение Барнса-Хата. Вместо расчета попарных расстояний для всех точек оно организует

данные в октодерево и аппроксимирует расстояния, снижая вычислительную сложность.

Точная реализация t-SNE имеет вычислительную сложность, пропорциональную квадрату числа точек данных, поскольку она требует расчета попарных расстояний. Это ограничивает ее масштабируемость для очень больших наборов данных. Для повышения эффективности приближение Барнса-Хата снижает сложность масштабирования до числа точек данных, умноженного на логарифм числа точек данных.

На эффективность t-SNE, оказывает влияние параметр растерянности, который контролирует баланс между фокусировкой на локальных и глобальных структурах. Он определяет эффективное количество соседей, рассматриваемых для каждой точки, с типичными значениями в диапазоне от пяти до пятидесяти.

T-SNE эффективно учитывает локальные структуры и кластеры, обрабатывает нелинейные отношения и обеспечивает визуально интерпретируемые вложения для сложных наборов данных. Однако метод вычислительно затратен для очень больших наборов данных и чувствителен к настройкам параметров, часто требуя тщательной настройки. Кроме того, он фокусируется на сохранении локальных структур, что может привести к искажениям в глобальных отношениях.

2.1.7. Аппроксимация и проекция однородного многообразия

Аппроксимация и проекция однородного многообразия (Uniform Manifold Approximation and Projection, UMAP) – это нелинейный алгоритм снижения размерности, который сохраняет как локальные, так и глобальные структуры многомерных данных. UMAP использует методы алгебраической топологии и

римановой геометрии для построения низкоразмерных вложений, что делает его мощным инструментом для визуализации, кластеризации и изучения признаков.

UMAP моделирует данные как лежащие на гладком римановом многообразии и ищет низкоразмерное представление, аппроксимируя топологическую структуру данных. Он строит взвешенный граф на основе локальных окрестностей и оптимизирует вложение, которое сохраняет эти отношения. В отличие от t-SNE, который фокусируется на сохранении локальных сходств, UMAP уравновешивает глобальные и локальные структуры.

UMAP аппроксимирует многообразную структуру с помощью нечетких симплициальных множеств, представляющих отношения данных как нечеткую топологическую структуру для кодирования связности. А также оптимизации низкоразмерного вложения, минимизируя расхождение между высокоразмерными и низкоразмерными графами, используя вероятностную структуру, вдохновленную перекрестной энтропией [132].

Алгоритм UMAP состоит из двух основных этапов:

1. Построение графа. Вычисляются парные расстояния между всеми парами точек данных в многомерном пространстве. Для каждой точки данных определяются ее ближайшие соседи, чтобы понять локальную структуру данных. Вычисляется нечеткое симплициальное множество: преобразовывается локальная связность каждой точки в представление нечеткого множества, фиксирующее вероятность связи между точками. Затем строится взвешенный граф k-ближайших соседей, представляющий топологическую структуру многомерных данных, путем объединение нечетких множеств.
2. Оптимизация графа. Инициализируется низкоразмерное вложение: случайно или, например, с использованием анализа главных компонент, размещаются точки данных в целевом маломерном пространстве. Вложение оптимизируется, путем

итеративной корректировки положения точек в низкоразмерном пространстве, чтобы минимизировать расхождение между многомерными и маломерными нечеткими множествами. Эта оптимизация направлена на сохранение топологической структуры данных. Затем используется стохастический градиентный спуск для эффективного схождения к оптимальному вложению, которое сохраняет многообразную структуру данных.

Вычислительная сложность построения графа k-ближайших соседей оценивается как O(NlogN) с использованием приближенных алгоритмов, таких как BallTree. Сложность оптимизации вложения посредством стохастического градиентного спуска масштабируется линейно с размером данных, O(Nd), где d – целевое измерение. Это делает UMAP подходящим для больших наборов данных, обеспечивая несколько более высокую производительность, чем t-SNE.

2.1.8. Нейроподобный метод

Профессор Краснов А. Е. предлагает нейроподобный метод снижения размерности спектральных данных, в частности, инфракрасных спектров. Целью метода является отображение многомерных оптических спектров в трехмерное пространство для улучшенной визуализации и анализа [8; 9; 33]. Основные шаги метода можно описать следующим образом:

1. Мультиплексирование. Каждая компонента спектра мультиплексируется на три канала: X, Y и Z.
2. Режекторная фильтрация. Для первого канала X проводится режекторная фильтрация для удаления неинформативных компонентов. Это делается путём

сравнения вариаций значений каждой компоненты с порогом. Компоненты, превышающие этот порог, сохраняются.

3. Гильберт-фильтрация. Во втором канале Y используется Гильберт-фильтрация для формирования первого вспомогательного сигнала.
4. Вторая режекторная фильтрация. Для третьего канала Z производится аналогичная обработка, чтобы выделить особенности компонентов спектров, связанные с их классификацией.
5. Ортогонализация Грамма-Шмидта. Для второго канала формируется второй дискретный сигнал на основе ортогонализации, что позволяет адаптировать коэффициенты усиления или ослабления компонентов спектра.
6. Формирование трехмерного образа. Наконец, для каждого спектра формируется трехмерный образ, который представляет собой проекцию многомерных данных в трехмерное пространство.

Метод позволяет эффективно уменьшать размерность данных с алгоритмической сложностью порядка N, что значительно быстрее, чем традиционные методы (например, метод главных компонент) [7]. Это делает его особенно полезным для работы с большим объемом спектральной информации, такой как в случае многозондовых фурье-спектрометров.

2.2. Методы предварительной обработки многомерных векторов

Спектральные данные, характеризующиеся высокой размерностью и сложными закономерностями, создают значительные проблемы для алгоритмов снижения размерности. Во многих областях применения, таких как химический анализ, характеристика материалов и биологические исследования, спектральные наборы

данных состоят из векторов с несколькими образцами и сильно коррелированными признаками. Методы предварительной обработки, включая дискретную свёртку, автокорреляцию, дискретную производную и градиентные вычисления, играют решающую роль в подготовке таких данных для снижения размерности. Эти методы подчеркивают значимые закономерности, уменьшают шум и улучшают разделимость признаков, позволяя алгоритмам работать более эффективно [29; 30; 35; 56; 58].

Алгоритмы снижения размерности работают в предположении, что данные содержат скрытые низкоразмерные структуры. Однако необработанные спектральные данные имеют недостатки в виде шума, смещения базовой линии и избыточной информации. Методы предварительной обработки решают эти проблемы, улучшая отношение сигнал/шум, выделяя структурные закономерности и нормализуя масштабы. Эти преобразования создают более подходящее входное пространство для алгоритмов, позволяя им определять значимые взаимосвязи и сохранять существенные связи во время проецирования.

Ниже описаны рассмотренные методы предварительной обработки.

2.2.1. Дискретная свёртка

Дискретная свёртка – это метод обработки сигналов, который применяет фильтр или ядро к данным, улучшая определенные шаблоны и уменьшая шум [40]. Для спектральных данных дискретная свёртка сглаживает колебания и выделяет общие тренды или пики. Сглаживающие фильтры уменьшают высокочастотный шум путем усреднения соседних точек. Это особенно полезно для данных спектроскопии с резкими пиками, вызванными ошибками измерений. Уменьшая шум и подчеркивая

шаблоны, дискретная свёртка может повысить стабильность метрик расстояния и функций ядра, используемых в таких алгоритмах, как PCA и UMAP.

Спектры, подвергнутые дискретной свёртке с собой же, вычисляются по следующей формуле:

$$R_c[k] = \sum_{n=0}^{N-1} A[n] \cdot A[k-n], \qquad k = 0,1,\dots,2N-2, \qquad A[n] = 0 \text{ для } n < 0 \text{ и } n \geq N \quad ,(10)$$

где A – исходный вектор значений спектра вещества;

R_c – обработанный дискретной свёрткой вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

2.2.2. Автокорреляция

Автокорреляция измеряет сходство между сигналом и его задержанной версией, фиксируя повторяющиеся шаблоны и периодические структуры [40]. В спектральных данных автокорреляция идентифицирует циклы, гармоники или повторяющиеся мотивы, связанные с молекулярными колебаниями или резонансами. Подчеркивание периодических шаблонов может упростить кластеризацию путем группировки спектров со схожими частотными характеристиками.

Автокорреляция спектров вычисляется по следующей формуле:

$$R_a[k] = \sum_{n=0}^{N-1-k} A[n] \cdot A[n+k], \qquad k = 0,1,\dots,2N-2, \qquad A[n] = 0 \text{ для } n < 0 \text{ и } n \geq N, \quad (11)$$

где A – исходный вектор значений спектра вещества;

R_a – обработанный автокорреляцией вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

2.2.3. Кумулятивная сумма

Расчет кумулятивной суммы преобразует данные, заменяя каждую точку суммой всех предыдущих точек [112]. Этот метод эффективен для выделения кумулятивных тенденций и сглаживания колебаний. Улучшение постепенных тенденций упрощает идентификацию долгосрочных закономерностей. Снижение шума путем усреднения по точкам данных повышает устойчивость к выбросам. Выделение сдвигов в тенденциях дает представление о переходах и спектральных изменениях. Предварительная обработка кумулятивной суммы повышает стабильность алгоритмов, чувствительных к небольшим изменениям, и особенно полезна для обнаружения глобальных закономерностей.

Кумулятивная сумма для спектров вычисляются следующим образом:

$$R_m[k] = \sum_{n=0}^{k} A[n], k = 0,1, \dots, N-1, \quad (12)$$

где A – исходный вектор значений спектра вещества;

R_m – обработанный кумулятивной суммой вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

2.2.4. Прямая дискретная разница первого порядка

Прямая дискретная разница первого порядка (дискретная производная) вычисляет изменение между последовательными точками данных, эффективно удаляя тенденции и сосредотачиваясь на изменениях [112]. Этот подход особенно полезен для спектральных данных с отклонениями базовой линии или систематическими смещениями. Устранение тенденций позволяет алгоритмам сосредоточиться на отклонениях, которые указывают на базовые закономерности. Дискретная производная преобразует спектральные данные в форму, которая выделяет высокочастотные компоненты, что упрощает обнаружение кластеров, переходов и аномалий.

Прямая дискретная разница первого порядка для спектров вычисляется по формуле:

$$R_d[k] = A[k+1] - A[k], k = 0,1,\dots,N-2, \quad (13)$$

где A – исходный вектор значений спектра вещества;

R_d – обработанный дискретной производной вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

2.2.5. Обратное преобразование

Обратное преобразование заменяет каждое значение данных его обратным значением, подчеркивая меньшие значения и подавляя большие [112]. Нормализация признаков снижает доминирование больших пиков, создавая сбалансированные распределения. Повышение чувствительности к небольшим вариациям улучшает обнаружение признаков. Уменьшение динамического диапазона упрощает метрики расстояния и оптимизирует графовые алгоритмы.

Спектры, для которых были вычислены обратные значения, обработаны следующим образом:

$$R_r[k] = \frac{1}{A[k]}, k = 0,1,\dots,N-1, \quad (14)$$

где A – исходный вектор значений спектра вещества;

R_r – обработанный обратным преобразованием вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

2.2.6. Квадратное преобразование

Квадратное преобразование применяет процесс поэлементного возведения в степень, увеличивая различия и подчеркивая пики [112]. Выделение интенсивности пиков делает доминирующие признаки более заметными. Подавление фонового шума увеличивает контраст между признаками. Улучшение разделимости кластеров путем преувеличения различий повышает чувствительность алгоритма.

Квадраты значений спектра рассчитывались по следующей формуле:

$$R_s[k] = (A[k])^2, k = 0,1, \dots, N-1, \quad (15)$$

где A – исходный вектор значений спектра вещества;

R_s – обработанный квадратным преобразованием вектор, подаваемый на вход алгоритма снижения размерности;

k – индекс элемента вектора;

N – длина вектора.

Предварительная обработка не только повышает качество данных, но и может снижать время вычисления алгоритмов снижения размерности. Алгоритмы снижения размерности часто включают такие операции, как расчеты расстояний, собственные разложения и оптимизацию градиентного спуска, которые плохо масштабируются с увеличением размерности. Предварительная обработка уменьшает избыточные признаки, сжимает информацию и сглаживает переходы, что приводит к более быстрой сходимости и улучшенной масштабируемости.

Помимо улучшения производительности алгоритма, предварительная обработка повышает интерпретируемость. Например, пики, выделенные градиентами,

соответствуют химическим связям или молекулярным колебаниям, что дает представление о сокращенных измерениях.

Методы предварительной обработки, такие как дискретная свёртка, автокорреляция, дискретная производная и градиентные вычисления, необходимы для подготовки спектральных данных к снижению размерности. Эти методы выделяют значимые закономерности, удаляют шум и подчеркивают критические особенности, позволяя алгоритмам эффективно раскрывать базовые структуры. Улучшая стабильность, масштабируемость и интерпретируемость, предварительная обработка преобразует необработанные спектральные данные в форму, которая максимизирует потенциал многообразных алгоритмов обучения. Этот подход не только повышает вычислительную эффективность, но и дает представление о взаимосвязях и свойствах веществ, анализируемых с помощью спектральных измерений.

2.3. Метрики расстояния и сходства

Алгоритмы снижения размерности направлены на упрощение многомерных данных путем проецирования их в пространства с меньшей размерностью, сохраняя при этом значимые связи между данными. Эффективность этих алгоритмов в значительной степени зависит от выбора метрик расстояния, поскольку они определяют, как количественно оценивается сходство или различие между точками. Метрики расстояния влияют не только на структуру встраивания, но и на интерпретируемость и стабильность сокращенного представления [39; 54; 59].

Метрики расстояния определяют геометрию пространств данных, влияя на вычисления определения соседства, построение графа и измерения сходства.

Алгоритмы снижения размерности зависят от этих вычислений для эффективного захвата локальных и глобальных отношений.

Выбор метрики расстояния влияет на:

- Чувствительность к масштабированию данных и шуму.
- Сохранение кластерных структур и глобальных шаблонов.
- Вычислительную эффективность.

Различные метрики подходят для разных типов данных, включая числовые, категориальные и смешанные наборы данных. Понимание характеристик каждой метрики имеет решающее значение для выбора наиболее подходящего алгоритма и предварительной обработки.

Ниже описаны рассмотренные метрики расстояния.

2.3.1. Евклидово расстояние

Евклидово расстояние является краеугольным камнем во многих вычислительных областях, предлагая простой, но мощный способ измерения пространственных отношений. Евклидово расстояние, названное в честь древнегреческого математика Евклида, является мерой расстояния по прямой между двумя точками в пространстве [123]. В двумерной плоскости это длина отрезка линии, соединяющего две точки. В пространствах большей размерности оно обобщается для измерения расстояния между двумя точками в n-мерном евклидовом пространстве. Евклидово расстояние применяется в машинном обучении, распознавании изображений и навигации.

Формула для евклидова расстояния в n-мерном пространстве выводится из теоремы Пифагора. Для двух точек P = (x_1, x_2, …, x_n) и Q = (y_1, y_2, …, y_n) евклидово расстояние вычисляется как:

$$D_{euc}(P,Q) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\,. \tag{16}$$

Преимущества метрики:

Простота понимания и расчета.

Применимость в различных измерениях без концептуальных изменений.

Ограничения:

Чувствительность к масштабу: признаки с большими диапазонами могут доминировать при расчете расстояния.

Увеличение вычислительных затрат с увеличением размерности.

2.3.2. Манхэттенское расстояние

Манхэттенское расстояние, также известное как расстояние «такси» или «городского квартала», – это метрика, используемая для измерения расстояния между двумя точками на клеточной сетке. Термин «манхэттенское расстояние» вдохновлен планировкой улиц в городе, подобном Манхэттену, где пути между двумя точками ограничены маршрутами, похожими на сетку [120].

Для двух точек P = (x_1, x_2, …, x_n) и Q = (y_1, y_2, …, y_n) манхэттенское расстояние в n-мерном пространстве вычисляется как:

$$D_{man}(P,Q) = \sum_{i=1}^{n} |x_i - y_i|\,. \tag{17}$$

Эта метрика особенно полезна, когда движения ограничены перпендикулярными направлениями, как это часто бывает в городском планировании, сетевой маршрутизации или задачах дискретной оптимизации.

Преимущества метрики:

Простота вычисления и интерпретации.

Идеально подходит для систем на основе сетки или там, где движение ограничено ортогональными траекториями.

Хорошо работает с категориальными данными при использовании в определенных моделях машинного обучения.

Ограничения:

Менее подходит для контекстов, требующих диагонального или свободного движения.

Может давать вводящие в заблуждение результаты в многомерных пространствах из-за «проклятия размерности».

2.3.3. Расстояние Хэмминга

Расстояние Хэмминга – это фундаментальное понятие в информатике, математике и теории информации. Расстояние Хэмминга между двумя строками одинаковой длины – это количество позиций, в которых соответствующие символы или биты различны. Эта метрика названа в честь Ричарда Хэмминга, американского математика и ученого-компьютерщика, который ввел ее в контексте кодов обнаружения и исправления ошибок [81]. Расстояние Хэмминга особенно полезно для двоичных данных и символических последовательностей, таких как последовательности ДНК, сетевые адреса и коды исправления ошибок.

Между двумя последовательностями S_1 и S_2 одинаковой длины n расстояние считается как количество не совпавших элементов при одинаковом индексе i:

$$D_{hamm}(S_1, S_2) = \sum_{i=1}^{n} \left(S_{1[i]} \neq S_{2[i]}\right). \quad (18)$$

Преимущества метрики:

Простота вычислений. Эффективно работает с бинарными и категориальными данными.

Хорошо подходит для последовательностей данных фиксированной длины.

Ограничения:

Требует, чтобы строки или последовательности были одинаковой длины.

Ограниченная применимость для числовых данных или данных с непрерывными атрибутами.

2.3.4. Расстояние Брея-Кёртиса

Расстояние Брея-Кёртиса – это метрика несходства, используемая для количественной оценки разницы между двумя неотрицательными распределениями данных. Она широко применяется в экологии, биологии и анализе данных для сравнения композиционных данных, таких как численность видов или распределение популяций. Эта метрика учитывает пропорциональные различия между двумя наборами данных, что делает ее особенно полезной для сравнения относительной численности, а не абсолютных значений.

В отличие от евклидовых или манхэттенских расстояний, которые фокусируются на геометрических или сетчатых пространствах, расстояние Брея-Кёртиса основано

на пропорциях, что делает его идеальным для композиционных данных, где общее количество имеет меньшее значение, чем относительные вклады [84].

Для двух векторов $X = (x_1, x_2, \ldots, x_n)$ и $Y = (y_1, y_2, \ldots, y_n)$ расстояние Брея-Кёртиса в n-мерном пространстве вычисляется как:

$$D_{bc}(X, Y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} |x_i + y_i|}. \quad (19)$$

Преимущества метрики:

Подходит для композиционных данных, где относительные вклады важнее абсолютных значений.

Нормализует различия в общей величине между наборами данных.

Ограничения:

Не учитывает геометрические отношения между точками данных, в отличие от евклидова расстояния.

2.3.5. Расстояние Канберры

Расстояние Канберры – это метрика несходства, используемая для измерения расстояния между двумя числовыми векторами. В отличие от евклидовых или манхэттенских расстояний, расстояние Канберры делает больший акцент на небольших различиях относительно величины элементов. Это делает его особенно полезным при анализе наборов данных с различными масштабами или когда относительное изменение важнее абсолютных различий.

Расстояние Канберры – это взвешенная версия расстояния Манхэттена, которая вычисляет сумму относительных различий между элементами двух векторов. Оно особенно чувствительно к небольшим изменениям, когда значения малы, так как

относительные различия больше. И наоборот, оно ослабляет различия, когда значения велики. Эта метрика неотрицательна и хорошо подходит для разреженных данных [121].

Для двух векторов X = (x_1, x_2, …, x_n) и Y = (y_1, y_2, …, y_n) расстояние Канберры в n-мерном пространстве вычисляется как:

$$D_{cb}(X,Y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}. \quad (20)$$

Преимущества метрики:

Эффективно обрабатывает наборы данных с различными масштабами, фокусируясь на относительных различиях.

Чувствительна к небольшим изменениям в малых значениях, что делает её пригодным для разреженных или составных данных.

Ограничения:

Может непропорционально сильно зависеть от малых значений в знаменателе, что приводит к большим расстояниям для незначительных различий.

2.3.6. Расстояние Чебышёва

Расстояние Чебышёва, также известное как метрика шахматной доски или метрика L_∞, является мерой наибольшей разницы между двумя векторами вдоль любого одного измерения. В отличие от других метрик расстояния, таких как евклидово или манхэттенское расстояние, которые объединяют различия по измерениям, расстояние Чебышёва фокусируется на наибольшем отклонении [105].

Для двух векторов X = (x_1, x_2, …, x_n) и Y = (y_1, y_2, …, y_n) расстояние Чебышёва в n-мерном пространстве вычисляется как это максимальная абсолютная разница между их соответствующими координатами:

$$D_{cheb}(P,Q) = \max_i |x_i - y_i| . \quad (21)$$

В двумерном пространстве эта метрика соответствует количеству ходов, необходимых королю в шахматах для перемещения с одной клетки на другую, поскольку король может сделать один шаг в любом направлении (по горизонтали, вертикали или диагонали).

Преимущества метрики:

Простота вычисления и интерпретации.

Подходит для многомерных пространств, где совокупные показатели могут ослаблять важные выбросы.

Ограничения:

Не учитывает кумулятивные различия по измерениям, что делает его менее подходящим для некоторых приложений машинного обучения.

Может чрезмерно упрощать измерение расстояния в системах, где все отклонения вносят одинаковый вклад.

2.3.7. Корреляционное расстояние

Корреляционное расстояние – это метрика, которая измеряет несходство между двумя векторами данных на основе их корреляции. Она количественно определяет, насколько сильно связаны два вектора, исследуя степень их совместного изменения. В отличие от традиционных метрик расстояния, таких как евклидово или манхэттенское

расстояние, расстояние корреляции фокусируется на связи между векторами, а не на их абсолютных значениях.

Расстояние корреляции выводится из коэффициента корреляции Пирсона, меры линейной связи между двумя векторами [137]. Если векторы идеально положительно коррелируют (r=1), корреляционное расстояние равно 0, что указывает на отсутствие несходства. И наоборот, если векторы идеально отрицательно коррелируют (r=−1), расстояние равно 2, что указывает на максимальное несходство.

Эта метрика широко используется в таких областях, как наука о данных, машинное обучение и биоинформатика, где взаимосвязи между переменными имеют большее значение, чем их абсолютные различия.

Для двух векторов X = (x_1, x_2, …, x_n) и Y = (y_1, y_2, …, y_n) корреляционное расстояние в n-мерном пространстве вычисляется как это максимальная абсолютная разница между их соответствующими координатами:

$$D_{corr}(X, Y) = 1 - r(X, Y), \tag{22}$$

где r – коэффициент корреляции Пирсона, определяемый как:

$$r(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{23}$$

где $\bar{x}$ и $\bar{y}$ – средние значения X и Y: $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$, $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$.

Преимущества метрики:

Фокусируется на взаимосвязи между переменными, а не на величине.

Хорошо работает для стандартизированных или нормализованных данных.

Устойчива к равномерному масштабированию или переводу значений данных.

Ограничения:

Предполагает линейные взаимосвязи; может не охватывать нелинейные зависимости.

Требует ненулевой дисперсии в обоих векторах, так как в противном случае корреляция не определена.

Чувствительна к выбросам, которые могут исказить коэффициент корреляции.

Относительно сложна в вычислении.

2.3.8. Косинусное расстояние

Косинусное расстояние – это мера несходства двух ненулевых векторов в многомерном пространстве. Оно количественно определяет, насколько различаются направления двух векторов, независимо от их величин. Эта метрика широко используется в текстовом анализе, машинном обучении и системах рекомендаций для сравнения ориентации векторов.

Косинусное расстояние выводится из косинусного сходства, которое измеряет косинус угла между двумя векторами. В то время как косинусное сходство оценивает близость двух векторов с точки зрения направления, косинусное расстояние фокусируется на их расхождении. Оно варьируется от 0 до 2: Косинусное расстояние 0 указывает на идентичные направления (максимальное сходство), 1 предполагает ортогональные векторы (отсутствие сходства), а расстояние равное 2, указывает на векторы, направленные в противоположные стороны (максимальное различие) [111].

Для двух векторов $X = (x_1, x_2, \ldots, x_n)$ и $Y = (y_1, y_2, \ldots, y_n)$ косинусное расстояние в n-мерном пространстве вычисляется как:

$$D_{cos}(X,Y) = 1 - S(X,Y), \tag{24}$$

где S – косинусное сходство, определяемое как:

$$S(X,Y) = \frac{X \cdot Y}{\|X\|_2 \|Y\|_2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{25}$$

Преимущества метрики:

Устойчива к различиям в величине, фокусируется на сходстве направленности.

Хорошо работает с данными высокой размерности.

Интуитивная интерпретация, особенно для разреженных или нормализованных данных.

Ограничения:

Требует ненулевых векторов; не определена, если какой-либо вектор имеет нулевую величину.

Может не улавливать различия на основе величины, если они важны.

Чувствительна к небольшим изменениям направления для почти параллельных векторов.

2.3.9. Стандартизированное евклидово расстояние

Стандартизированное евклидово расстояние – это разновидность метрики евклидова расстояния, которая учитывает дисперсию каждого измерения в наборе данных [32]. Стандартизируя измерения, эта метрика гарантирует, что признаки с большими дисперсиями не будут непропорционально влиять на расчет расстояния. Это особенно полезно при анализе данных, когда признаки имеют разные масштабы или единицы. Такое масштабирование гарантирует, что все измерения вносят одинаковый вклад в расчет расстояния, независимо от их исходных масштабов или единиц.

Для двух векторов X = (x_1, x_2, …, x_n) и Y = (y_1, y_2, …, y_n) стандартизированное евклидово расстояние в n-мерном пространстве вычисляется как:

$$D_{seuc}(X,Y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i}}, \quad (26)$$

где σ – дисперсия i-ых элементов.

Преимущества метрики:

Корректирует смещение, вносимое признаками с большими дисперсиями.

Подходит для наборов данных со смешанными шкалами или единицами.

Повышает надежность алгоритмов на основе расстояний, таких как кластеризация.

Ограничения:

Требует расчета дисперсий для каждого измерения.

Чувствительна к выбросам, поскольку дисперсии могут быть искажены экстремальными значениями.

Вычислительно более сложна, чем простое евклидово расстояние.

2.3.10. Квадратное евклидово расстояние

Квадратное евклидово расстояния – это разновидность метрики евклидового расстояния, в которой квадраты разностей между соответствующими измерениями суммируются без извлечения квадратного корня [32]. Эта метрика вычислительно проще и часто используется в приложениях, где не требуются точные расстояния, но важны относительные расстояния, например, в алгоритмах кластеризации.

Для двух точек P = (x_1, x_2, …, x_n) и Q = (y_1, y_2, …, y_n) квадратное евклидово расстояние вычисляется как:

$$D_{sqeuc}(P, Q) = \sum_{i=1}^{n} (x_i - y_i)^2 . \quad (27)$$

Преимущества метрики:

Быстрее вычисляется, чем традиционное евклидово расстояние, из-за отсутствия операции квадратного корня.

Подходит для относительных сравнений, сохраняя тот же порядок расстояний, что и евклидово расстояние.

Идеально подходит для алгоритмов, требующих повторяющихся вычислений расстояний, таких как кластеризация.

Ограничения:

Результаты нельзя напрямую интерпретировать как расстояния из-за отсутствия квадратного корня.

Квадраты разностей усиливают влияние выбросов, потенциально искажая результаты.

Метрики расстояний формируют основу алгоритмов снижения размерности, формируя вложения и влияя на эффективность. Каждая метрика имеет свои сильные стороны и ограничения, что делает необходимым согласование выбора метрик с характеристиками данных и алгоритмическими требованиями. Поскольку сложность данных продолжает расти, дальнейшие исследования адаптивных и гибридных метрик обещают улучшить масштабируемость, надежность и точность в анализе многомерных данных.

Метрики расстояния влияют на алгоритмы уменьшения размерности следующим образом:

- Такие алгоритмы, как Isomap и Spectral Embedding, используют метрики расстояния для формирования графов соседства.
- Метрики влияют на веса, назначенные ребрам графа, влияя на вычисления собственных значений и оптимизацию вложения.
- Такие алгоритмы, как t-SNE и UMAP, напрямую зависят от определений соседства, где метрики определяют вероятности и сходства.
- Сложность метрики влияет на вычислительные затраты, ограничивая масштабируемость для больших наборов данных.

2.4. Методы кластеризации

Задача кластеризации точек в низкоразмерном пространстве может быть сформулирована следующим образом.

Дано:

- Множество спектров $Z = \{z_i\}_{i=1}^{N}$, где $z_i \in R^d$ – представления спектров в низкоразмерном пространстве после отображения T (формула 1).
- Количество кластеров K, которое может быть задано или определяться алгоритмом автоматически.

Задача:

Построить разбиение множества Z на подмножества S

$$S = \{S_1, S_2, \dots, S_k\}, \qquad \bigcup_{k=1}^{K} S_k = Z, \qquad S_i \cap S_j = \emptyset \ (i \neq j), \tag{28}$$

такое, что внутри каждого кластера объекты схожи, а между кластерами – различны.

Критерии:

Для методов на основе центроидов критерием является минимизация внутрикластерной дисперсии.

$$\min_{S, \mu_1, \dots, \mu_K} J(S, \mu_1, \dots, \mu_K), \tag{29}$$

где μ_K – центр k-го кластера.

Для плотностных методов критерием будет являться максимизация числа точек, объединенных в плотностные компоненты при фиксированных радиусах окрестности и минимальном числе точек для формирования кластера и выделение точек, которые не принадлежат ни одному кластеру, как шумовых.

Ограничения:

- Каждая точка принадлежит либо одному кластеру, либо определяется как выброс.
- Для устойчивости требуется невысокая чувствительность результата к шуму и выбросам.

2.4.1. K-средних

Среди множества доступных алгоритмов кластеризации кластеризация K-средних (K-means) выделяется как один из самых популярных и широко используемых методов благодаря своей простоте, эффективности и универсальности [149].

K-means – это итеративный алгоритм кластеризации, разработанный для разбиения набора данных на отдельные кластеры, где количество кластеров – указанный пользователем параметр. Целью алгоритма является минимизация внутрикластерной дисперсии или инерции, измеряемой как сумма квадратов расстояний между точками данных (x_i) и их соответствующими центроидами кластера (μ_j):

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left|\left| x_i - \mu_j \right|\right|^2 \right). \quad (30)$$

Алгоритм состоит из следующих шагов:

1. Инициализация: выбор начальных центроидов случайным образом или с использованием эвристического метода, например K-means++.
2. Шаг назначения: назначение каждой точки данных ближайшему центроиду на основе метрики расстояния.
3. Шаг обновления: пересчет центроидов как среднего значения всех точек данных, назначенных каждому кластеру.
4. Итерация: повторение шагов назначения и обновления до сходимости, определяемой минимальными изменениями в центроидах или назначениях.

Несмотря на свою простоту, K-means имеют ряд преимуществ, включая простоту реализации, масштабируемость и вычислительную эффективность. Однако он также имеет ограничения, такие как чувствительность к начальным центроидам, тенденция

к сходимости к локальным минимумам и предположение о сферических кластерах одинакового размера.

2.4.2. Сдвиг среднего значения

Сдвиг среднего значения (MeanShift) – это алгоритм кластеризации на основе центроида, который итеративно сдвигает точки данных в сторону режима функции плотности. В отличие от алгоритмов типа K-means, MeanShift не предполагает предопределенного количества кластеров. Вместо этого он идентифицирует кластеры, находя пики в функции плотности вероятности данных [93].

Алгоритм работы MeanShift:

1. Оценка плотности ядра. Алгоритм начинается с оценки плотности данных с использованием функции ядра, обычно гауссовского ядра. Функция плотности вычисляется вокруг каждой точки данных в пределах указанного параметра допуска.
2. Вектор среднего сдвига. Для каждой точки данных алгоритм вычисляет вектор среднего сдвига (mean shift), который указывает в сторону области максимальной плотности. Математически вектор сдвига задается следующим образом:

$$m(x) = \frac{\sum_{x_i \in N(x)} x_i K(x - x_i)}{\sum_{x_i \in N(x)} K(x - x_i)}, \tag{31}$$

где K – функция ядра;

x – центроид;

$N(x)$ – окрестность вокруг x, определяемая параметром допуска.

3. Сходимость. Алгоритм итеративно обновляет положение каждой точки данных с использованием вектора среднего сдвига до сходимости, где точки данных группируются вокруг пиков плотности.
4. Формирование кластера. Точки, сходящиеся к одному и тому же пику плотности, группируются в один кластер.

Преимущества MeanShift:

MeanShift не требует предварительного указания количества кластеров.

Алгоритм хорошо работает в сценариях, где кластеры имеют несферическую или сложную форму.

Устойчивость к шуму. Сосредоточившись на пиках плотности, MeanShift может эффективно игнорировать изолированные точки данных (выбросы).

Ограничения MeanShift:

Сложность вычислений. MeanShift требует больших вычислительных затрат, особенно для больших наборов данных, из-за повторных оценок плотности ядра и поиска ближайшего соседа.

Хранение и обработка плотных наборов данных может привести к высокому использованию памяти.

Чувствительность к параметру допуска. Параметр допуска существенно влияет на результаты. Небольшой допуск способность может привести к переобучению (слишком большому числу кластеров), тогда как большое значение может объединить отдельные кластеры.

2.4.3. Основанная на плотности пространственная кластеризация приложений с шумами

Пространственная кластеризация приложений с шумами на основе плотности DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – это мощный алгоритм кластеризации, который отлично справляется с определением кластеров различной формы и обработкой шума в наборах данных. DBSCAN – это алгоритм кластеризации на основе плотности, который определяет кластеры как плотные области в данных, разделенные областями с меньшей плотностью. В отличие от методов на основе центроидов, DBSCAN не предполагает фиксированного количества кластеров или определенных форм, что делает его идеальным для наборов данных со сложной структурой [100].

Основные параметры DBSCAN: ε – максимальное расстояние между двумя точками данных, чтобы они считались соседями и min_samples – минимальное количество точек, необходимое для формирования плотной области (кластера).

Точки данных с не менее чем min_samples соседями на расстоянии ε называются основными точками.

Точки, которые находятся на границе кластера, но сами не соответствуют критерию количества соседей, называются граничными точками.

Точки, которые не являются ни основными, ни граничными точками и лежат в областях с низкой плотностью, называются шумовыми.

Алгоритм работы DBSCAN:

1. Для каждой точки данных алгоритм определяет ее ε-окрестность (все точки в пределах расстояния ε).

2. Точки классифицируются как основные, граничные или шумовые на основе ε-окрестности и минимального количества соседей.
3. Начиная с не посещённой основной точки, DBSCAN итеративно расширяет кластер, включая все достижимые по плотности точки.
4. Процесс продолжается до тех пор, пока не будут посещены все точки, что сформирует отдельные кластеры и шумовые точки.

Преимущества DBSCAN:

В отличие от алгоритмов, предполагающих сферические кластеры, DBSCAN может идентифицировать кластеры произвольных форм и размеров.

DBSCAN явно идентифицирует выбросы как шумовые точки, улучшая качество кластеризации.

Нет необходимости указывать количество кластеров, количество кластеров определяется динамически на основе распределения данных.

Алгоритм эффективен для наборов данных с различной плотностью.

Ограничения алгоритма:

Чувствительность к параметрам. Выбор параметров существенно влияет на результаты кластеризации. Неправильный выбор параметров может привести к избыточной или недостаточной кластеризации.

Для очень больших наборов данных DBSCAN может быть вычислительно интенсивным из-за поиска соседей для каждой точки.

DBSCAN испытывает трудности с наборами данных, содержащими кластеры с сильно различающейся плотностью.

2.4.4. Упорядочение точек для обнаружения кластерной структуры

Упорядочение точек для обнаружения кластерной структуры – OPTICS (Ordering Points To Identify the Clustering Structure) – это алгоритм кластеризации на основе плотности, который обобщает DBSCAN для идентификации кластеров с различной плотностью. OPTICS – это алгоритм кластеризации на основе плотности, разработанный для идентификации кластеров произвольной формы и плотности. В отличие от DBSCAN, который опирается на фиксированные параметры плотности, OPTICS генерирует график достижимости, который фиксирует структуру кластеризации в диапазоне пороговых значений плотности [67]. Такой подход делает OPTICS идеальным для наборов данных с различной плотностью кластеров.

Ключевыми параметрами алгоритма является ε – максимальное расстояние для двух точек, которые будут считаться соседями. В OPTICS это служит верхней границей, а не фиксированным параметром. А также min_samples – минимальное количество точек, необходимых для формирования кластера.

Минимальное расстояние, необходимое для достижения точки из другой основной точки называется расстоянием достижимости.

Алгоритм работы OPTICS:

1. Вычисление окрестностей. Для каждой точки данных вычисляется ее ε-окрестность, чтобы определить, квалифицируется ли она как основная точка на основе min_samples.
2. Упорядочение точек. Проход по набору данных, начиная с произвольных точек, чтобы итеративно расширить плотные регионы. Точки упорядочиваются по увеличению расстояния достижимости, отдавая приоритет обходу через области с высокой плотностью.

3. Построение графика достижимости. Генерация графика достижимости, кодирующего ландшафт плотности данных. Плотные кластеры проявляются как «долины», в то время как разреженные регионы и шум отображаются как «горы». Чем ниже «долина», тем плотнее находятся точки
4. Извлечение кластеров. Применение ручных или автоматизированных методов пороговой обработки для сегментации кластеров из графика достижимости.

На рисунке 1 показан пример графика достижимости и результатов кластеризации исследуемых данных методами OPTICS и DBSCAN с разными параметрами допущения.

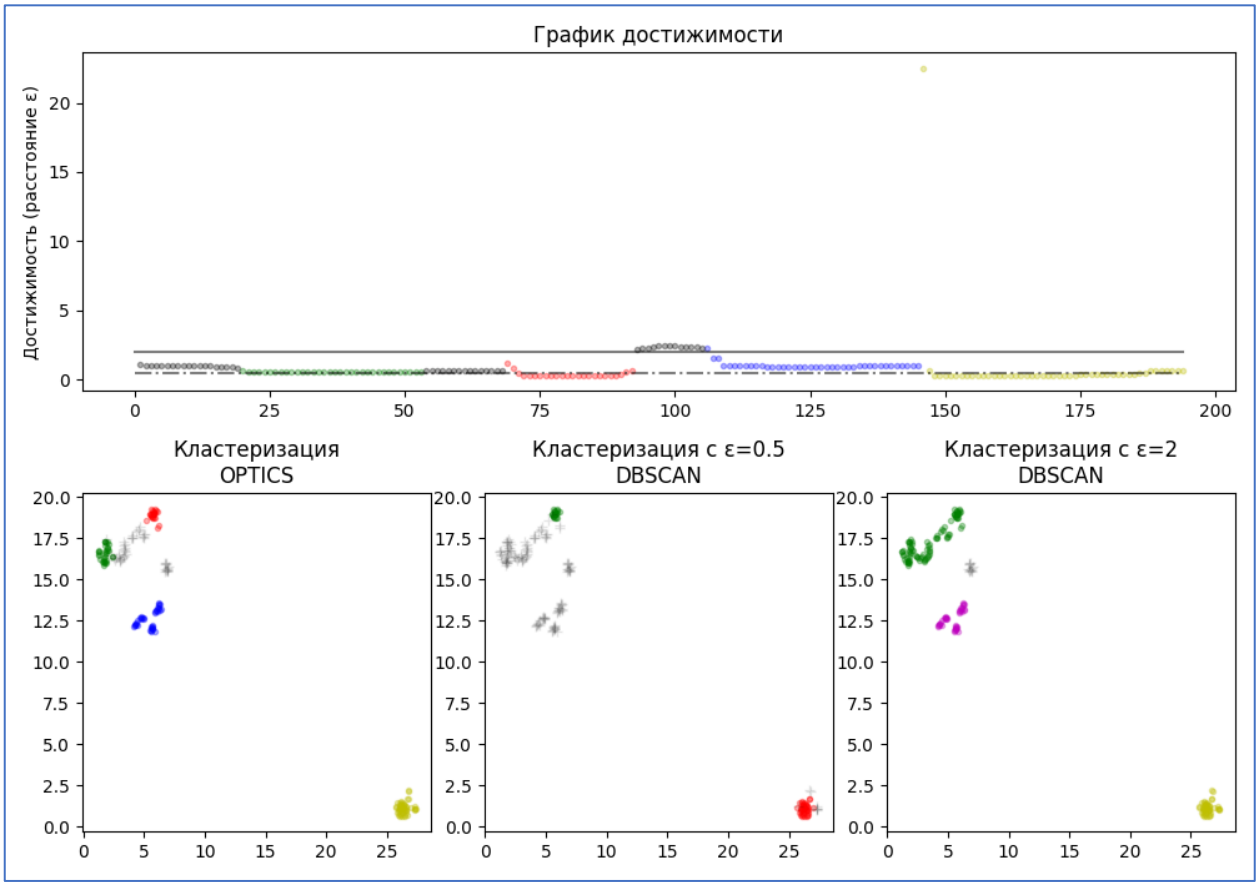


Рисунок 1 – График достижимости и демонстрация работы алгоритмов OPTICS и DBSCAN

Преимущества OPTICS:

Адаптивность к различным плотностям. В отличие от DBSCAN, OPTICS может легко идентифицировать кластеры с различными плотностями, учитывая сложные структуры данных.

Идентификация произвольных форм кластеров. Алгоритм не ограничен сферической или выпуклой геометрией кластера, что делает его универсальным в различных областях.

Устойчивость к шуму. OPTICS отфильтровывает шумовые точки, распознавая их как выбросы в областях с низкой плотностью.

Единичное выполнение для комплексного анализа. График достижимости устраняет необходимость в нескольких запусках алгоритма при разных значениях, объединяя усилия.

Ограничения OPTICS:

Вычислительные издержки. Парное вычисление расстояний достижимости требует больших временных и временных затрат, особенно для больших наборов данных.

Чувствительность параметров. Хотя алгоритм и более гибкий, чем DBSCAN, выбор ε и min_samples все равно влияет на результаты кластеризации и вычислительную эффективность.

Проблемы интерпретации кластера. Извлечение значимых кластеров из графика достижимости требует экспертных знаний в данной области и тщательного анализа.

2.4.5. Иерархическая пространственная кластеризация приложений с шумами на основе плотности

Иерархическая пространственная кластеризация приложений с шумами на основе плотности, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) развивает сильные стороны DBSCAN, устраняя его ограничения и расширяя его возможности для наборов данных с кластерами различной плотности. HDBSCAN – это расширение DBSCAN, разработанное для обработки наборов данных с кластерами различной плотности. Он представляет иерархический подход к кластеризации, создавая дерево возможных кластеров на основе плотности и позволяя извлекать наиболее значимые кластеры без необходимости определения параметра ε [87].

Расстоянием до ближайшего соседа, которое удовлетворяет критерию min_samples, называется основное расстояние. Это определяет плотность локального соседства.

Расстояние взаимной достижимости объединяет основные расстояния двух точек с прямым расстоянием между ними, гарантируя, что более плотные регионы будут приоритетными при кластеризации.

Алгоритм работы HDBSCAN:

1. Построение графика взаимной достижимости. Вычисление расстояния взаимной достижимости между всеми точками и представление набора данных в виде графика.
2. Построение минимального остовного дерева. Получение дерева из графика взаимной достижимости, чтобы зафиксировать существенную связность между точками.

3. Создание иерархии кластеров. Постепенно обрезать ребра в остовном дереве, чтобы сформировать иерархическое дерево кластеризации.
4. Генерация сжатого дерева. Извлечение значимых кластеров путем обрезки иерархического дерева на основе стабильности, удаляя переходные или шумные кластеры.
5. Извлечение кластера. Определение кластеров с различной плотностью без необходимости глобального порога плотности.

Преимущества алгоритма:

Обрабатывает кластеры с различной плотностью, преодолевая ограничение DBSCAN по определению параметра ε.

Эффективно определяет и исключает шумовые точки, улучшая качество кластеров.

Иерархические данные. Алгоритм создает иерархию кластеров, что позволяет глубже понять структуру данных.

Ограничения алгоритма:

Сложность вычислений. Построение графика взаимной достижимости и иерархического дерева может быть ресурсоемким для больших наборов данных.

Хотя алгоритм и менее чувствителен к параметрам, чем DBSCAN, на результаты все равно может влиять выбор параметра min_samples.

Анализ и извлечение значимых кластеров из иерархического дерева требуют знания предметной области и тщательной интерпретации.

2.4.6. Спектральная кластеризация

Спектральная кластеризация (Spectral Clustering) – это основанный на графах алгоритм, который использует спектр (собственные значения и собственные векторы) матрицы подобия, полученной из данных, для выполнения снижения размерности перед кластеризацией [94]. Этот подход особенно эффективен для данных, которые не очень хорошо подходят для традиционных методов, таких как K-means, из-за невыпуклых или сложных форм кластеров.

Алгоритм работы спектральной кластеризации:

1. Построение графа подобия, где узлы представляют точки данных, а веса ребер представляют парные сходства, часто вычисляемые с использованием ядер Гаусса или методов ближайшего соседа.
2. Вычисление матрицы Лапласа, полученной из матрицы подобия, кодирует структуру данных. Она может быть ненормализованной или нормализованной в зависимости от задачи кластеризации.
3. Выполнение собственной декомпозиции матрицы Лапласа и извлечение собственных векторов, соответствующих наименьшим собственным значениям.
4. Собственные векторы используются для вложения данных в пространство с меньшей размерностью, где кластеры более различимы.
5. Применение стандартного алгоритма кластеризации, такого как K-means, в этом преобразованном пространстве.

Преимущества алгоритма:

Обработка сложных форм кластеров. Спектральная кластеризация отлично справляется с обнаружением невыпуклых и произвольно сформированных кластеров, которые другие алгоритмы с трудом идентифицируют.

Гибкие меры сходства. Метод позволяет настраивать метрики сходства, что делает его адаптируемым к различным типам данных, включая изображения, графики и текст.

Снижение размерности. Вкладывая данные в пространство с меньшей размерностью, спектральная кластеризация упрощает задачу кластеризации, сохраняя при этом важную структурную информацию.

Ограничения алгоритма:

Вычислительная сложность. Разложение собственных значений требует больших вычислительных затрат, особенно для больших наборов данных.

Масштабируемость. Алгоритм ограничен требованиями к памяти для построения и хранения матриц сходства и Лапласа.

Чувствительность параметров. Результаты могут различаться в зависимости от выбора меры сходства, количества кластеров и выбранных собственных векторов.

2.5. Выводы по второй главе

По результатам анализа, выполненного во второй главе, можно сделать следующие выводы. Были систематически рассмотрены линейные и нелинейные (многообразное обучение) методы снижения размерности. Детальный обзор показал, что выбор метода должен опираться на природу данных: для близких к линейной структуре PCA обеспечивает простоту и интерпретируемость; t-SNE и UMAP демонстрируют высокую визуальную информативность при сохранении локальной структуры, но требуют внимательной настройки параметров. Нейросетевые решения подтверждают свою перспективность с точки зрения вычислительной эффективности и масштабируемости при корректной архитектурной настройке.

В разделе, посвящённом предварительной обработке, показано, что операции дискретной свёртки, автокорреляции и дискретной производной существенно улучшают соотношение сигнал/шум и выделяют информативные компоненты спектра. Эти преобразования также повышают стабильность метрик расстояния и повышают воспроизводимость и устойчивость снижения размерности к выбросам и матричным эффектам.

Выбор метрик расстояния показан как фундаментальный фактор, определяющий геометрию данных и корректность применения алгоритмов снижения размерности и кластеризации. Подчёркивается, что разные метрики (евклидова, манхэттенская, корреляционная, Канберры, косинусная и др.) по-разному отражают сходство объектов, и выбор метрики должен быть согласован с природой данных.

Методы кластеризации описаны как завершающий этап анализа, направленный на выявление однородных групп спектров. Методы на основе центроидов (K-means) просты и эффективны при сферической структуре кластеров; плотностные алгоритмы (DBSCAN, OPTICS, HDBSCAN) лучше выявляют сложные и разрежённые структуры и отделяют шум; спектральная кластеризация сочетает преимущества графовых представлений, но чувствительна к объёму данных и выбору меры сходства. Каждый метод требует соизмерения с задачей: точность разбиения, устойчивость к выбросам и интерпретируемость результатов оказываются взаимоисключающими в части требований и потому нуждаются в компромиссных решениях.

В практическом аспекте основной вывод сводится к следующему: системная интеграция предварительной обработки и модифицированного алгоритма снижения размерности является критически важным условием для эффективной кластеризации и интерпретации спектральных данных, а выбранный алгоритм обеспечивает баланс между точностью, устойчивостью и интерпретируемостью получаемых представлений.

Глава 3. Анализ методов исследования

Исследование влияния выбора различных подходов на каждом шаге алгоритма на качество классификации было проведено на спектральных данных четырех видов бензинов, по 50 образцов в каждом, полученных на образцах Татнефть с помощью ИК Фурье-спектрометра АФ-3 [31].

3.1. Анализ предварительной обработки

В химической аналитике, в частности в анализе спектральных данных, ключевую роль играют предварительная обработка данных, снижение размерности и кластеризация. Эти этапы обеспечивают выделение информативных признаков и выявление структур в данных. Далее будут рассмотрены результаты различных методов предварительной обработки спектральных данных веществ. Применяются такие подходы, как дискретная свёртка, автокорреляция, дискретная производная, обратные величины и возведение в квадрат. Каждый из этих методов позволяет по-разному подчеркнуть особенности спектров и подготовить их к дальнейшему анализу.

На рисунке 2 представлен график исходных значений отклика спектров.

Визуально даже при таком масштабе видно, что явно выделяется спектр АИ-80 (бирюзовый) высокими пиками значений около 400 $см^{-1}$ и 1400 $см^{-1}$.

Образец АИ-92 (синий) также отличается схожими, но несколько меньшими пиками около 400 $см^{-1}$ и относительно более выраженным откликом около 2900 $см^{-1}$.

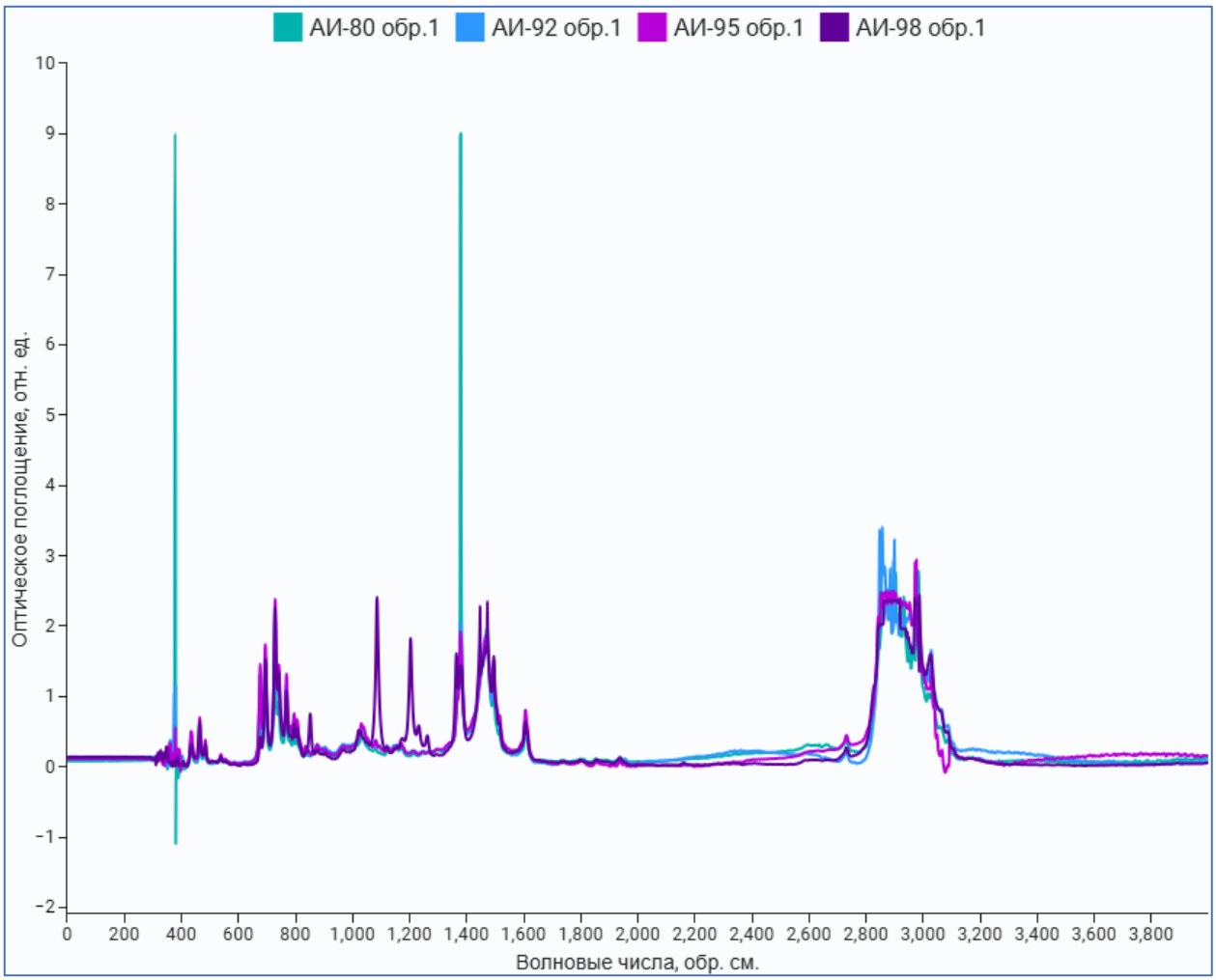


Рисунок 2 – График отклика спектров в исходном виде

Также образец АИ-98 (темно-фиолетовый) отличается откликами в нескольких областях от 800 $см^{-1}$ до 1300 $см^{-1}$, что хорошо видно на рисунке 3. В остальных фрагментах спектры слабо различимы.

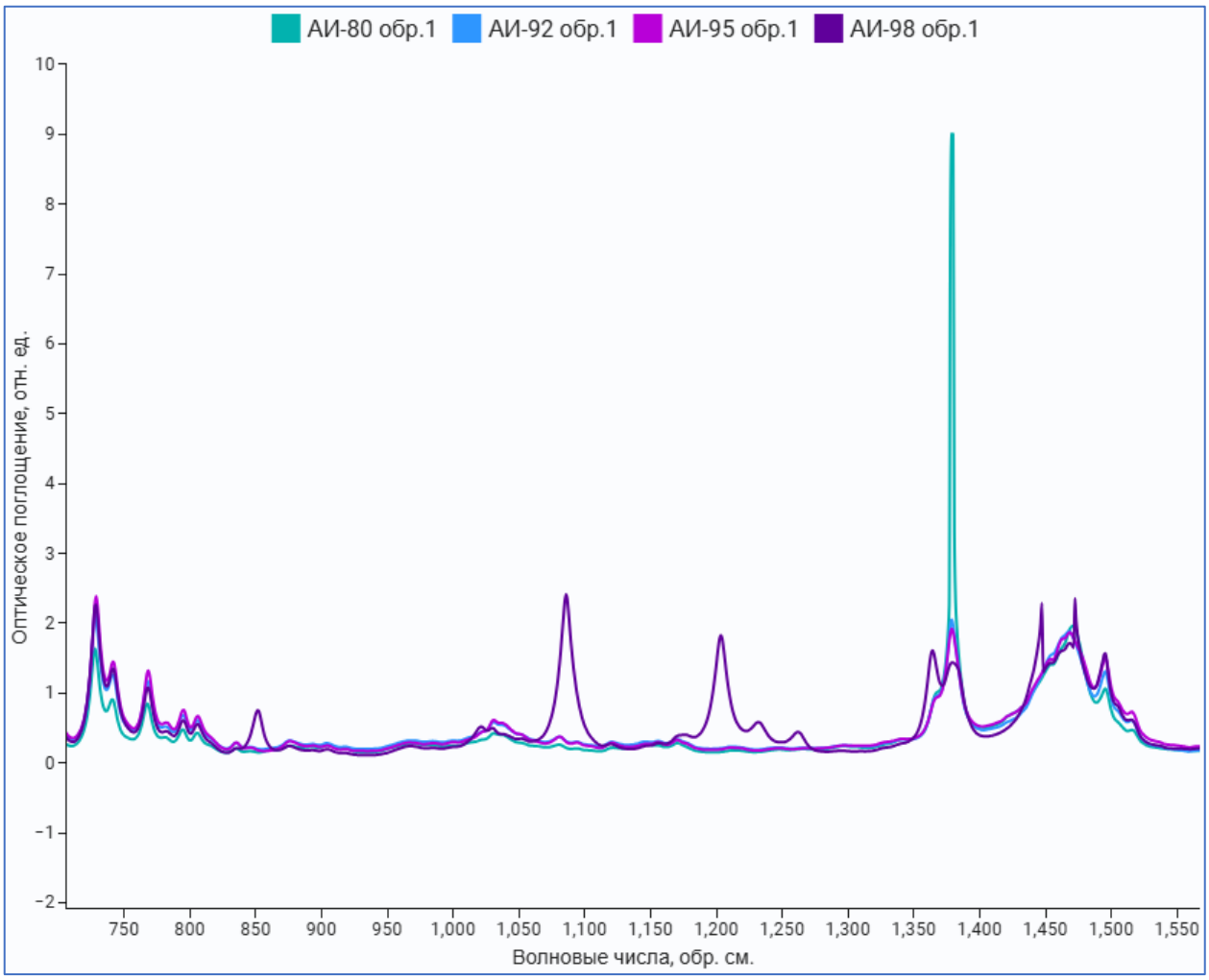


Рисунок 3 – Фрагмент графика отклика спектров в исходном виде

График нескольких полученных значений отклика бензинов после применения дискретной свёртки представлен на рисунке 4.

Визуально на графике можно выделить, что образец АИ-95 (фиолетовый), в отличие от остальных, на отрезке от 6200 $см^{-1}$ до 6800 $см^{-1}$ совершает более явный рост значений, когда остальные образцы совершают лишь небольшой рост в конце.

Образец АИ-98 (фиолетовый) показывает большее количество мелких пиков и колебаний в диапазоне от 1800 $см^{-1}$ до 2800 $см^{-1}$.

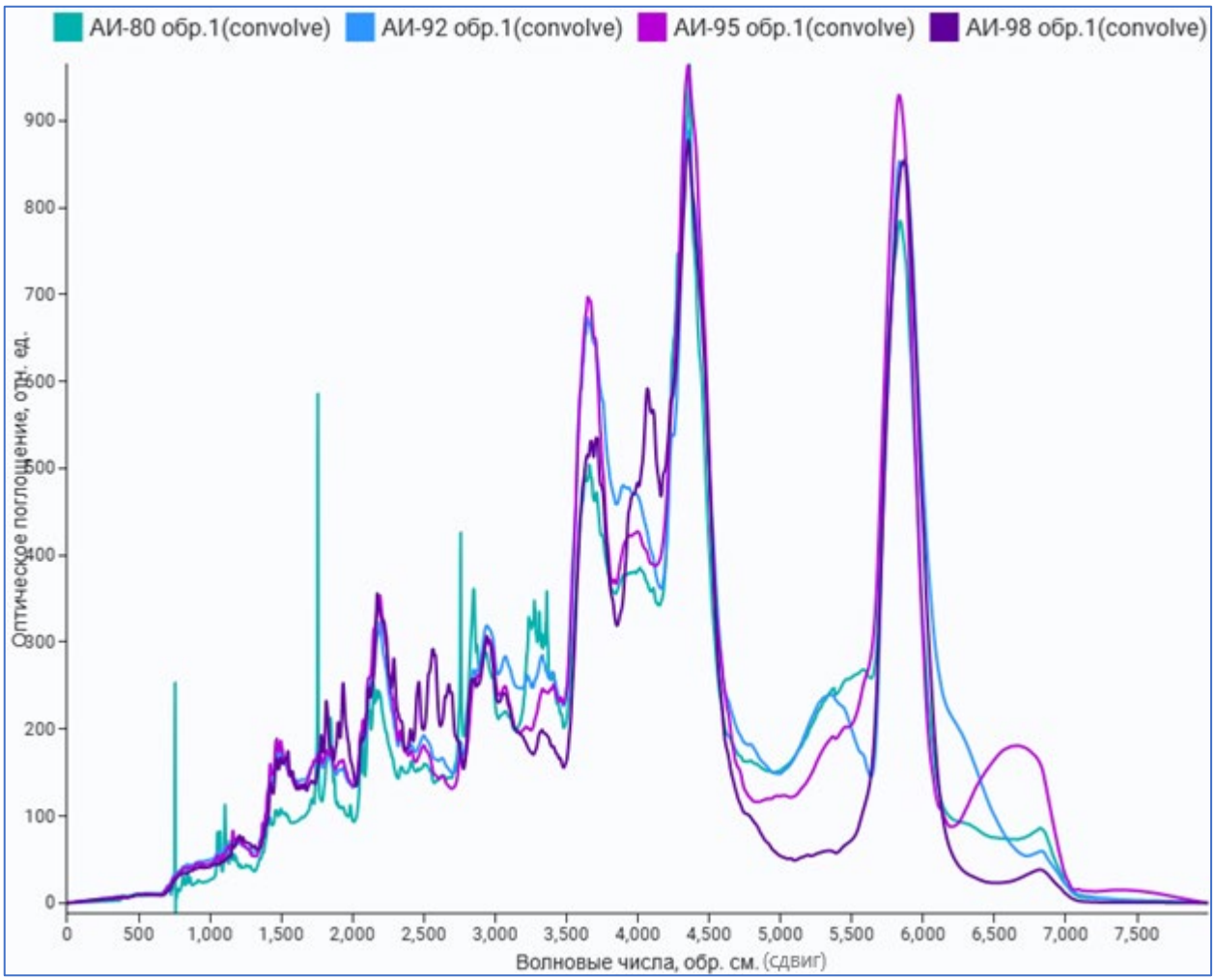


Рисунок 4 – График дискретной свёртки отклика спектров

На рисунке 5 представлен результат автокорреляции отклика веществ.

У образца АИ-80 (бирюзовый) наблюдаются пики около 3000 $см^{-1}$, остальные образцы имеют более сглаженные пики в этом диапазоне.

В той же области можно отметить, что значения образца АИ-98 (фиолетовый) не растут до 3200 $см^{-1}$ в отличие от АИ-92 и АИ-95.

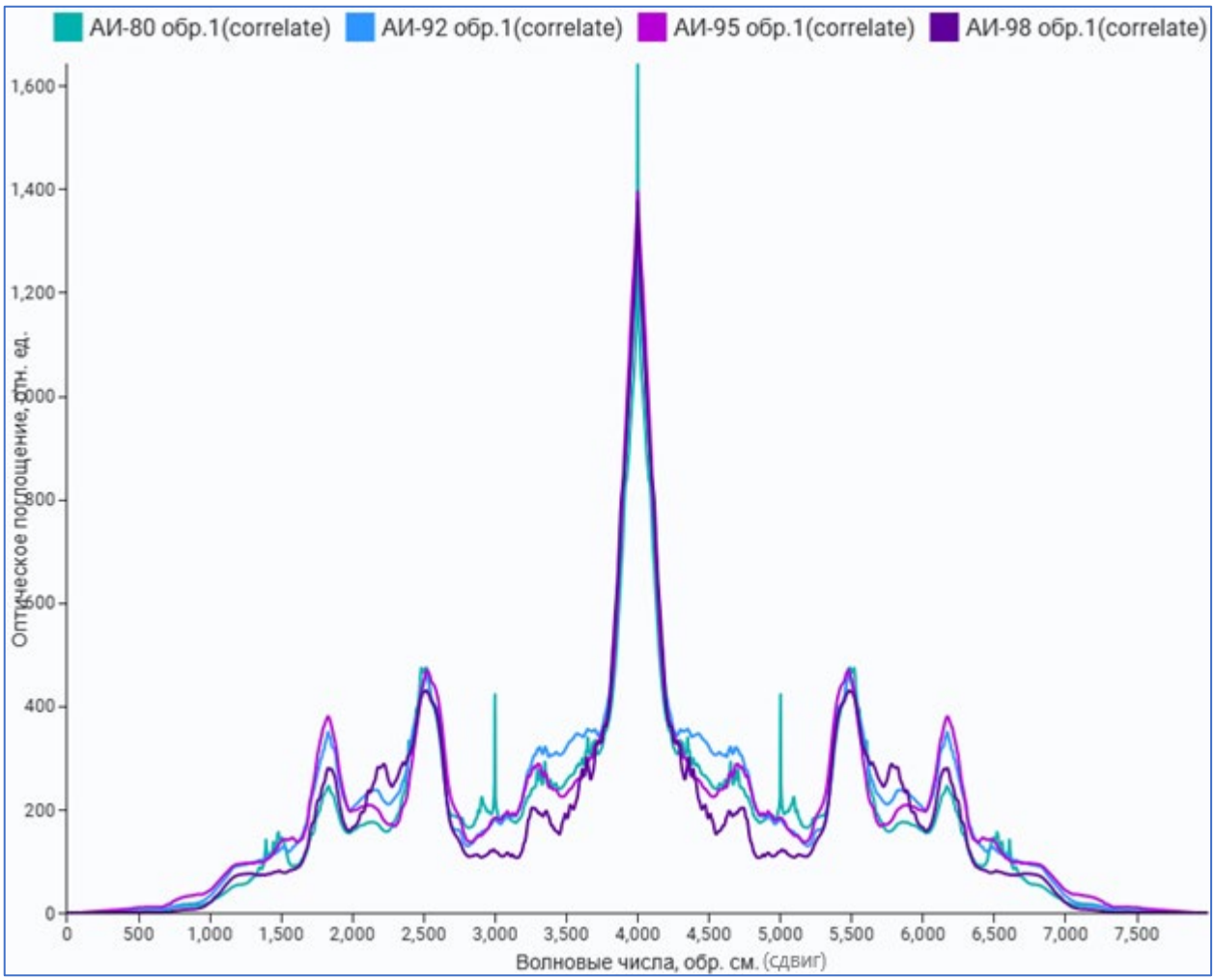


Рисунок 5 – График автокорреляции отклика спектров

График кумулятивной суммы откликов представлен на рисунке 6.

В данном случае можно отметить, что образец АИ-80 (бирюзовый) показывает самый большой локальный рост около 400 $см^{-1}$ и 1400 $см^{-1}$.

Также относительно больший рост показывает образец АИ-98 (темно-фиолетовый) в области от 100 $см^{-1}$ до 1300 $см^{-1}$.

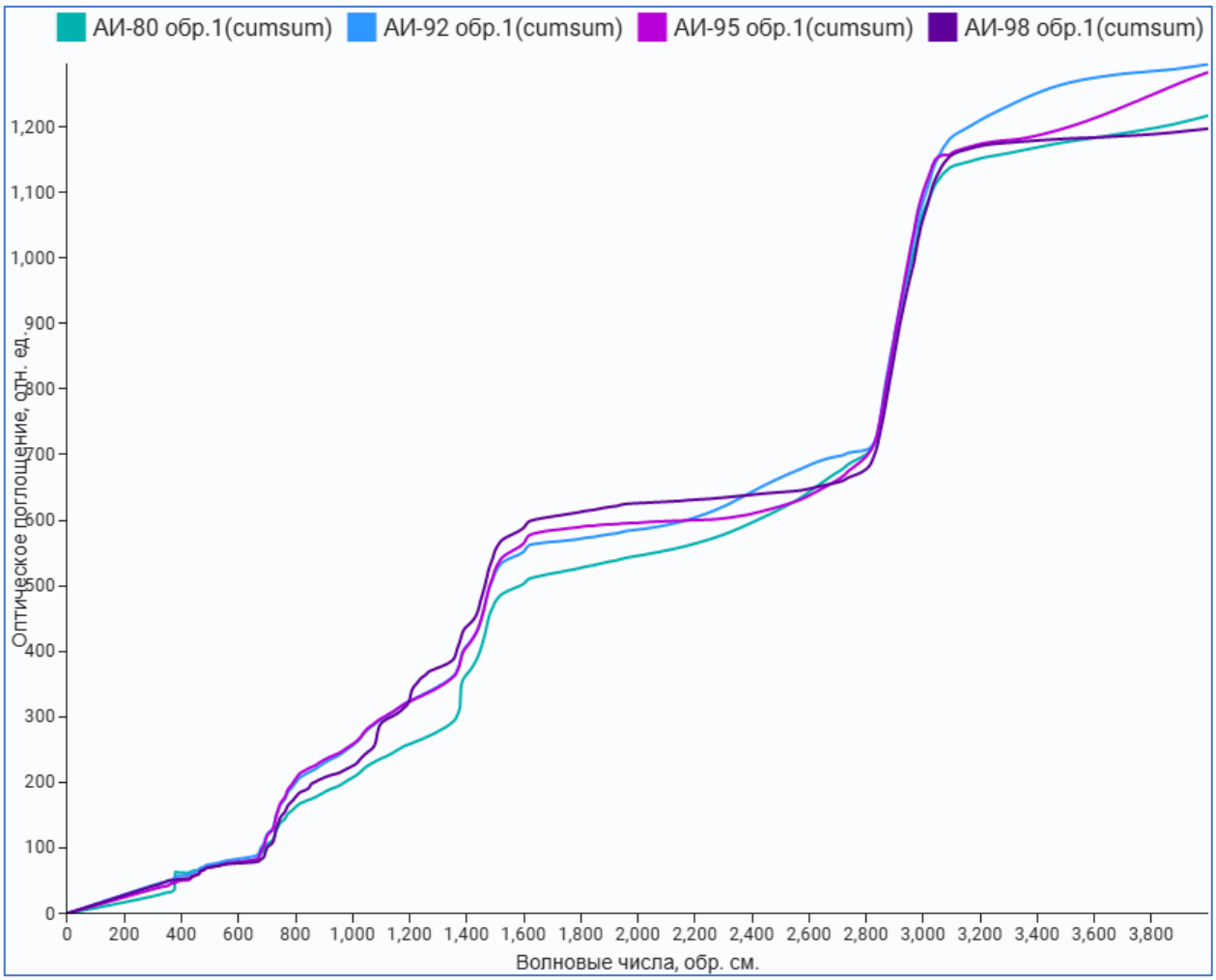


Рисунок 6 – График кумулятивной суммы отклика спектров

На рисунке 7 представлены данные после применения преобразования дискретной производной, что выделяет быстрые изменения в спектре и подчеркивает особенности градиента сигнала.

В диапазоне 2800-3000 $см^{-1}$ видны небольшие флуктуации образца АИ-92 (синий) и несколько позже у АИ-95 (фиолетовый) и АИ-95 (темно-фиолетовый). Образец АИ-98 также имеет выбросы около 1500 $см^{-1}$.

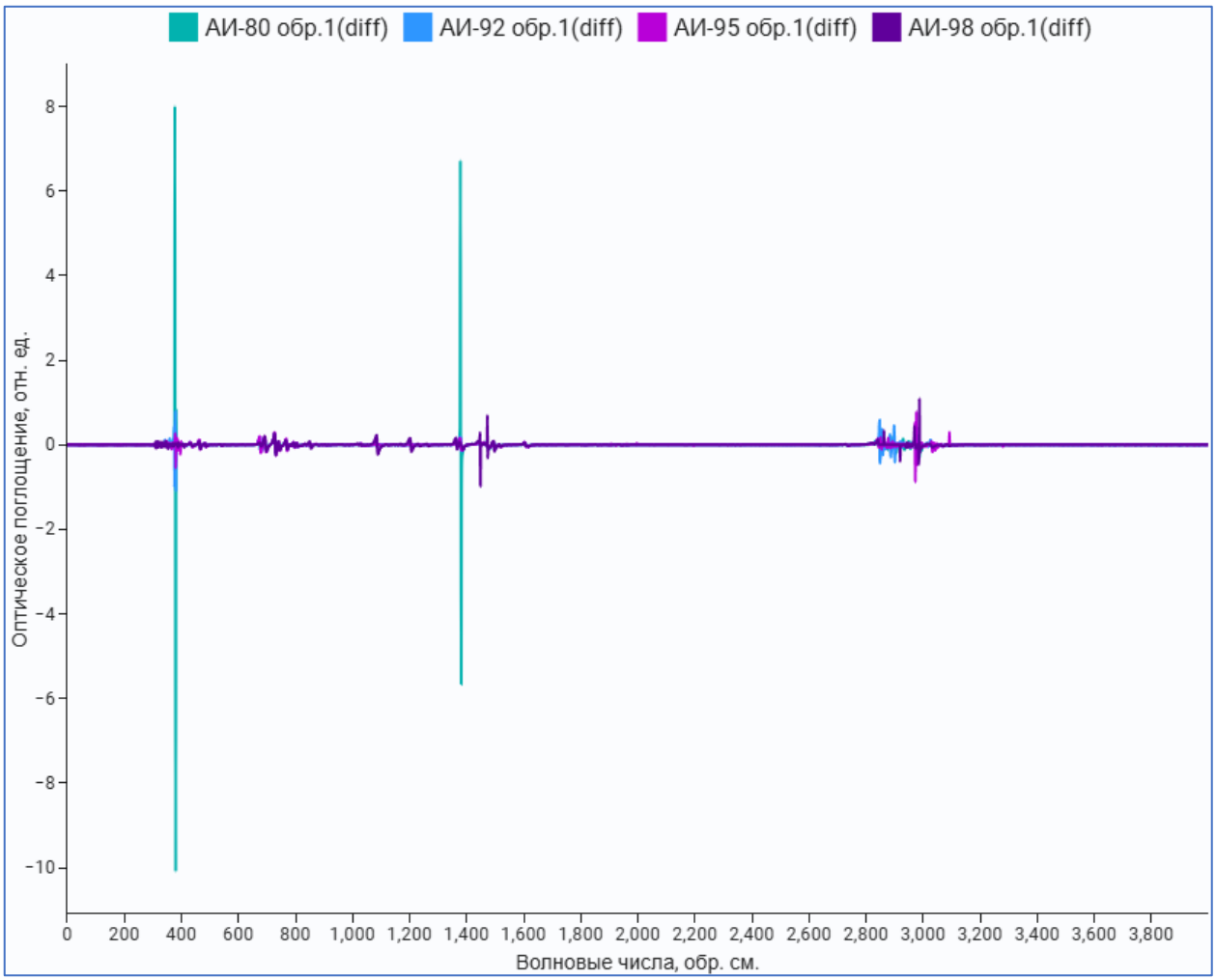


Рисунок 7 – График дискретной производной отклика спектров

АИ-80 (бирюзовый) демонстрирует заметные выбросы в диапазонах около 400 $см^{-1}$ и 1300 $см^{-1}$, которые сильно отличаются от других образцов.

Образец АИ-92 также имеет относительно небольшие выбросы около 400 $см^{-1}$ и еще меньшие выбросы здесь у образца АИ-95. Это можно наблюдать на рисунке 8.

На данном графике можно заметить, как визуально отличимые различия «перетекают» от низкооктановых бензинов к высокооктановым. В области у 400 $см^{-1}$ образец с октановым числом 80 имел явно выраженные отличия, несколько меньшие отличия были у образца с октановым числом 92 и совсем небольшие у 95, образец с октановым числом 98 же здесь вообще не имел никаких выбросов. С другой стороны,

в области от 2800 $см^{-1}$ до 3000 $см^{-1}$ образец с октановым числом 80 не имел выбросов, образец 92 имел небольшие выбросы, а высокооктановые образцы имели большие выбросы.

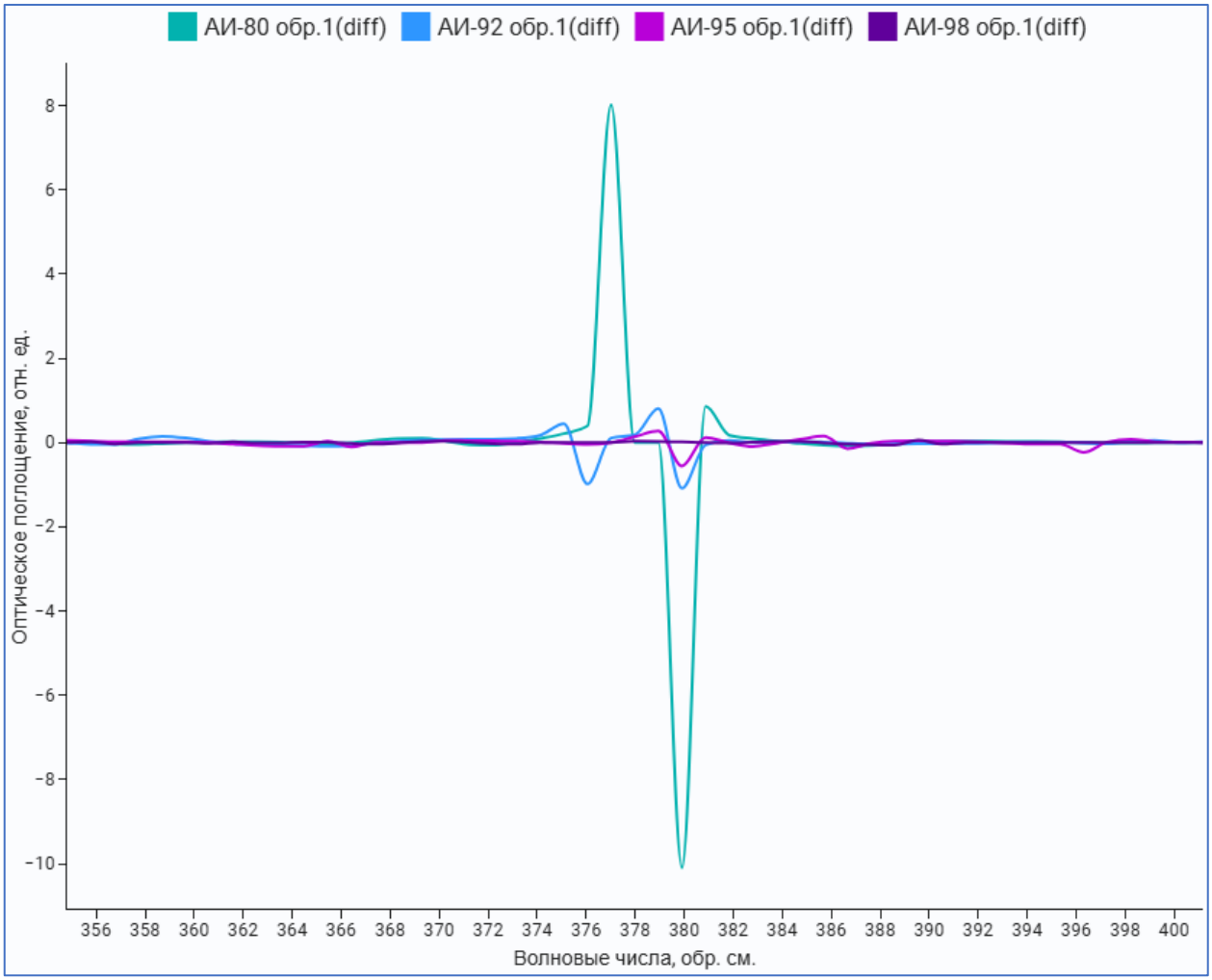


Рисунок 8 – График дискретной производной отклика спектров, фрагмент

На рисунке 9 изображены графики обратных величин спектральных распределений спектров. Такой подход позволяет сильнее выделить колебания на низких значениях и уделять меньше внимания выбросам с высокими номинальными значениями.

АИ-80 (бирюзовый) имеет заметные отрицательные выбросы в области 300-500 $см^{-1}$, когда у образцов АИ-92 (синий) и АИ-98 (темно-фиолетовый) выбросы положительные.

Образец АИ-95 (фиолетовый) выделяется значительными выбросами в диапазоне от 1800 $см^{-1}$ до 2300 $см^{-1}$ и около 3100 $см^{-1}$.

Образец АИ-98 также отличается немного большими значениями в областях от 500 $см^{-1}$ до 700 $см^{-1}$ и от 3200 $см^{-1}$ до 4000 $см^{-1}$.

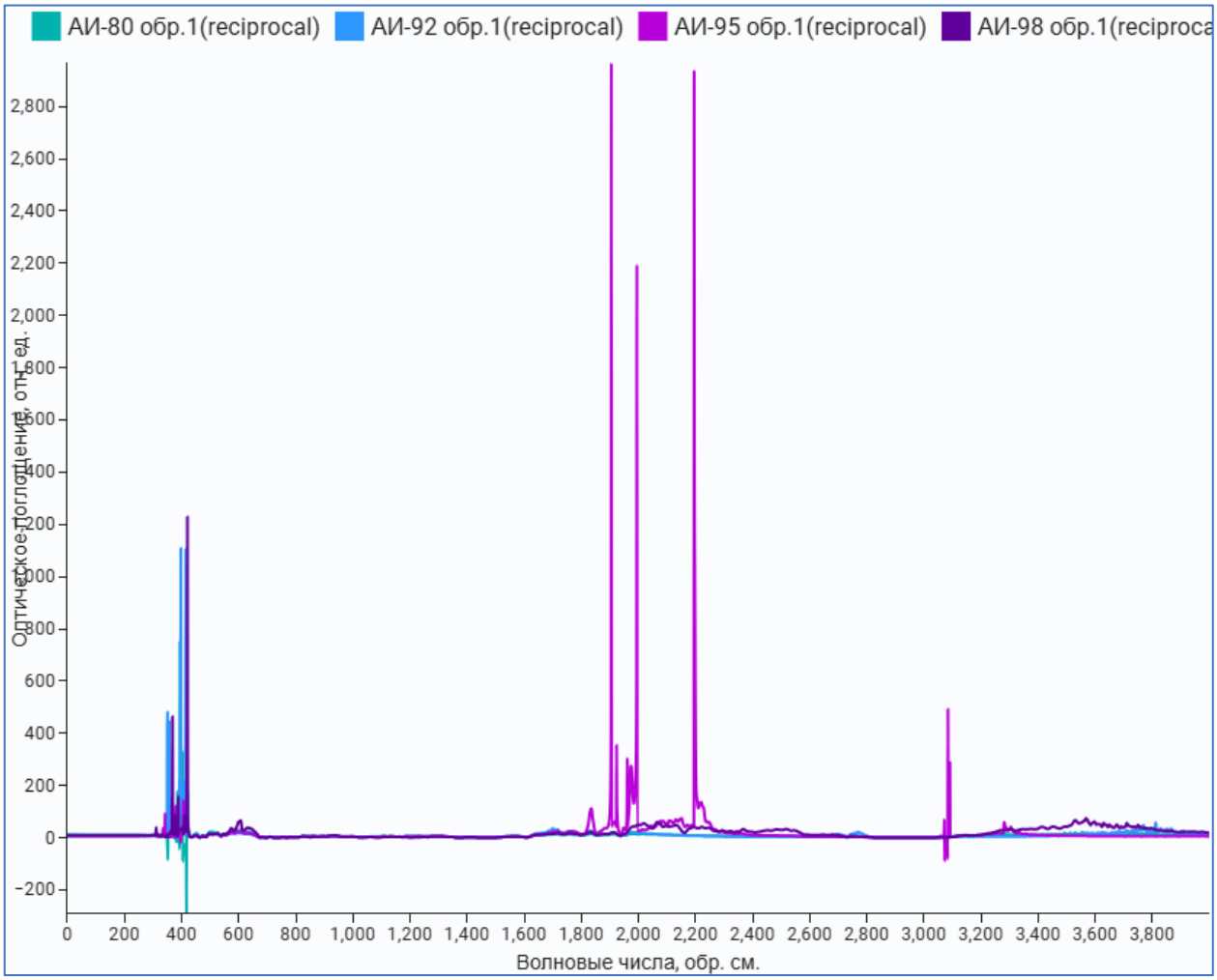


Рисунок 9 – График обратных значений отклика спектров

Результаты последнего варианта предварительной обработки – вычисление квадратов значений – показано на рисунке 10. Такой подход, наоборот, гораздо сильнее

акцентирует внимание на сильных спектральных откликах вещества и снижает влияние слабых колебаний.

Здесь, как и на рисунке с необработанными откликами видны высокие пики значений спектра АИ-80 (бирюзовый) около 400 $см^{-1}$ и 1400 $см^{-1}$.

Образец АИ-92 (синий) также отличается несколько более выраженным откликом около 2900 $см^{-1}$ относительно остальных.

Образец АИ-98 (темно-фиолетовый) также отличается теми же откликами в нескольких областях от 800 $см^{-1}$ до 1300 $см^{-1}$, что заметны уже без приближения графика.

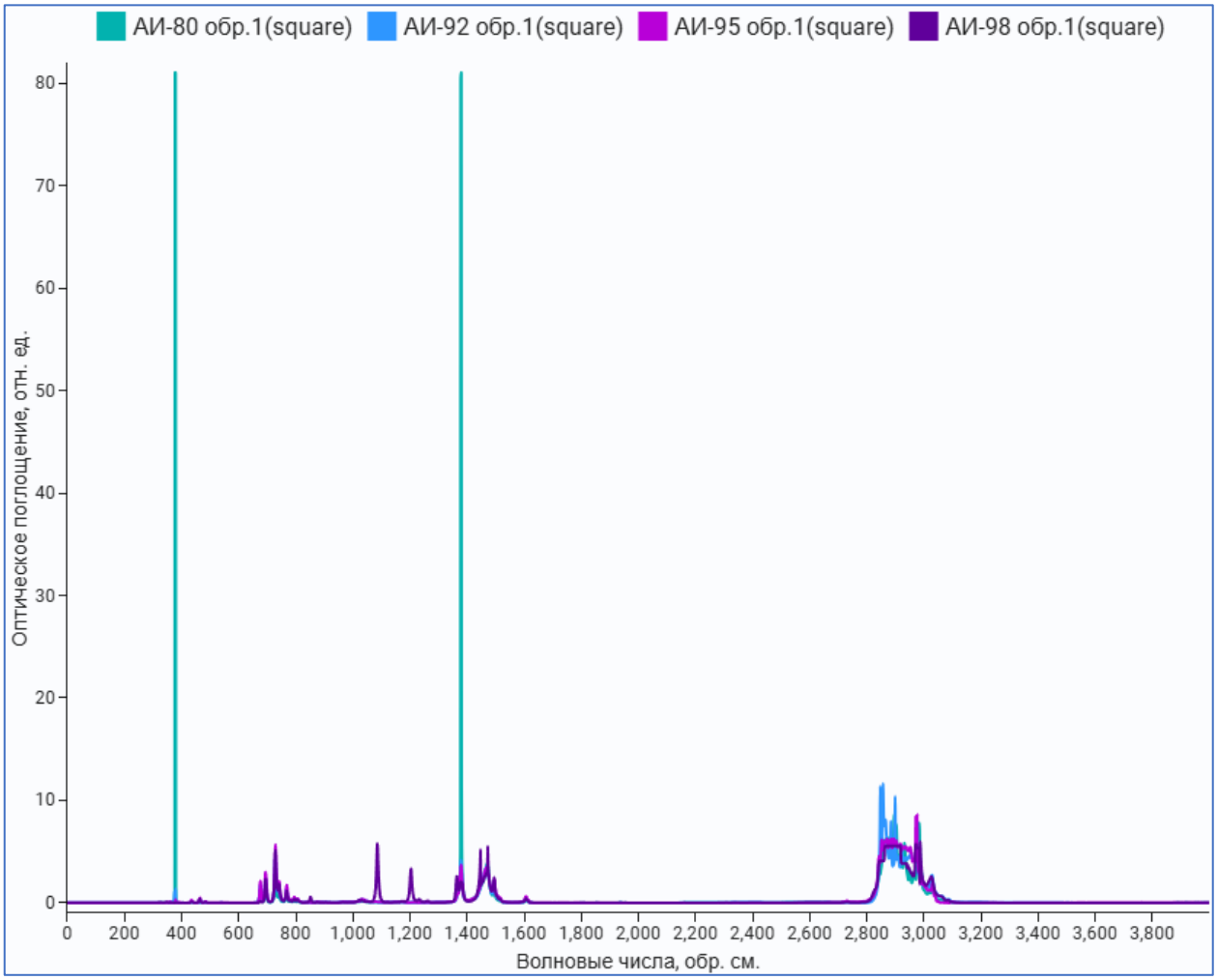


Рисунок 10 – График квадратов значений отклика спектров

Проведенный анализ спектральных данных с применением различных методов предварительной обработки демонстрирует их влияние на выделение признаков и подготовку данных к дальнейшему анализу. Полученные результаты позволяют утверждать, что выбор метода обработки оказывает существенное влияние на форму и структуру данных, что в дальнейшем может повлиять на результаты кластеризации и их интерпретацию.

После этапа предварительной обработки данные подаются на вход алгоритмам снижения размерности с различными метриками. Такой подход дает возможность оценить влияние различных методов обработки на результаты анализа, а также выявить наиболее информативные признаки и закономерности в данных. Разные подходы конкретных алгоритмов и применение разных метрик может найти закономерности в данных там, где их сложно заметить глазом исследователя. Так что даже кажущиеся неудачными методы предварительной обработки могут оказаться весьма успешными [56].

3.2. Анализ применения алгоритмов снижения размерности

Далее приведена оценка результатов работы алгоритмов снижения размерности с использованием различных метрик на различных типов предварительно обработанных данных. Помимо визуальной оценки, результаты снижения размерности будут оценены с помощью коэффициента силуэта кластера (SC) и индекса Дэвиса-Боулдина (DBI).

Коэффициент силуэта кластера – это метрика, используемая для оценки качества кластеризации, позволяющая количественно определить степень компактности и разделимости кластеров [144]. Он был предложен Питером Руссо в 1987 году и

основывается на анализе двух ключевых параметров для каждого объекта: среднего расстояния до других точек внутри того же кластера (a) и минимального среднего расстояния до точек ближайшего соседнего кластера (b). Опытным путем было выявлено, что эти расстояния лучше всего вычислять с помощью метрики Канберры [60]. Коэффициент силуэта для отдельного объекта i вычисляется по формуле:

$$SC(i) = \frac{b(i) - a(i)}{maxa(i), b(i)}. \quad (32)$$

Значение коэффициента варьируется в диапазоне от -1 до 1. Чем ближе значение к 1, тем выше внутрикластерная компактность и межкластерная разделимость. Если коэффициент близок к 0, это указывает на пересечение кластеров или неоптимальное их разделение, а отрицательные значения сигнализируют о возможной ошибочной кластеризации. Среднее значение коэффициента силуэта по всем объектам даёт интегральную оценку качества разбиения, что делает его удобным инструментом для сравнения различных алгоритмов или параметров кластеризации.

Однако метрика имеет ограничения: высокая вычислительная сложность при работе с большими данными из-за необходимости расчёта попарных расстояний, а также чувствительность к форме кластеров. Для кластеров сложной геометрии или низкой плотности коэффициент может давать заниженные оценки, несмотря на содержательную значимость группировки [20].

Индекс Дэвиса-Боулдина – это метрика оценки качества кластеризации, основанная на анализе внутрикластерной компактности и межкластерной разделимости [66]. Предложенный в 1979 году Дэвидом Дэвисом и Дональдом Боулдином индекс вычисляется как среднее значение схожести между каждым кластером и его ближайшим аналогом. Для его расчёта используются две ключевые характеристики: внутрикластерное рассеяние (среднее расстояние между точками

внутри кластера и его центроидом) и межкластерное расстояние (например, евклидово расстояние между центроидами кластеров). Для кластера C_i его схожесть с другим кластером C_j определяется как отношение суммы внутрикластерных расстояний S_i и S_j к расстоянию между их центроидами d_{ij}:

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}. \quad (33)$$

Индекс Дэвиса-Боулдина представляет собой среднее значение максимальных значений R_{ij} для N кластеров:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} \max_{i \neq j} R_{ij}\,. \quad (34)$$

Чем меньше значение индекса, тем выше качество кластеризации: компактные, хорошо разделённые кластеры минимизируют схожесть между собой. Высокие значения, напротив, указывают на перекрытие кластеров или их неоднородность. Основное преимущество DBI заключается в его вычислительной эффективности, так как он опирается на центроиды и не требует расчёта попарных расстояний между всеми объектами, что делает его применимым для больших объемов данных.

Однако DBI имеет ряд ограничений: он предполагает сферическую форму кластеров, что снижает его эффективность для данных сложной геометрии. Кроме того, индекс может давать некорректные оценки при существенной разнице в плотности кластеров или наличии шума. Эти особенности важно учитывать при его использовании в задачах, где кластеры имеют невыпуклую структуру или неравномерное распределение [95].

Экспериментальное сравнение алгоритмов DBSCAN и K-Means на случайном наборе данных (669 точек, разбитых на 3 кластера) показало, что DBSCAN демонстрирует лучший результат по индексу Дэвиса-Боулдина (0,05 против 0,075 у K-

means). Однако визуальный анализ результатов свидетельствует, что K-means обеспечивает более геометрически точное разбиение. Авторы делают вывод о необходимости выбора метода кластеризации в зависимости от специфики данных, подчеркивая важность эвристического подхода [44].

3.2.1. Оценка изометрического отображения

Метод Isomap показывал разные результаты на разных входных данных: от удовлетворительных результатов при использовании автокорреляции до отличных результатов при использовании данных, обработанных дискретной производной. Ниже будут показаны результаты применения алгоритма на тех данных, на которых был достигнут лучший результат при использовании конкретной метрики.

На рисунке 11 показаны лучшие результаты применения алгоритма Isomap с применением метрик Чебышёва на необработанных данных (слева) и Хэмминга на данных, обработанных дискретной производной (справа). Оценка визуальной группировки данных может быть проведена следующим образом.

В первом случае точки, принадлежащие классам АИ-92 (синий) и АИ-95 (бирюзовый) сильно смешались друг с другом, хотя точки АИ-95 сгруппировались лучше. Точки АИ-80 (фиолетовый) Также неплохо сгруппированы, но часть из них смешалась с точками АИ-92 и АИ-95. Точкам АИ-98 (желтый) удалось сформировать удаленный сжатый кластер. DBI в данном случае равен 0.992718, а SC 0.351776.

Во втором случае никаким точкам не удалось сформировать своего удаленного сжатого кластера. DBI в данном случае равен 2.74996, а SC 0.340507.

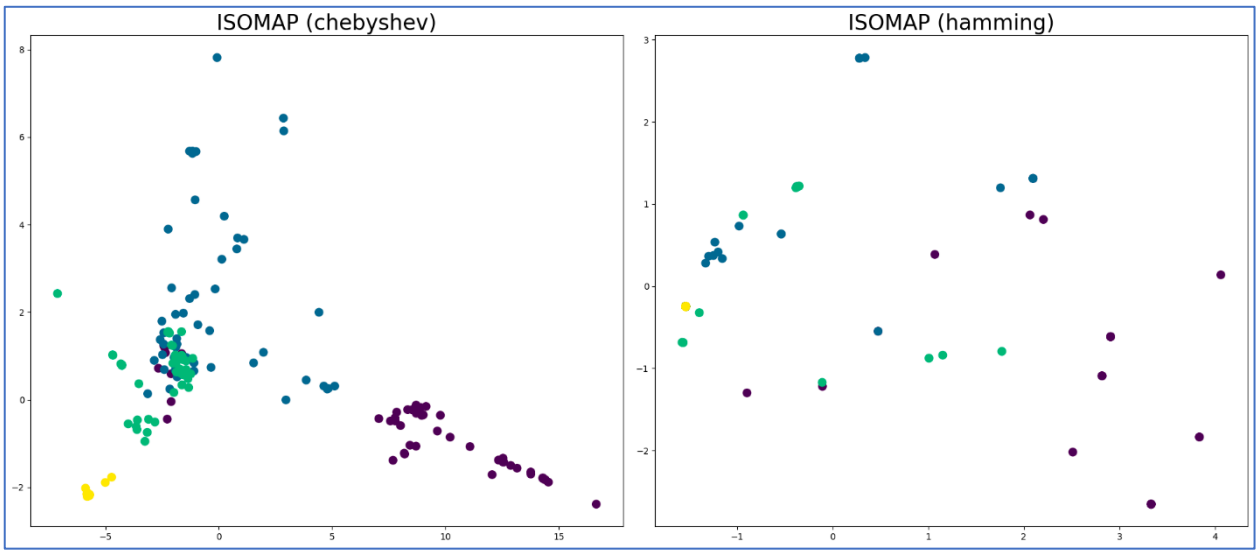


Рисунок 11 – Результаты работы алгоритма снижения размерности ISOMAP (chebyshev, hamming)

Несколько лучшие результаты показаны на рисунке 12. Здесь, при использовании косинусной метрики на данных, обработанных дискретной производной, (слева сверху) всем точками удалось сформировать хоть и не плотные, но различимые кластеры. Однако точки АИ-95 (бирюзовый) разбились на два кластера, разделенных другим. DBI в данном случае равен 1.132996, а SC 0.504984.

Использование манхэттенской метрики на необработанных данных (справа сверху) дало результат, где точки классов АИ-80 (фиолетовый), АИ-92 (синий) и АИ-95 (бирюзовый) хорошо сгруппированы, но их кластеры слабо отличимы друг от друга. Точки АИ-98 (желтый) сформировали плотный кластер, достаточно удаленный от других. DBI в данном случае равен 1.244369, а SC 0.420771.

Использование квадратной евклидовой метрики на необработанных данных (слева снизу) дало схожий результат, где точки классов АИ-80 (фиолетовый), АИ-92 (синий) и АИ-95 (бирюзовый) сбились в близкие слабо отличимые классы, а АИ-98 (желтый) отдаленный от них сжатый кластер. DBI в данном случае равен 0.702482, а SC 0.613864.

Корреляционная метрика на данных дискретной свёртки (справа снизу) в целом дала тот же результат, но здесь кластеры получились несколько более сжатыми и различимыми. DBI в данном случае равен 0.50587, а SC 0.609797.

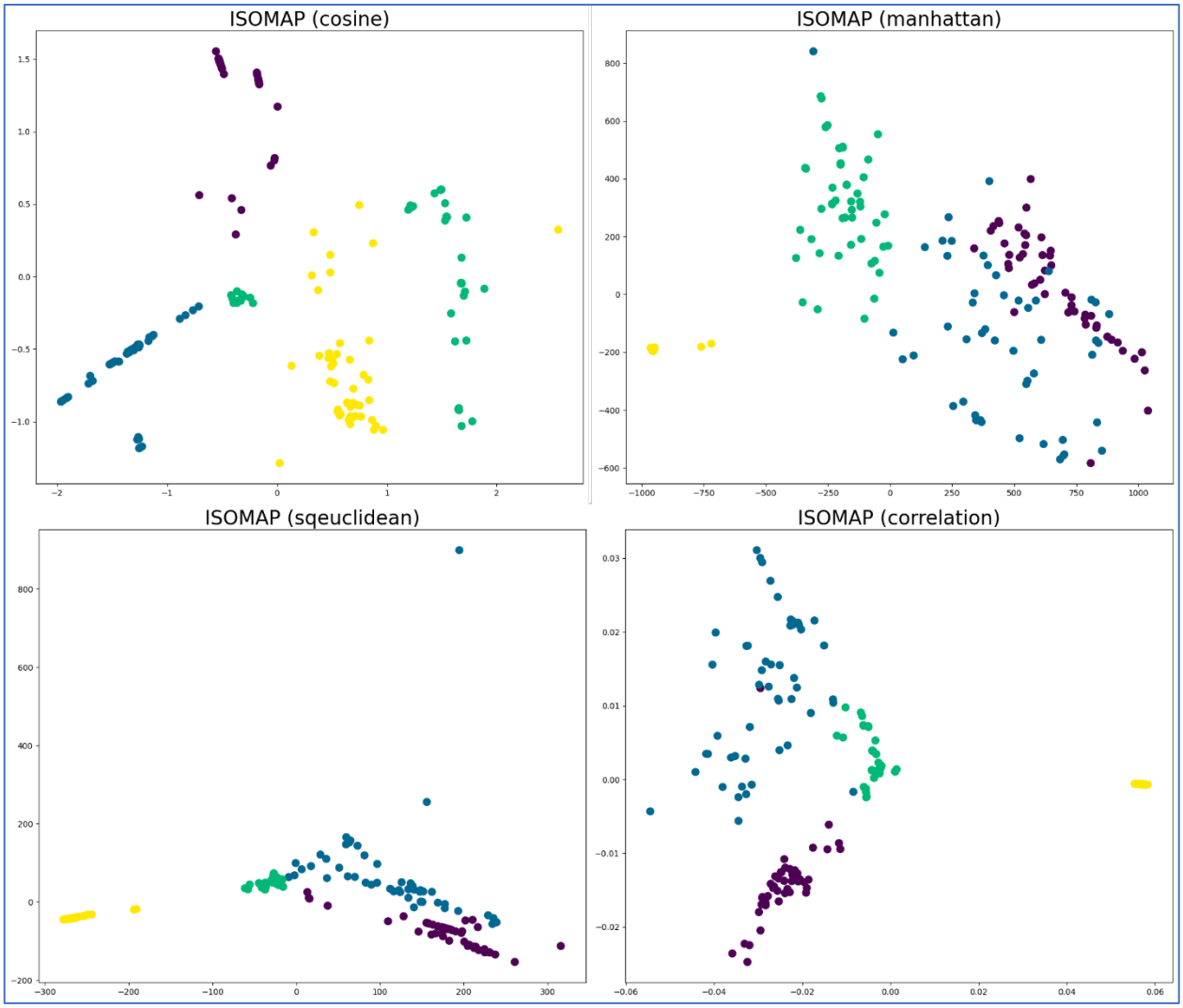


Рисунок 12 – Результаты работы алгоритма снижения размерности ISOMAP (cosine, manhattan, sqeuclidean, correlation)

Лучшие результаты применения алгоритма Isomap показаны на рисунке 13.

С помощью евклидовой метрики на необработанных данных удалось достичь результата, схожими с предыдущими, однако здесь кластеры получились еще более сжатыми и различимыми. DBI в данном случае равен 0.602591, а SC 0.573585.

Еще лучше результаты получились при использовании метрики Канберры на данных, обработанных дискретной производной. При применении этой метрики большинство точек распределились вдоль трех линий, образовав довольно хорошо различимые, хоть и разреженные кластеры. Почти все точки АИ-98 (желтый) разбились на три группы вдоль своей линии, некоторые из них отклонились в сторону центра и другой линии. Точки АИ-95 (бирюзовый) разбились на две группы: одна отчетливо распределилась вдоль своей линии, вторая расположилась около центра со стороны линии АИ-98. Все точки АИ-80 (фиолетовый) отчетливо расположились вдоль совей линии. Точки АИ-92 (синий) образовали плотный кластер около центра со стороны АИ-80. DBI в данном случае равен 0.680483, а SC 0.602114.

Использование метрики Брея-Кёртиса на данных, обработанных дискретной производной, также позволило достичь очень хороших результатов. Здесь точкам АИ-80 (фиолетовый) и АИ-92 (синий) удалось образовать плотные отличимые кластеры. У точек АИ-95 (бирюзовый) кластер получился менее плотным и несколько смешанным с точками АИ-98 (желтый), которые образовали очень разреженный, но довольно отличимые кластер. DBI в данном случае равен 0.590789, а SC 0.452809.

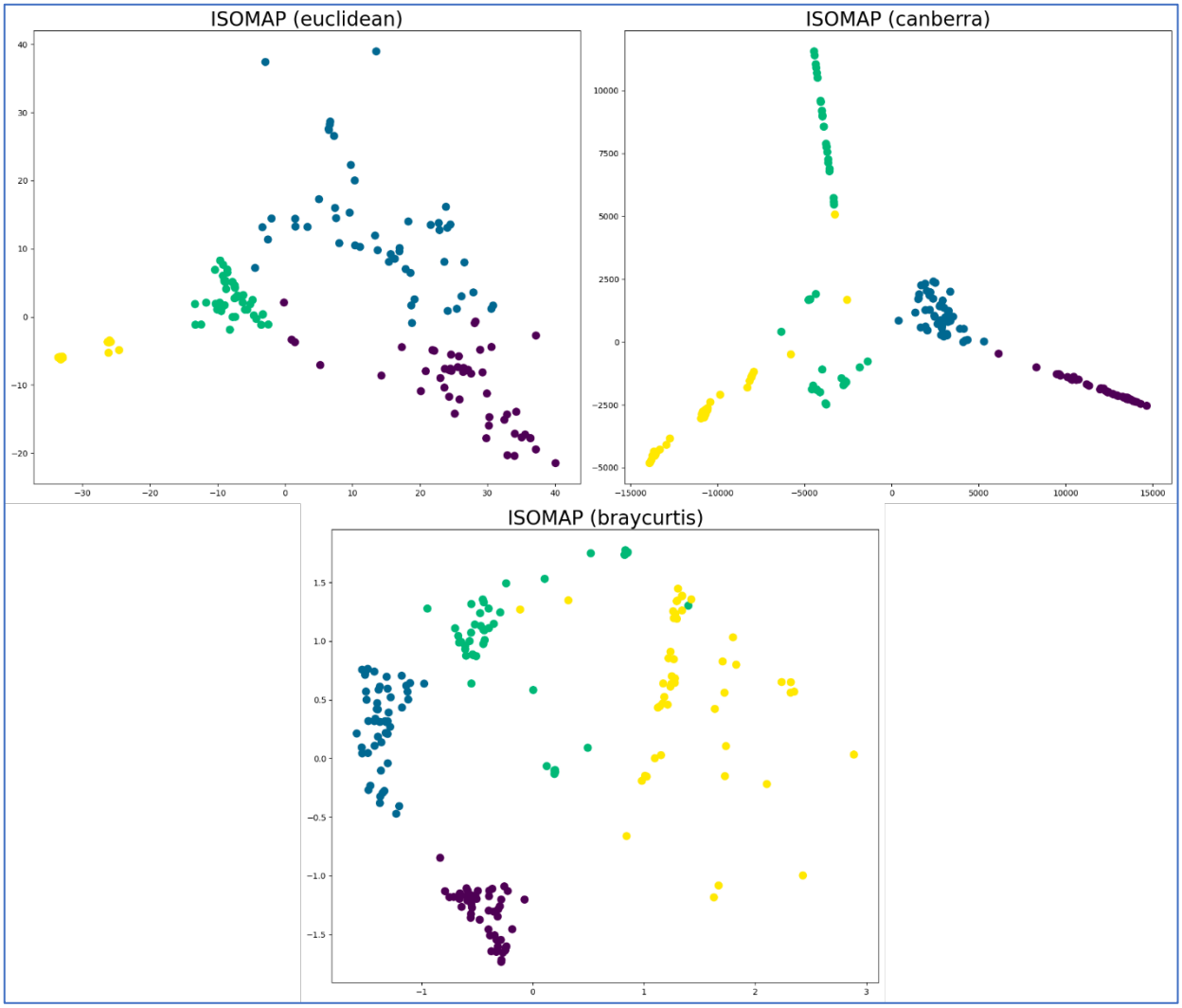


Рисунок 13 – Результаты работы алгоритма снижения размерности ISOMAP (euclidean, canberra, braycurtis)

По оценке SC, приведенной в таблице 1 ISOMAP демонстрирует умеренную производительность с результатами до 0.6313 для необработанных данных с использованием метрики Брея-Кёртиса. Корреляционная метрика в среднем дает лучшие результаты. Лучше всего алгоритм справляется с необработанными данными.

Таблица 1 – Оценка SC ISOMAP

data\metric	braycurtis	canberra	chebyshev	correlation	c o s i n e	euclidean	hamming	manhattan	sqeuclidean
convolve	0.362	0.345	0.310	0.610	0.536	0.312	-0.055	0.312	0.379
correlate	0.255	0.264	0.316	0.454	0.402	0.255	-0.057	0.283	0.307
cumsum	0.274	0.292	0.279	0.519	0.492	0.274	0.172	0.280	0.288
diff	0.453	0.602	0.100	0.505	0.505	0.228	0.341	0.150	0.270
normal	0.631	0.510	0.352	0.568	0.563	0.574	0.302	0.421	0.614
reciprocal	0.271	0.510	-0.107	0.312	0.297	0.111	0.300	0.231	-0.028
square	0.417	0.495	0.212	0.395	0.393	0.186	0.119	0.483	0.365

По оценке DBI, приведенной в таблице 2 алгоритм лучше всего показал себя на данных свёртки с использованием корреляционной метрики с оценкой 0.5059. В среднем, алгоритм дает лучшие результаты при использовании корреляционной метрики или необработанных данных.

Таблица 2 – Оценка DBI ISOMAP

data	braycurtis	canberra	chebyshev	correlation	c o s i n e	euclidean	hamming	manhattan	sqeuclidean
convolve	2.173	2.693	1.825	0.506	0.625	1.937	14.486	2.633	1.825
correlate	4.307	2.565	2.094	0.976	0.899	2.382	21.348	3.418	2.164
cumsum	3.481	2.432	3.850	0.717	1.029	3.374	5.205	3.285	2.380
diff	0.591	0.681	1.385	1.133	1.133	1.082	2.750	1.756	1.212
normal	0.699	3.618	0.993	0.766	0.764	0.603	1.510	1.244	0.703
reciprocal	2.553	3.618	5.303	1.599	1.785	3.685	1.516	1.559	2.500
square	1.944	3.909	1.479	1.983	1.860	1.156	5.807	1.094	2.856

3.2.2. Оценка локального линейного вложения

Метод снижения размерности локально-линейного вложения свои лучшие результаты на необработанных входных данных, а также входных данных, обработанных дискретной свёрткой, автокорреляцией, квадратным преобразованием в зависимости от используемого метода. Ниже будут показаны результаты применения алгоритма на тех данных, на которых был достигнут лучший результат в случае использования конкретного алгоритма.

На рисунке 14 представлены результаты стандартного алгоритма на данных, обработанных дискретной свёрткой. Видно, что кластеры АИ-80 (фиолетовый), АИ-92 (синий) и АИ-95 (бирюзовый) оказались сильно сжатыми и распределенными вдоль общей линии. Хотя кластеры практически сливаются в один, видно, что они распределены в разных частях общей линии, но границу обозначить невозможно. Кластер АИ-98 (желтый) расположился вдали от общей линии трех других, и сжался в одну точку. DBI в данном случае равен 0.575881, а SC 0.562767.

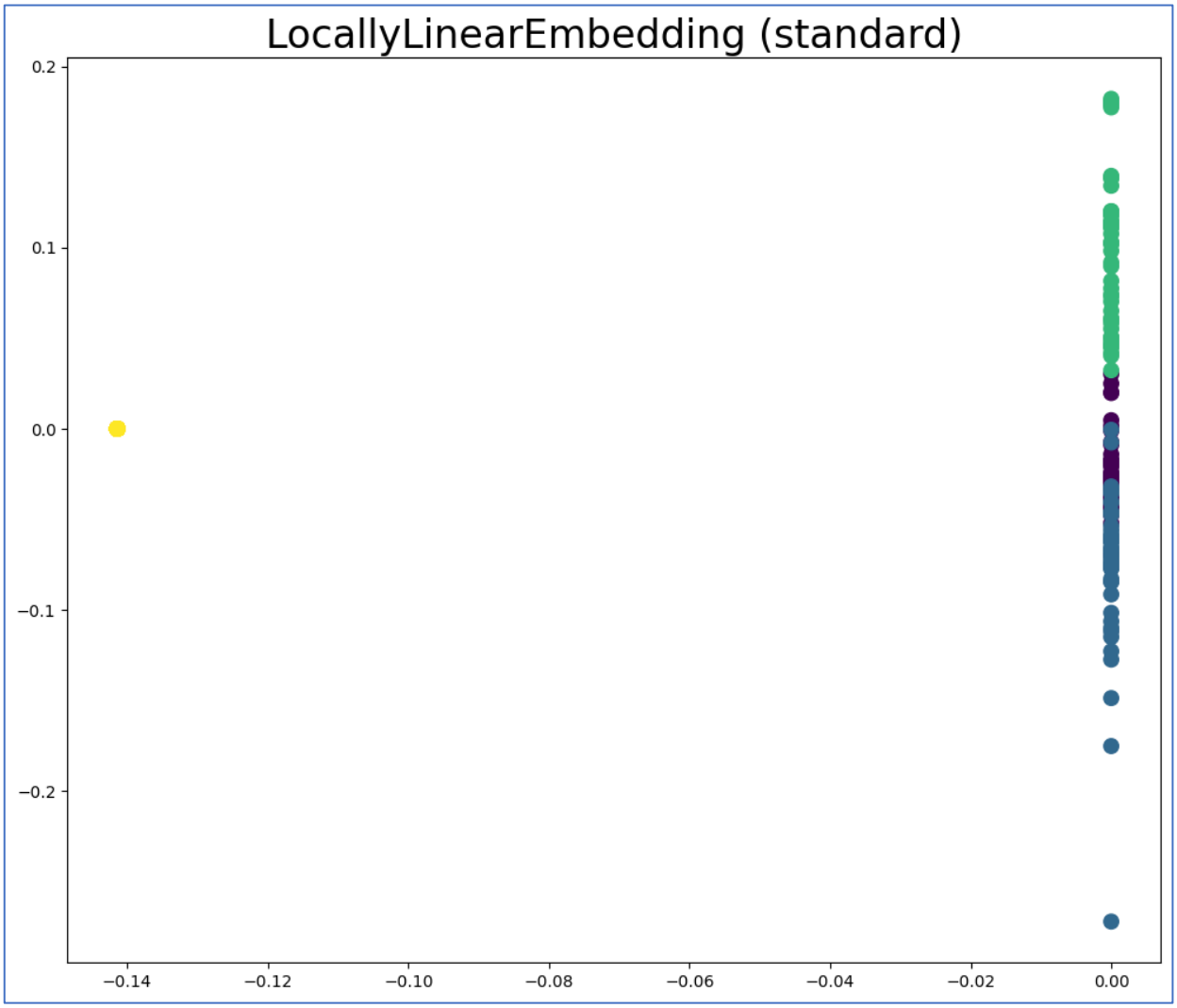


Рисунок 14 – Результаты работы алгоритма снижения размерности Locally Linear Embedding (стандартный)

Модифицированный алгоритм, примененный к необработанным данным, показал лучшие результаты, показанные на рисунке 15. Точки АИ-98 (желтый) и АИ-95 (бирюзовый) образовали очень плотные и отличимые кластеры, а образцы АИ-80 (фиолетовый) и АИ-92 (синий) образовали более разреженные и менее отличимые кластеры. DBI в данном случае равен 0.489917, а SC 0.576118.

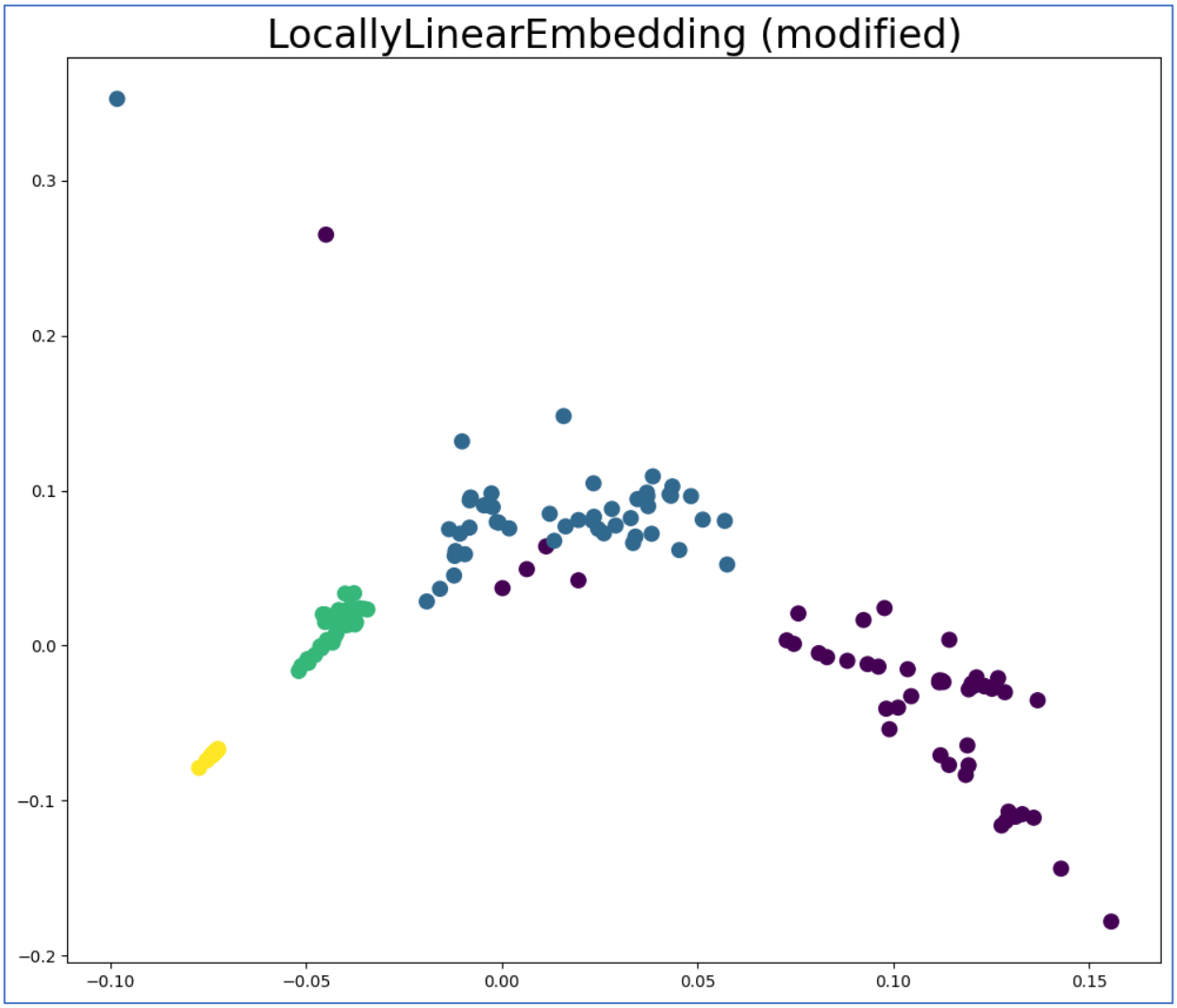


Рисунок 15 – Результаты работы алгоритма снижения размерности Locally Linear Embedding (модифицированный)

На рисунке 16 показаны результаты локально-линейного вложения Гессе необработанных данных. Результаты схожи с результатами стандартного алгоритма,

однако здесь все 4 кластера выстроились вдоль одной линии, и они менее различимы вдоль нее. DBI в данном случае равен 1.202844, а SC 0.443729.

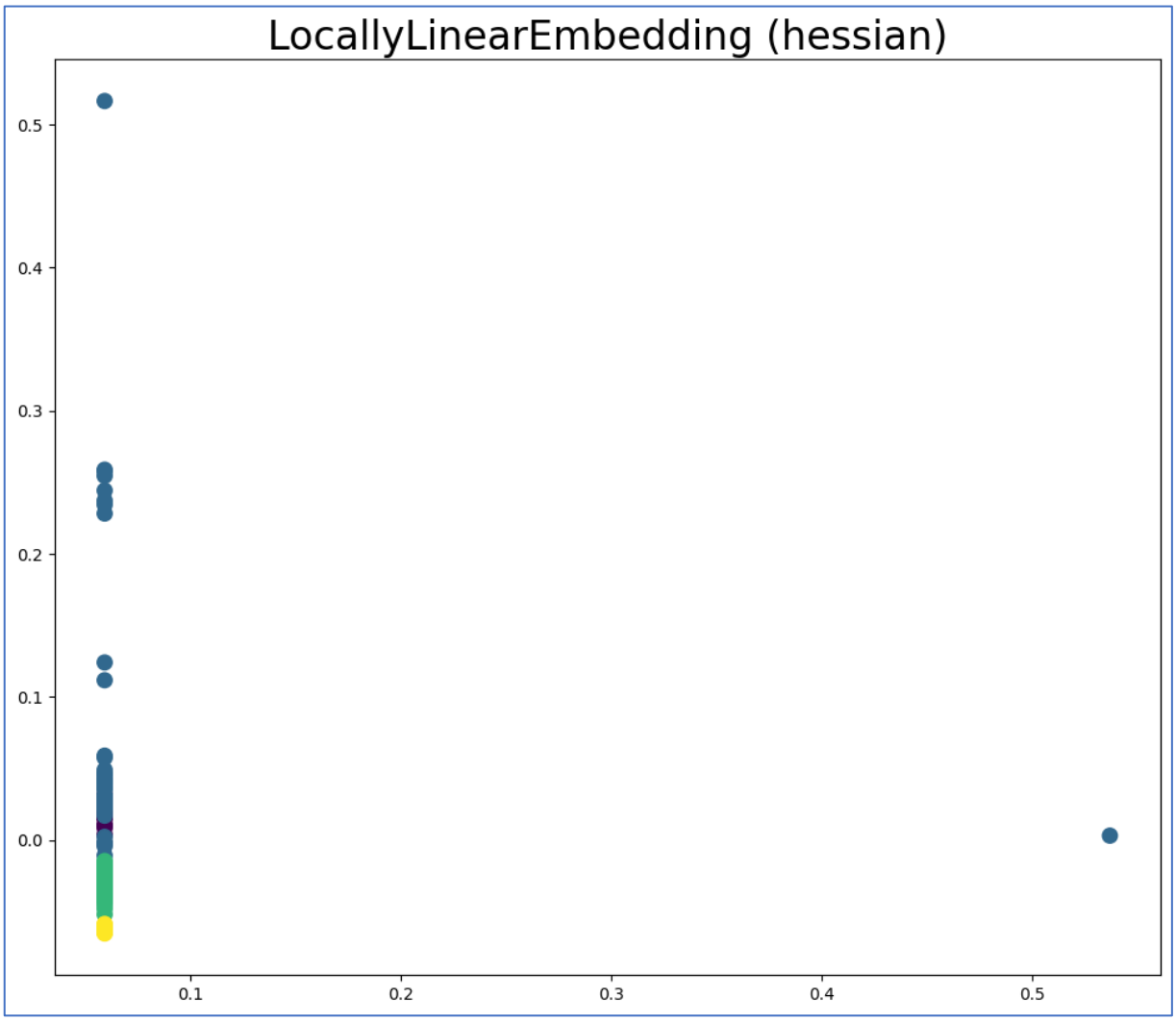


Рисунок 16 – Результаты работы алгоритма снижения размерности Locally Linear Embedding (Гессе)

На рисунке 17 показаны результаты применения алгоритма локального выравнивания касательного пространства на необработанных данных. Применение этого алгоритма дало неудовлетворительный результат, идентичный варианту Гессе. DBI в данном случае равен 1.333207, а SC 0.443254.

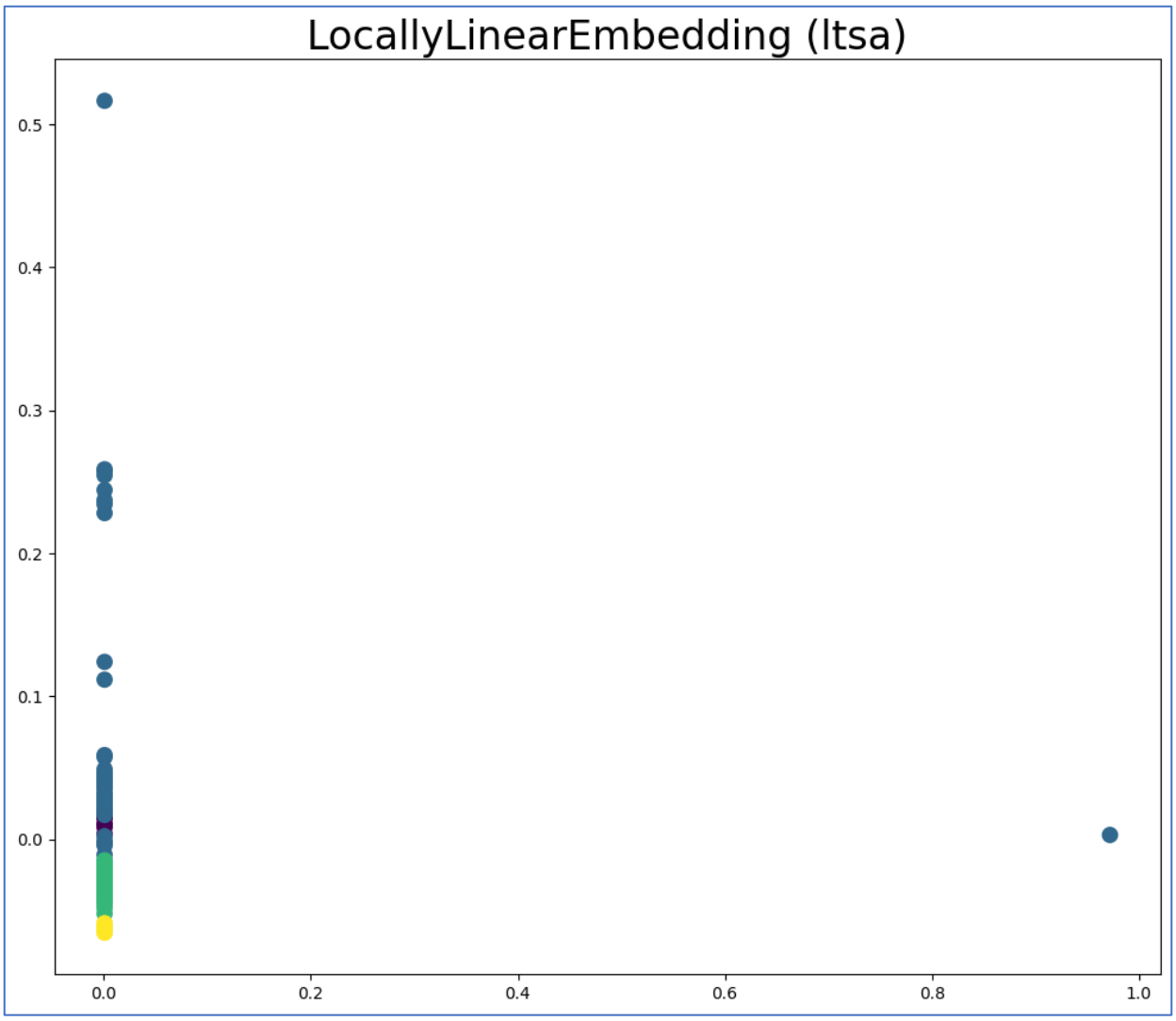


Рисунок 17 – Результаты работы алгоритма снижения размерности Locally Linear Embedding (касательного пространства)

По оценке SC, приведенной в таблице 3, алгоритм лучше всего показал себя в стандартной вариации на данных свёртки и в модифицированном варианте на необработанных данных с оценками 0.6225 и 0.6149 соответственно. Алгоритм в среднем дает приемлемые результаты в стандартном варианте или на необработанных данных.

Таблица 3 – Оценка SC LLE

data\algorithm	hessian	ltsa	modified	standard
convolve	0.278	0.278	0.342	0.623
correlate	0.229	0.263	0.336	0.484
cumsum	0.294	0.294	0.298	0.349
diff	0.118	-0.027	0.226	0.422
normal	0.444	0.443	0.615	0.491
reciprocal	-0.257	-0.261	-0.147	0.295
square	-0.240	-0.240	0.380	0.322

По оценке DBI, приведенной в таблице 4, алгоритм лучше всего показал себя в варианте Гёссе на данных свёртки с оценкой 0.3746, что противоречит оценке SC. Алгоритм в среднем дает хорошие результаты в стандартном варианте или с использованием необработанных данных.

Таблица 4 – Оценка DBI LLE

data\algorithm	hessian	ltsa	modified	standard
convolve	0.375	1.286	2.596	0.508
correlate	1.485	1.477	2.427	0.862
cumsum	1.920	2.461	1.761	1.579
diff	1.676	2.346	1.066	0.716
normal	1.203	1.201	0.490	0.682
reciprocal	3.141	4.368	2.284	1.813
square	2.262	2.346	1.403	1.033

3.2.3. Оценка многомерного масштабирования

Алгоритм MDS показал приемлемый результат только на необработанных данных. Результаты продемонстрированы на рисунке 18 и напоминают некоторые результаты алгоритма ISOMAP. Точки классов АИ-80 (фиолетовый) и АИ-92 (синий) практически слились в один очень разреженный кластер. Точкам АИ-95 (бирюзовый) удалось сформировать более сжатый кластер, но он также слился с предыдущими. А

образцам АИ-98 (желтый) удалось сформировать отдаленный от них кластер, хоть и тоже не сильно сжатый. DBI в данном случае равен 1.998305, а SC 0.241635.

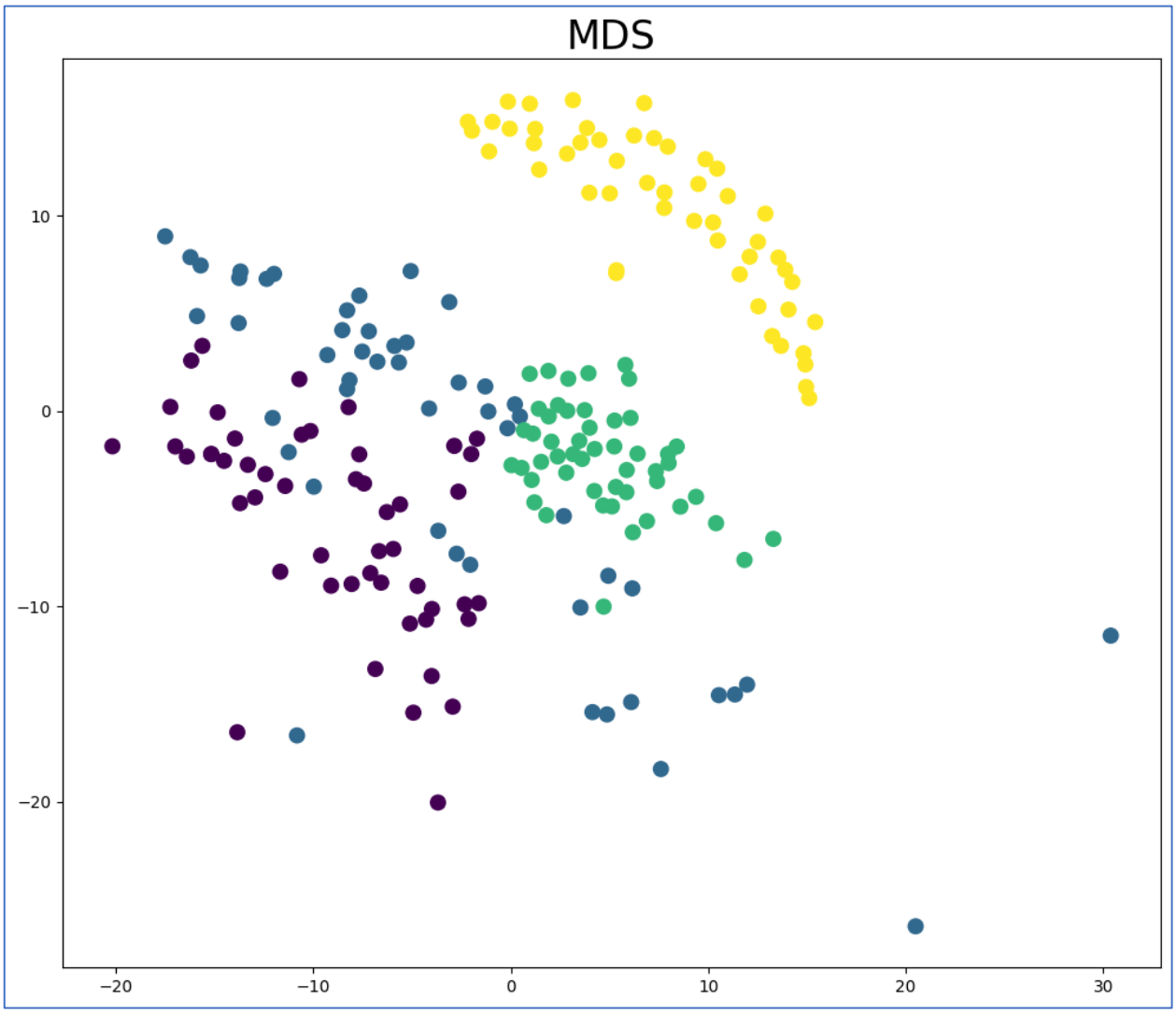


Рисунок 18 – Результаты работы алгоритма снижения размерности MDS

По оценке SC, приведенной в таблице 5, алгоритм MDS показал себя одним из худших алгоритмов с оценкой не более 0.25. Лучше всего алгоритм показал себя на свёртке и необработанных данных с оценками 0.2530 и 0.2416 соответственно.

Таблица 5 – Оценка SC MDS

data\algorithm	MDS
convolve	0.253
correlate	0.220
cumsum	0.139
diff	0.096
normal	0.242
reciprocal	0.031
square	0.089

По оценке DBI, приведенной в таблице 6, алгоритм лучше всего показал себя на необработанных данных и свёртке с оценками 1.9983 и 2.2009 соответственно.

Таблица 6 – Оценка DBI MDS

data\algorithm	MDS
convolve	2.201
correlate	2.611
cumsum	3.139
diff	5.388
normal	1.998
reciprocal	13.744
square	5.584

3.2.4. Оценка анализа главных компонент

Алгоритм PCA показал более-менее приемлемые результаты на данных дискретной свёртки, автокорреляции, кумулятивной суммы и необработанных.

На рисунке 19 продемонстрированы результаты работы алгоритма PCA также на необработанных данных, идентичные результатам MDS. Кластеры АИ-80 (фиолетовый) и АИ-92 (синий) смешанны и очень разреженные, кластер АИ-95 (бирюзовый) близок к ним и более сжатый. А АИ-98 (желтый) сформировал

отдаленный от них сжатый кластер. DBI в данном случае равен 1.699086, а SC 0.360091.

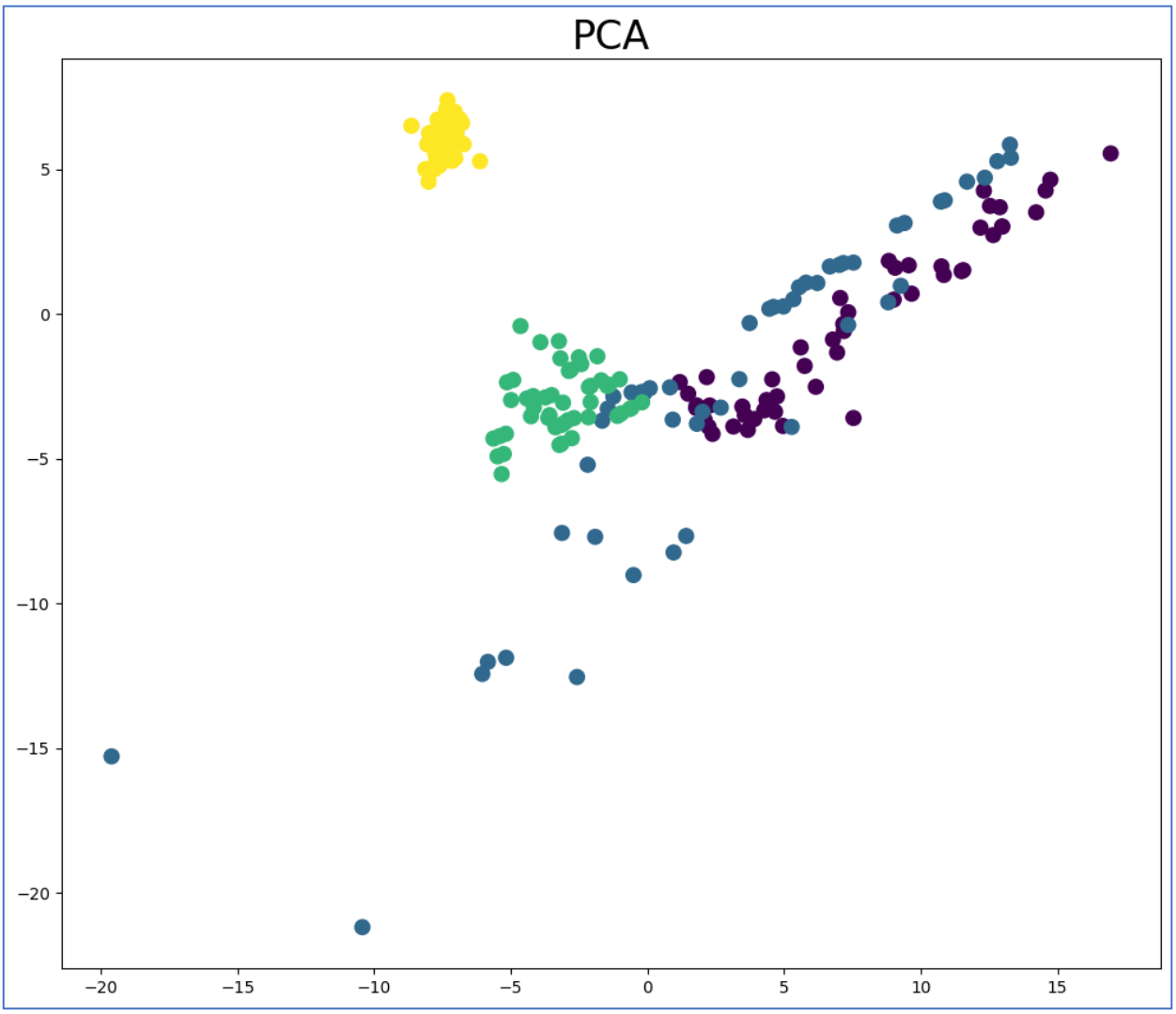


Рисунок 19 – Результаты работы алгоритма снижения размерности PCA

По оценке SC, приведенной в таблице 7, алгоритм в пике показал несколько лучшие результаты, чем MDS. Лучше всего алгоритм показал себя на необработанных данных и свёртке с оценками 0.3600 и 0.3284 соответственно.

Таблица 7 – Оценка SC PCA

data\algorithm	PCA
convolve	0.328
correlate	0.299
cumsum	0.278
diff	0.085
normal	0.360
reciprocal	-0.257
square	0.036

На первый взгляд оценка DBI алгоритма противоречит оценке SC, но на самом деле оценки коррелируют, за исключением результатов на обратных и данных, обработанных дискретной производной, в которых алгоритм дает совершенно неприемлемый результат. В данном случае оценка SC оказалась более достоверной. По оценке DBI, приведенной в таблице 8, алгоритм лучше всего показал себя на необработанных данных с оценкой 1.6990.

Таблица 8 – Оценка DBI PCA

data\algorithm	PCA
convolve	2.401
correlate	2.541
cumsum	3.137
diff	1.743
normal	1.699
reciprocal	2.284
square	5.463

3.2.5. Оценка нейроподобного метода

Нейроподобный метод снижения размерности лучше всего показал себя на необработанных данных и на кумулятивной сумме.

На рисунке 20 продемонстрирован лучший результат работы алгоритма на необработанных данных. С помощью этого алгоритма были получены схожие с предыдущими результаты. Точки АИ-98 (желтый) сформировали плотный отличимый кластер, точки АИ-95 (бирюзовый) сформировали характерный, но не плотный кластер. А точки АИ-80 (фиолетовый) и АИ-92 (синий) смешались друг с другом. DBI в данном случае равен 1.784445, а SC 0.166019.

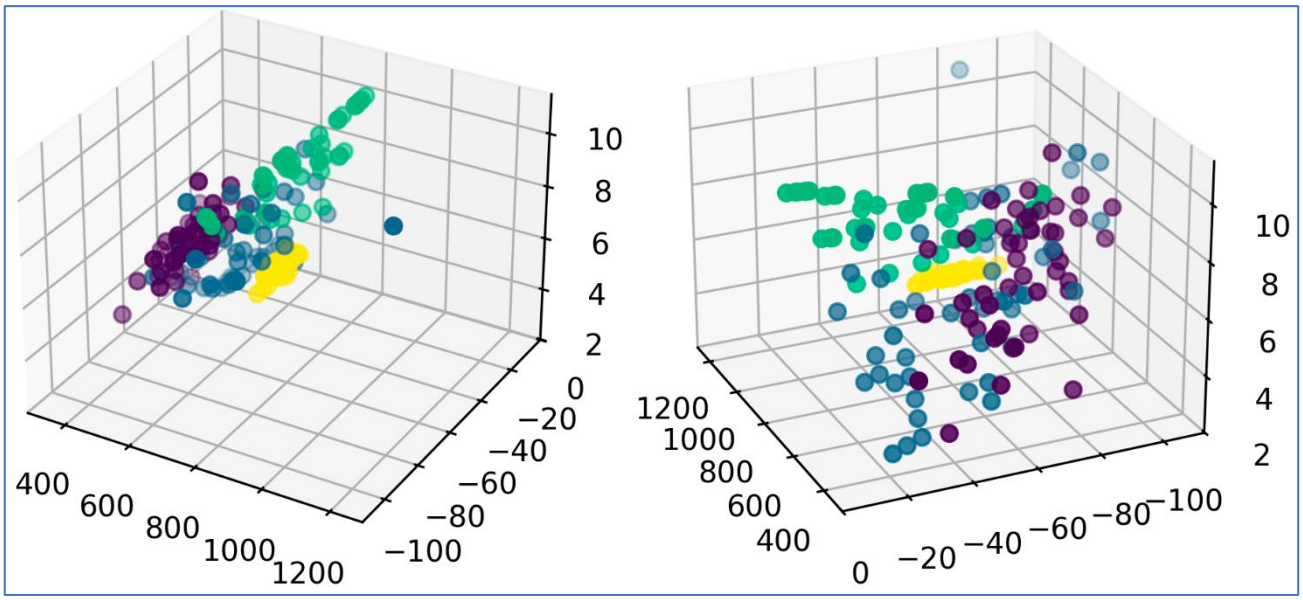


Рисунок 20 – Результат работы нейроподобного метода снижения размерности

По оценке SC, приведенной в таблице 9, алгоритм дал очень плохие результаты во всех случаях, кроме данных свёртки, где оценка составила 0.3126.

Таблица 9 – Оценка SC нейроподобного метода снижения размерности

data\algorithm	**HM**
convolve	0.313
correlate	0.144
cumsum	0.070
diff	0.121
normal	0.166
reciprocal	0.129
square	0.105

Оценка DBI, приведенная в таблице 10, в данном случае оказалась более достоверной, оценив работу алгоритма на необработанных данных как лучший результат с оценкой 1.7844.

Таблица 10 – Оценка DBI нейроподобного метода снижения размерности

data\algorithm	HM
convolve	4.470
correlate	4.480
cumsum	2.864
diff	6.004
normal	1.784
reciprocal	5.883
square	2.547

3.2.6. Оценка спектрального вложения

Алгоритм Spectral Embedding показал по большей части неудовлетворительные результаты, хотя использование некоторых вариантов алгоритма на необработанных данных может дать более-менее применимые результаты.

На рисунке 21 показан результат алгоритма с использованием косинусного сходства. Здесь точки АИ-98 (желтый) сформировали плотный кластер, отдаленный от других, точки АИ-80 (фиолетовый), АИ-92 (синий) и АИ-95 (бирюзовый) сформировали по большей части довольно отличимые, но несколько слившиеся кластеры. Точки АИ-95 сгруппировались более плотно, чем остальные. DBI в данном случае равен 0.913617, а SC 0.47851.

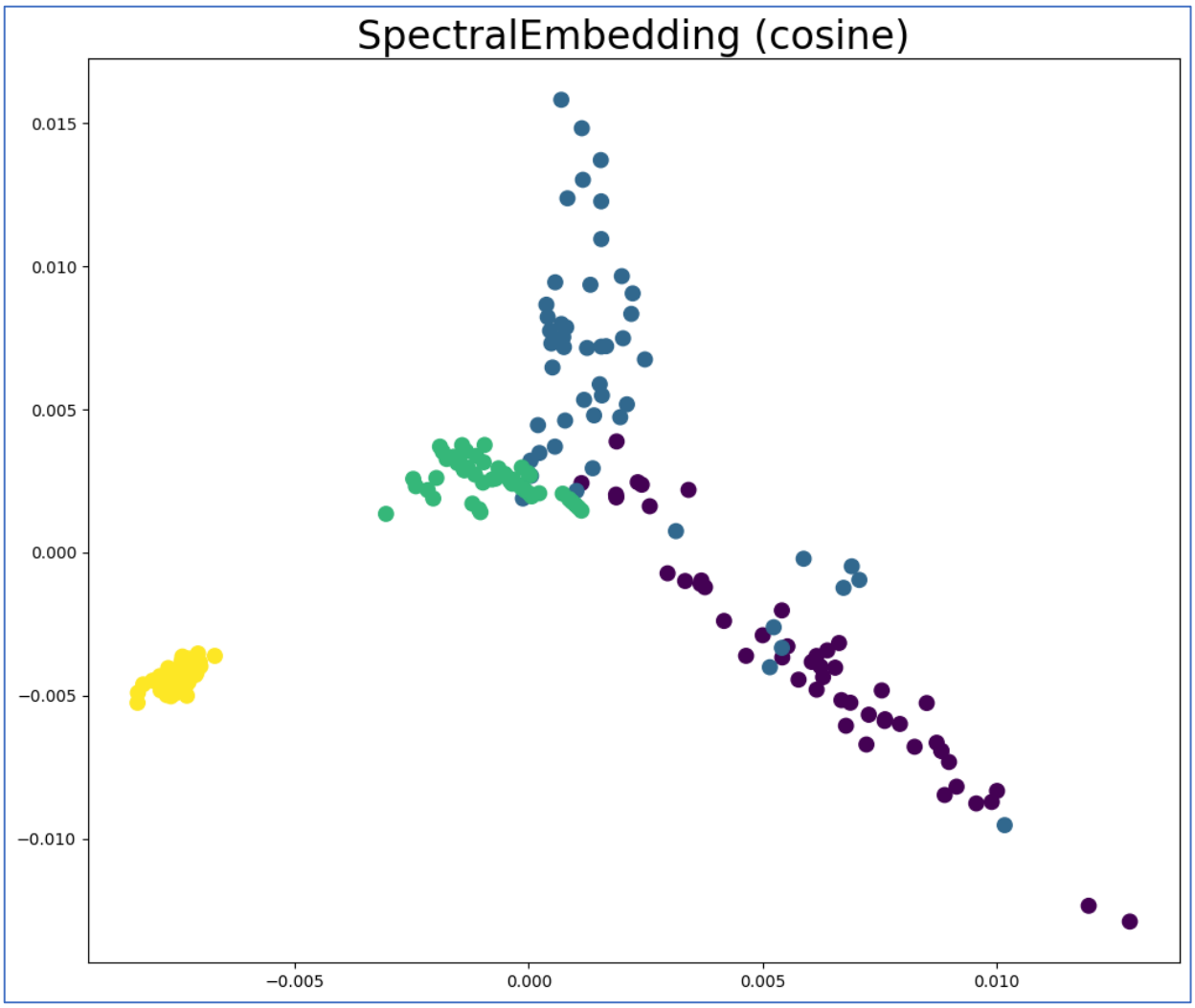


Рисунок 21 – Результаты работы алгоритма снижения размерности Spectral Embedding (косинусное сходство)

Результат использования ядра оператора Лапласа показан на рисунке 22. Здесь кластер из точек АИ-98 (желтый) получился менее плотным, но все так же отдаленным. Точки АИ-80 (фиолетовый), АИ-92 (синий) и АИ-95 (бирюзовый) образовали крупный разреженный кластер, хоть и расположились по большей части в разных его частях. DBI в данном случае равен 1.937176, а SC 0.37783.

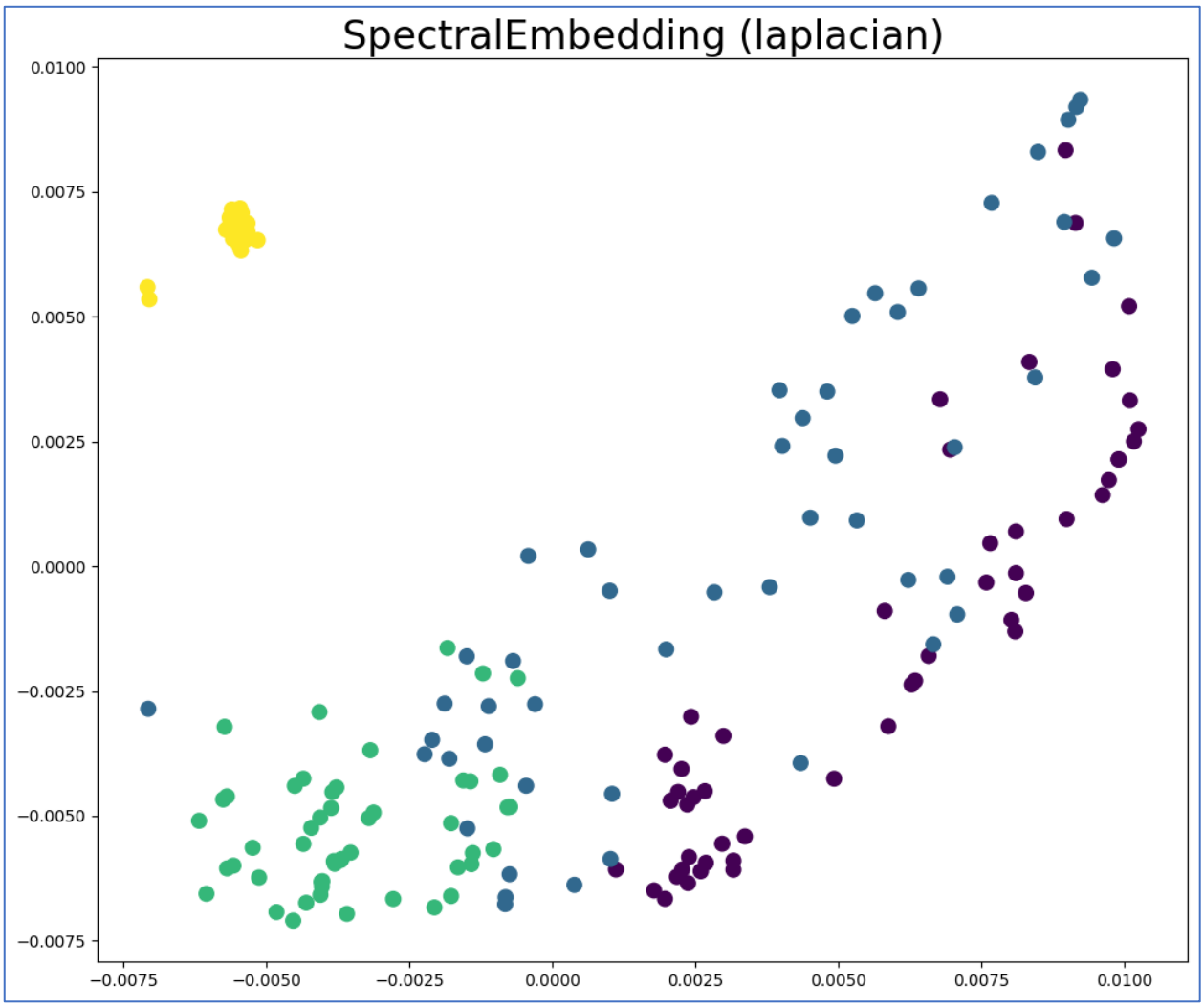


Рисунок 22 – Результаты работы алгоритма снижения размерности Spectral Embedding (ядро оператора Лапласа)

Рисунок 23 демонстрирует результат использования линейного ядра. Здесь результаты схожи с использованием косинусного сходства. Смешанные кластеры здесь получились несколько более отличимыми. DBI в данном случае равен 0.988299, а SC 0.462024.

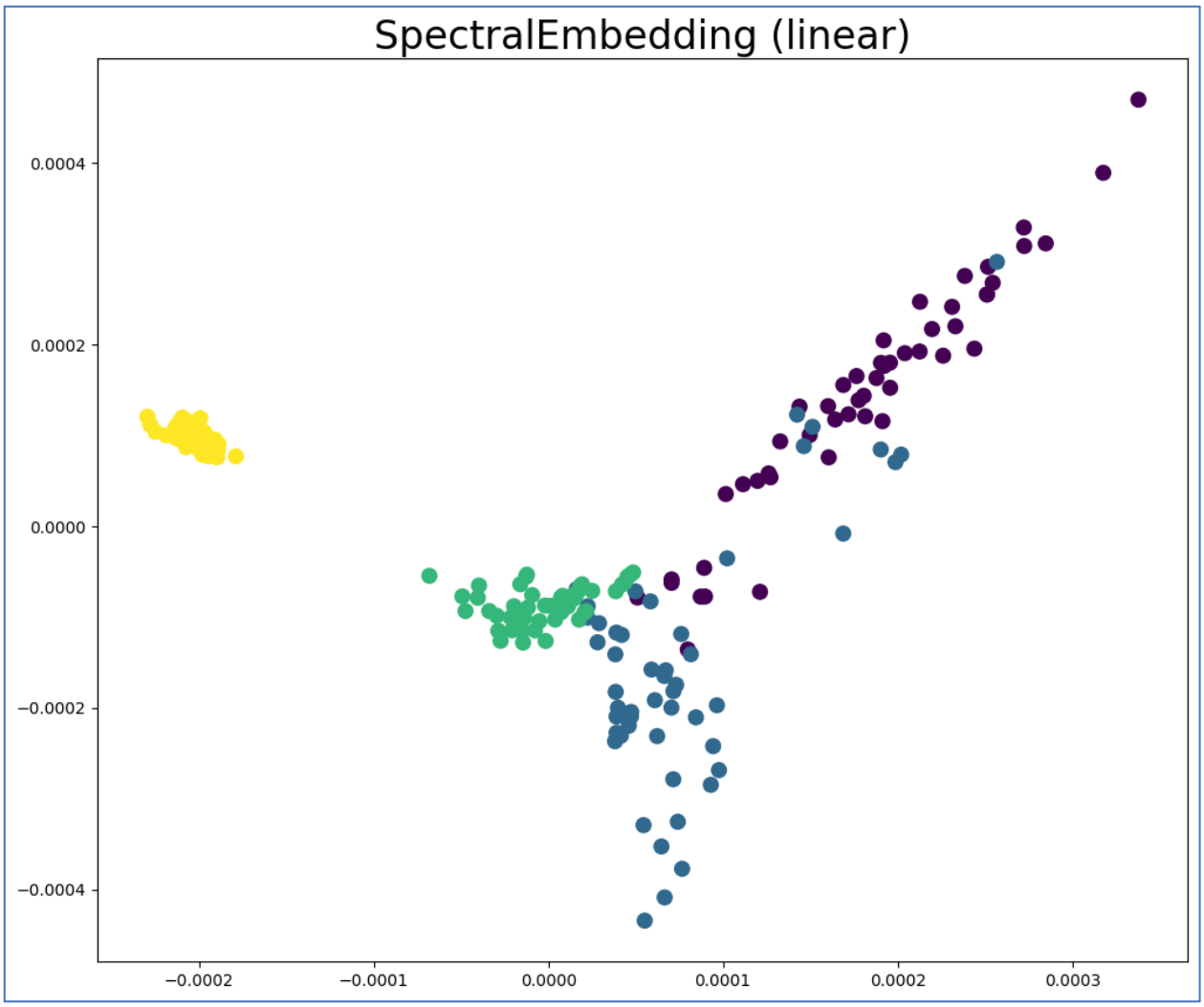


Рисунок 23 – Результаты работы алгоритма снижения размерности Spectral Embedding (линейное ядро)

На рисунке 24 показан результат работы алгоритма с использованием полиномиального ядра. Здесь точки АИ-80 (фиолетовый) и АИ-92 (синий) еще сильнее смешались друг с другом, а точкам АИ-95 (бирюзовый) удалось остаться разреженным, но отличимым кластером. DBI в данном случае равен 1.748541, а SC 0.358198.

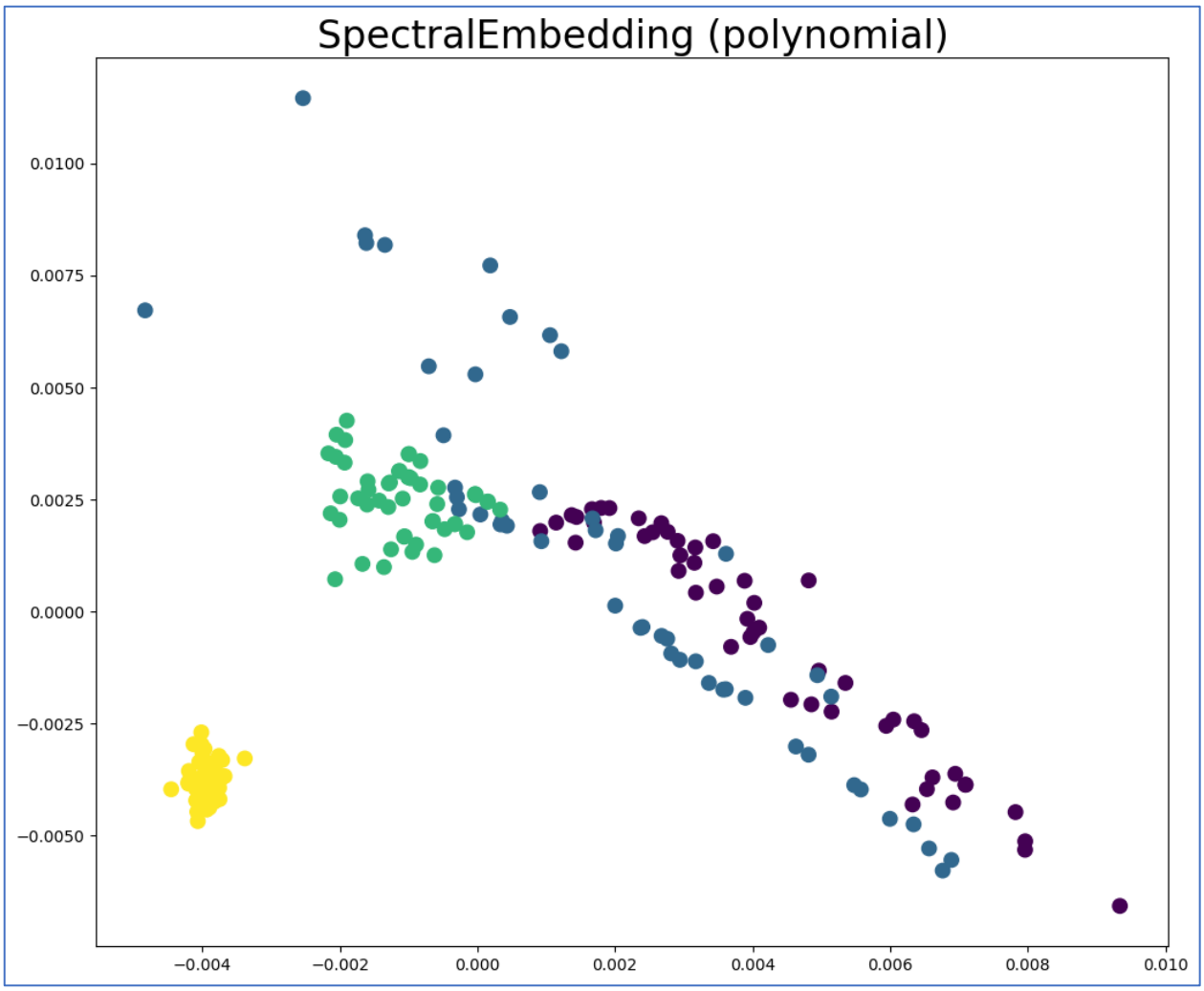


Рисунок 24 – Результаты работы алгоритма снижения размерности Spectral Embedding (полиномиальное ядро)

На рисунке 25 с использованием ядра радиальной базисной функции показан результат едва отличимый от полученного с помощью полиномиального ядра. DBI в данном случае равен 1.900452, а SC 0.360718.

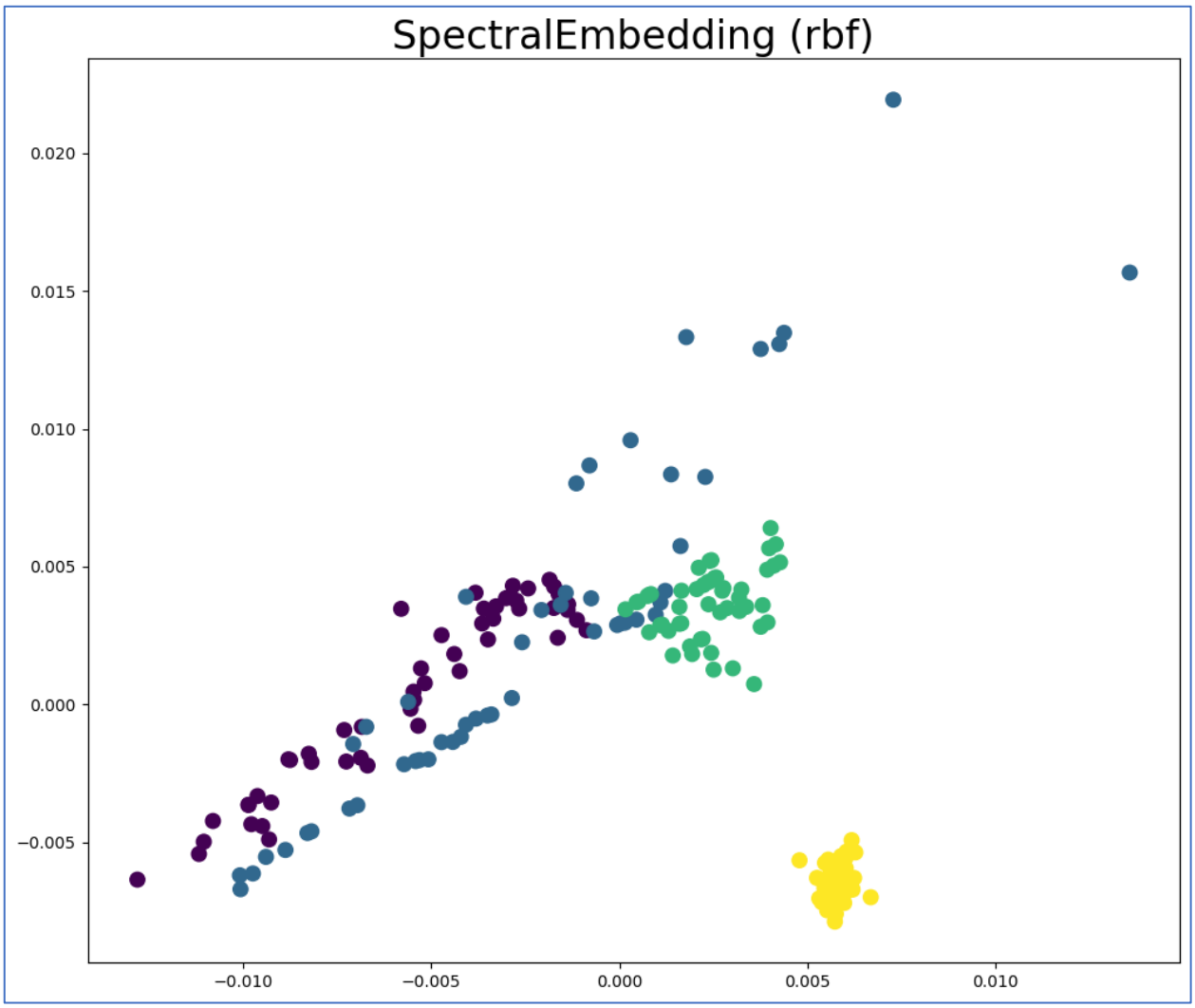


Рисунок 25 – Результаты работы алгоритма снижения размерности Spectral Embedding (ядро радиальной базисной функции)

Использование сигмоидального ядра дало похожий результат, но точки АИ-80 (фиолетовый) и АИ-92 (синий) расположились еще более разреженно, результат показан на рисунке 26. DBI в данном случае равен 2.184099, а SC 0.371953.

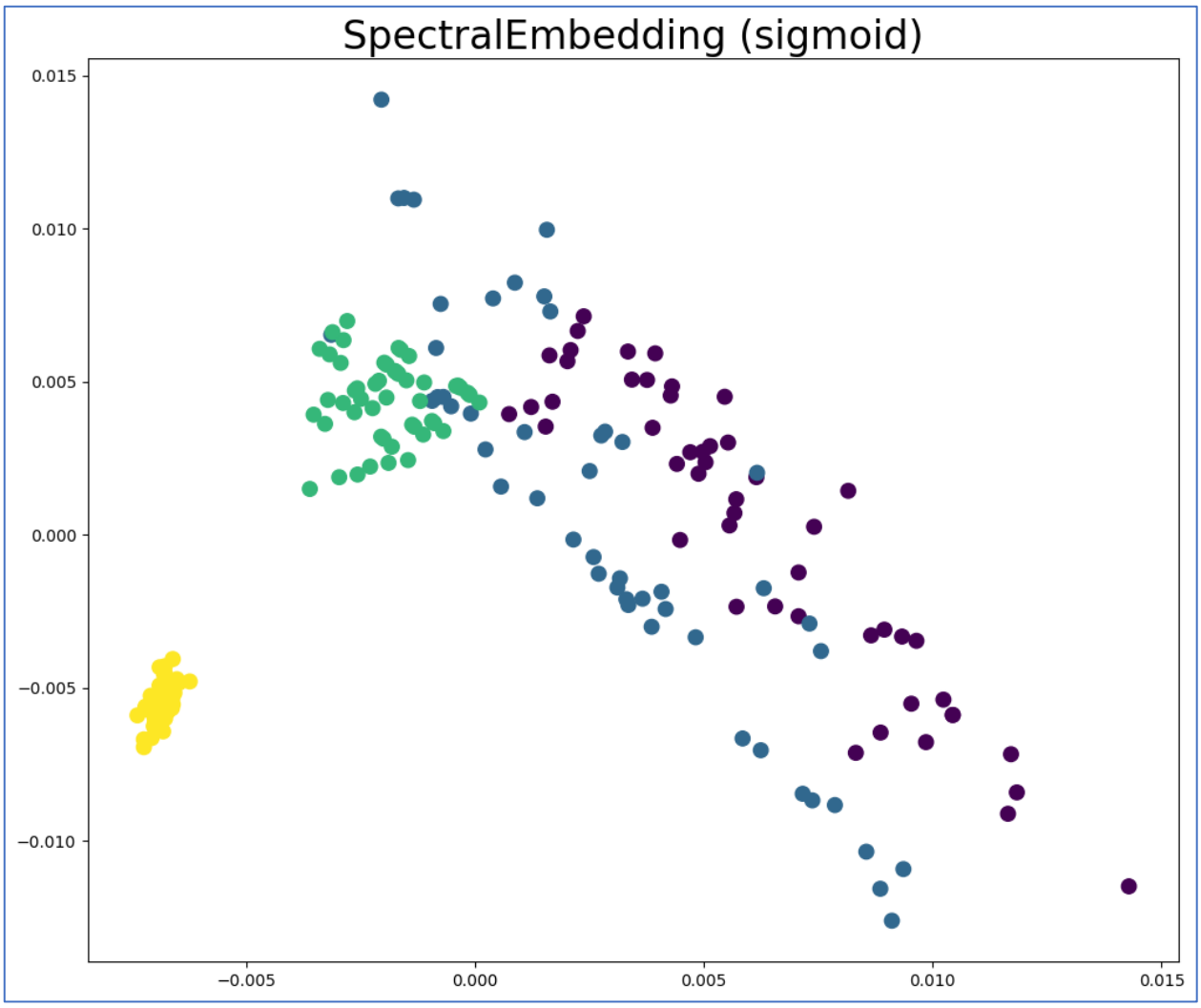


Рисунок 26 – Результаты работы алгоритма снижения размерности Spectral Embedding (сигмоидальное ядро)

По оценке SC, приведенной в таблице 11, алгоритм действительно показал лучшие результаты на необработанных данных. Наибольшие оценки получили результаты алгоритма с использованием косинусного сходства и линейного ядра – 0.4785 и 0.4620 соответственно.

Таблица 11 – Оценка SC SE

data\metric	cosine	laplacian	linear	polynomial	rbf	sigmoid
convolve	0.323	0.259	0.304	—	0.010	-0.034
correlate	0.212	0.263	0.232	—	0.142	-0.034
cumsum	0.266	0.246	0.178	—	0.260	-0.034
diff	—	0.221	—	0.063	0.068	0.067
normal	0.479	0.378	0.462	0.358	0.361	0.372
reciprocal	—	-0.238	—	—	-0.017	—
square	0.082	0.330	0.165	-0.153	0.030	0.167

Оценка DBI, приведенная в таблице 12, показывает схожие оценки: варианты алгоритма с косинусным сходством и линейным ядром оценены в 0.9136 и 0.9882 соответственно. Также DBI ложно хорошо оценивает вариант алгоритма с сигмоидальным ядром на квадратных данных.

Таблица 12 – Оценка DBI SE

data\metric	cosine	laplacian	linear	polynomial	rbf	sigmoid
convolve	1.877	2.143	1.817	—	1.913	12.018
correlate	2.676	2.112	1.990	—	1.921	12.018
cumsum	4.103	1.960	3.748	—	2.628	12.018
diff	—	2.033	—	1.738	1.790	1.749
normal	0.914	1.937	0.988	1.749	1.900	2.184
reciprocal	—	2.034	—	—	1.949	—
square	1.667	2.361	1.679	2.998	3.847	1.420

3.2.7. Оценка т-распределенного стохастического вложения соседей

Алгоритм t-SNE позволил получить хорошие результаты для всех типов обработанных данных, кроме квадратного и обратного преобразования, но даже на этих данных результаты были приемлемыми. На рисунке 27 показан лучший результат, полученный на данных, обработанных свёрткой, при использовании алгоритма с корреляционной метрикой. Видно, что кластеры получились разреженными, но хорошо отличимыми и удаленными друг от друга. Точки АИ-98 (желтый) сильнее

удалены от остальных кластеров. Среди образцов АИ-80 (фиолетовый) и АИ-92 (синий) имеются по одному выбросу в другие кластеры, которые, очевидно, будут единственным препятствием для алгоритмов кластеризации для достижения идеального результата. DBI в данном случае равен 0.447688, а SC 0.62724.

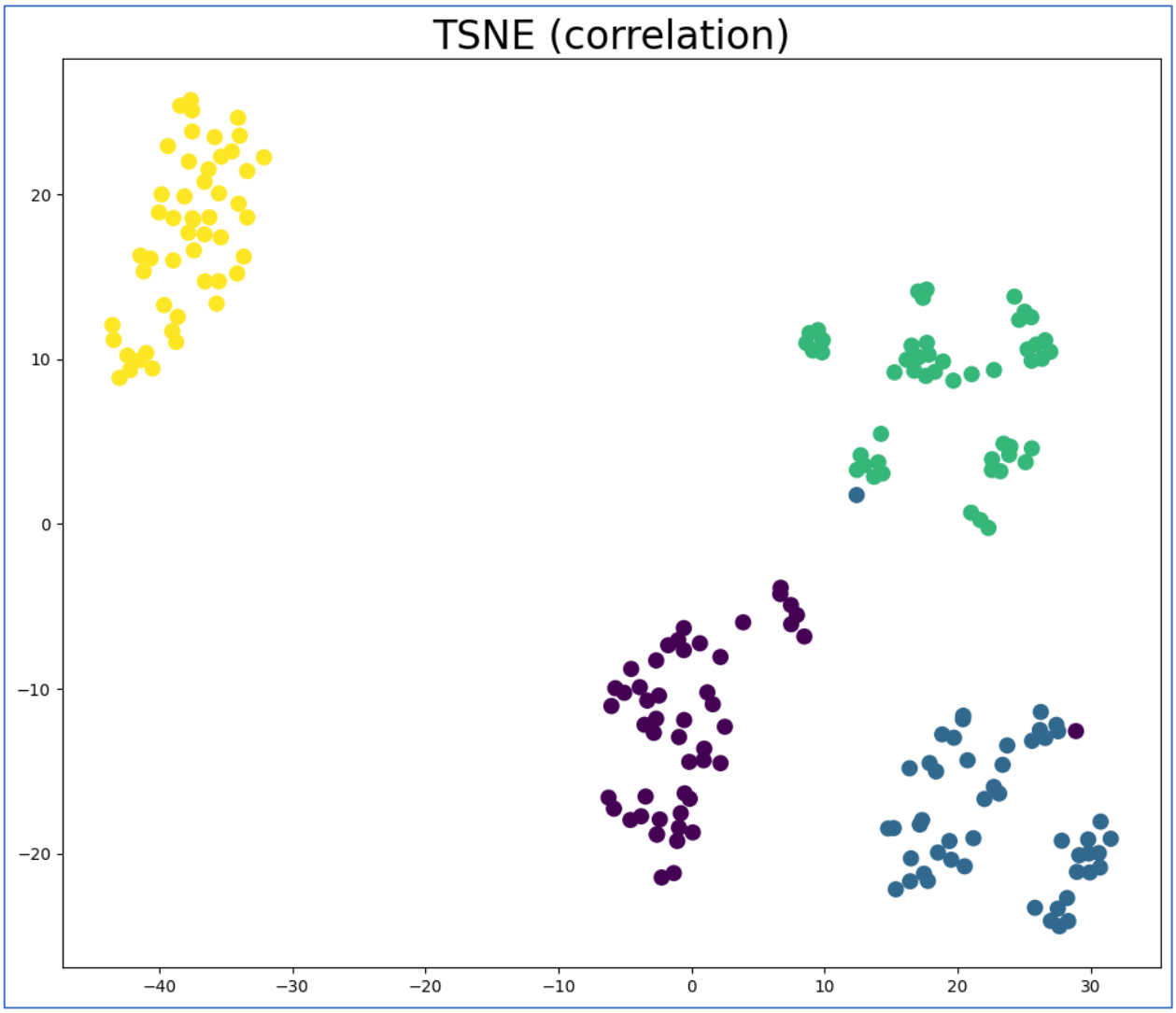


Рисунок 27 – Результаты работы алгоритма снижения размерности t-SNE для данных свёртки (корреляционная метрика)

Следующие результаты показаны на рисунке 28, где применена та же самая корреляционная метрика к данным, обработанным дискретной производной. Здесь образцы АИ-98 (желтый) образовали различимый, но очень разреженный кластер.

Остальные кластеры также получились довольно разреженными и разбитыми на несколько частей, перекрывающих друг друга. DBI в данном случае равен 1.210241, а SC 0.351844.

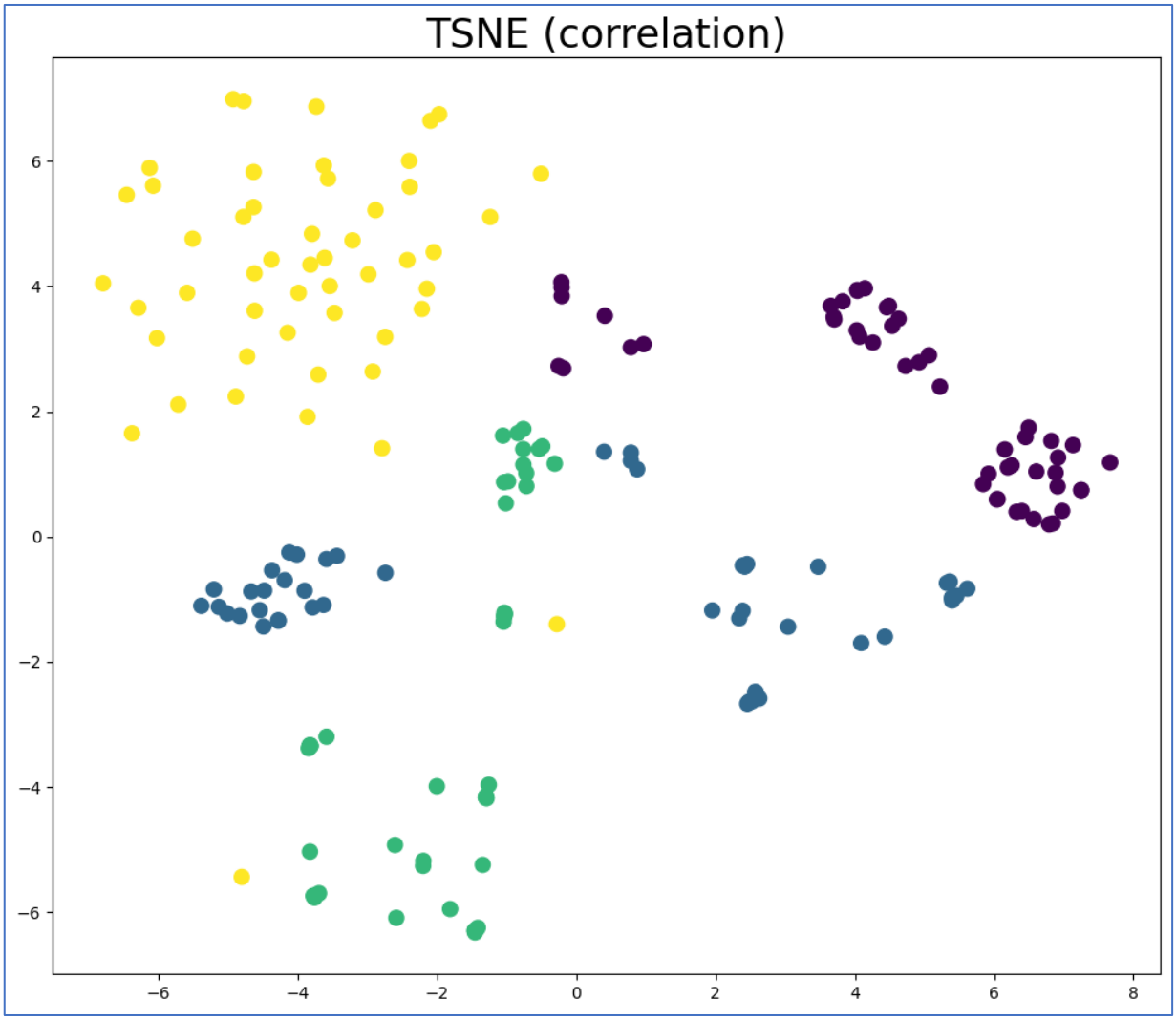


Рисунок 28 – Результаты работы алгоритма снижения размерности t-SNE для данных, обработанных дискретной производной (корреляционная метрика)

В следующем примере на рисунке 29 видно, что, при использовании метрики Канберры для данных, обработанных дискретной производной, точки довольно явно разбились на кластеры, причем равноудаленные и с примерно одинаковой плотностью. Здесь также имеются единичные выбросы в «чужие» кластеры, но у

образцов АИ-80 (фиолетовый) и АИ-98 (желтый), а образцы АИ-95 (бирюзовый) разбились на 2 кластера, что может вызвать проблемы у кластеризации без известного количества классов. DBI в данном случае равен 0.571966, а SC 0.629415.

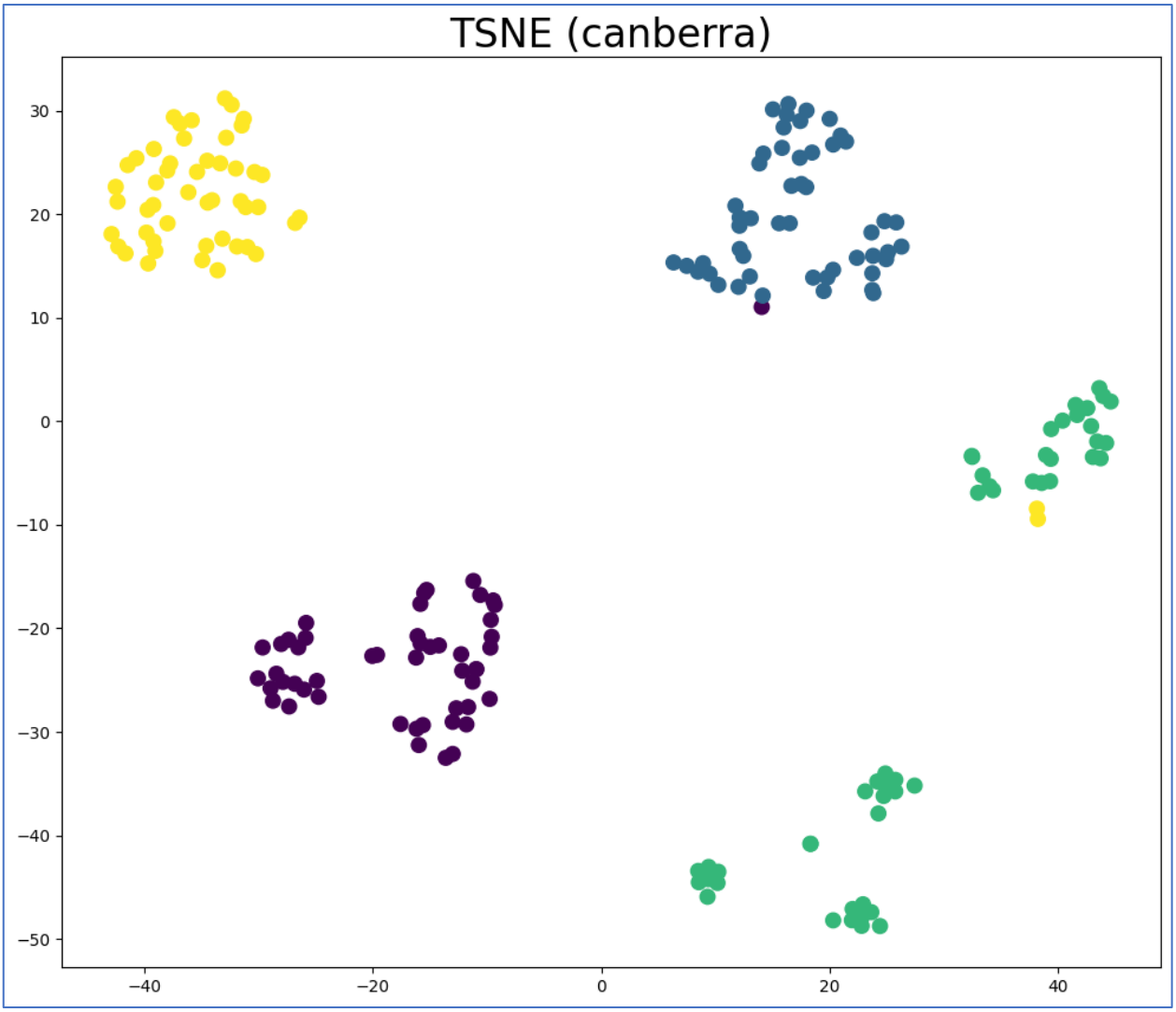


Рисунок 29 – Результаты работы алгоритма снижения размерности t-SNE для данных, обработанных дискретной производной (расстояние Канберры)

Необычные результаты получились на необработанных данных с применением косинусной метрики, продемонстрированные на рисунке 30. Здесь класс АИ-98 (желтый) сформировал отчетливый удаленный от остальных кластер. Рядом расположились кластеры АИ-95 (бирюзовый) и АИ-92 (синий), которые получились

разреженными и несколько слились друг с другом. Далее расположился кластер АИ-80 (фиолетовый), к которому примкнули несколько точек АИ-92, этот кластер разбился надвое и также имеет несколько «потерявшихся» точек. DBI в данном случае равен 0.908416, а SC 0.335447.

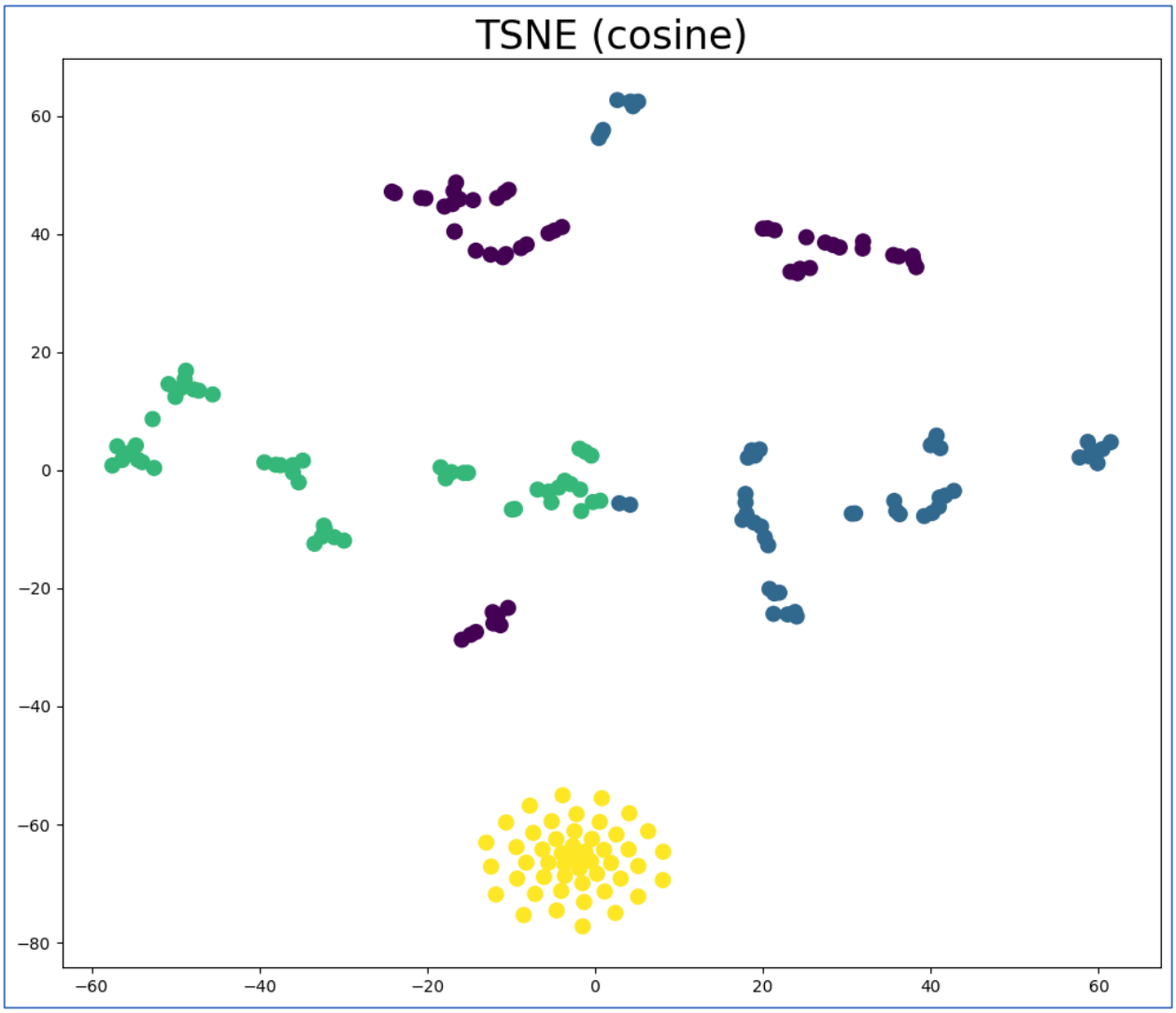


Рисунок 30 – Результаты работы алгоритма снижения размерности t-SNE для необработанных данных (косинусное расстояние)

Важно отметить, что результаты работы алгоритма сильно зависят как от метода предварительной обработки данных, так и от выбранной метрики расстояния. Только правильная совокупность этих параметров даст хороший результат.

По оценке SC, приведенной в таблице 13, алгоритм t-SNE даже в худшей своей половине результатов в среднем получает лучшие оценки, чем большинство других алгоритмов в лучшем случае (>0.3). Хорошие оценки получили около четверти всех результатов, лучшую оценку получила комбинация метрики Канберры с данными, обработанных дискретной производной – 0.6294.

Таблица 13 – Оценка SC t-SNE

data\metric	braycurtis	canberra	chebyshev	correlation	cosine	euclidean	hamming	manhattan
convolve	0.338	0.327	0.368	0.627	0.570	0.342	0.045	0.339
correlate	0.325	0.274	0.284	0.551	0.486	0.324	0.065	0.310
cumsum	0.222	0.252	0.280	0.546	0.450	0.261	0.060	0.264
diff	0.500	0.629	0.165	0.373	0.356	0.295	0.331	0.354
normal	0.617	0.542	0.416	0.606	0.571	0.490	0.364	0.552
reciprocal	0.344	0.489	0.106	0.265	0.241	0.194	0.381	0.334
square	0.403	0.486	0.361	0.369	0.362	0.287	0.132	0.384

Оценка DBI, приведенная в таблице 14, не противоречит оценке SC, но более «щедро» оценила работу алгоритма, здесь более трети вариантов получили хорошие оценки. Лучше всего по оценке DBI показала себя комбинация корреляционной метрики на данных свёртки с оценкой 0.4262.

Таблица 14 – Оценка DBI t-SNE

data\metric	braycurtis	canberra	chebyshev	correlation	cosine	euclidean	hamming	manhattan
convolve	1.861	1.584	2.121	0.426	0.508	2.353	2.351	1.993
correlate	2.858	2.067	2.155	0.652	0.758	2.577	3.155	3.058
cumsum	2.774	2.313	3.145	0.601	0.753	2.771	7.333	2.772
diff	0.829	0.572	1.235	0.903	0.914	0.882	1.314	1.069
normal	0.841	0.859	0.885	0.606	0.571	0.744	1.436	0.806
reciprocal	1.533	0.867	2.176	2.624	2.160	2.551	1.136	1.268
square	1.090	1.100	0.972	0.778	0.841	1.136	2.973	1.081

3.2.8. Оценка аппроксимации и проекции однородного многообразия

Алгоритм UMAP дает результаты схожие с теми, что получились при использовании алгоритма t-SNE. В среднем алгоритм дает несколько худшие результаты при лучшем подборе параметров, но в некоторых случаях достигает лучших результатов, и также имеет лучшие показатели при не лучшем подборе параметров, что позволяет исследователю легче добиваться хороших результатов. На рисунке 31 показан лучший результат алгоритма, полученный на данных после свёртки при использовании корреляционной метрики. На рисунке видно, что группа спектров АИ-98 (желтый) образовала сильно отдаленный от остальных плотный кластер, а остальные три группа образовали хоть и близкие друг к другу, но также плотные и различимые кластеры. Две точки группы АИ-80 (фиолетовый) попали в другие кластеры. DBI в данном случае равен 0.469292, а SC 0.625737.

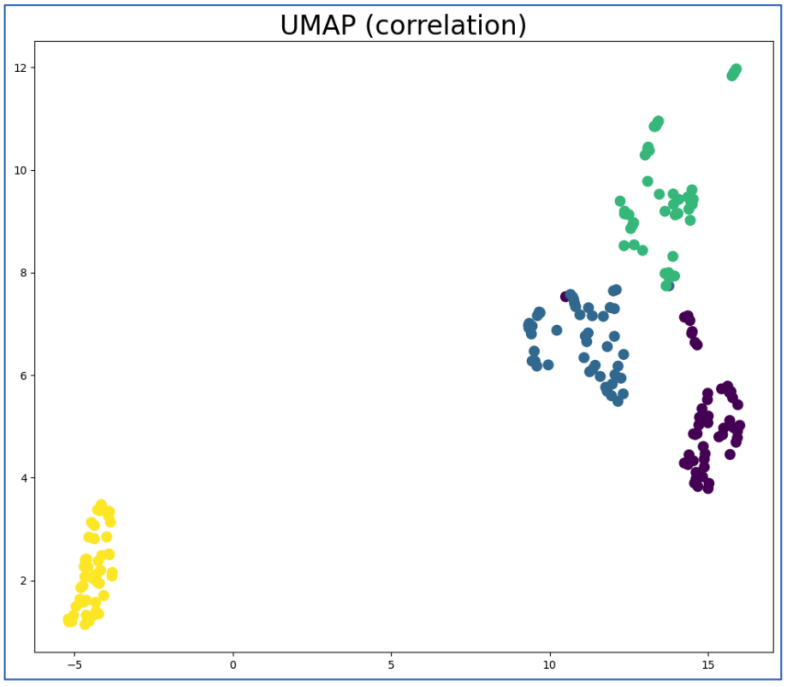


Рисунок 31 – Результаты работы алгоритма снижения размерности UMAP для данных свёртки (корреляционная метрика)

На рисунке 32 показаны результаты с применением той же корреляционной метрики на данных, обработанных дискретной производной. В данном случае только образцы АИ-98 (желтый) сформировали плотный кластер. Образцы АИ-92 (синий) также сформировали отличимый, но разреженный и смешанный с другими кластер. Точки АИ-95 (бирюзовый) разбились на два кластера, расположенных по разные стороны от кластера АИ-98. А образцы АИ-80 (фиолетовый) разбились на 3 части, одна из которых смешана с тремя другими классами, а две остальные получились плотными, но являются самыми удаленными кластерами друг от друга. DBI в данном случае равен 2.913638, а SC 0.266667.

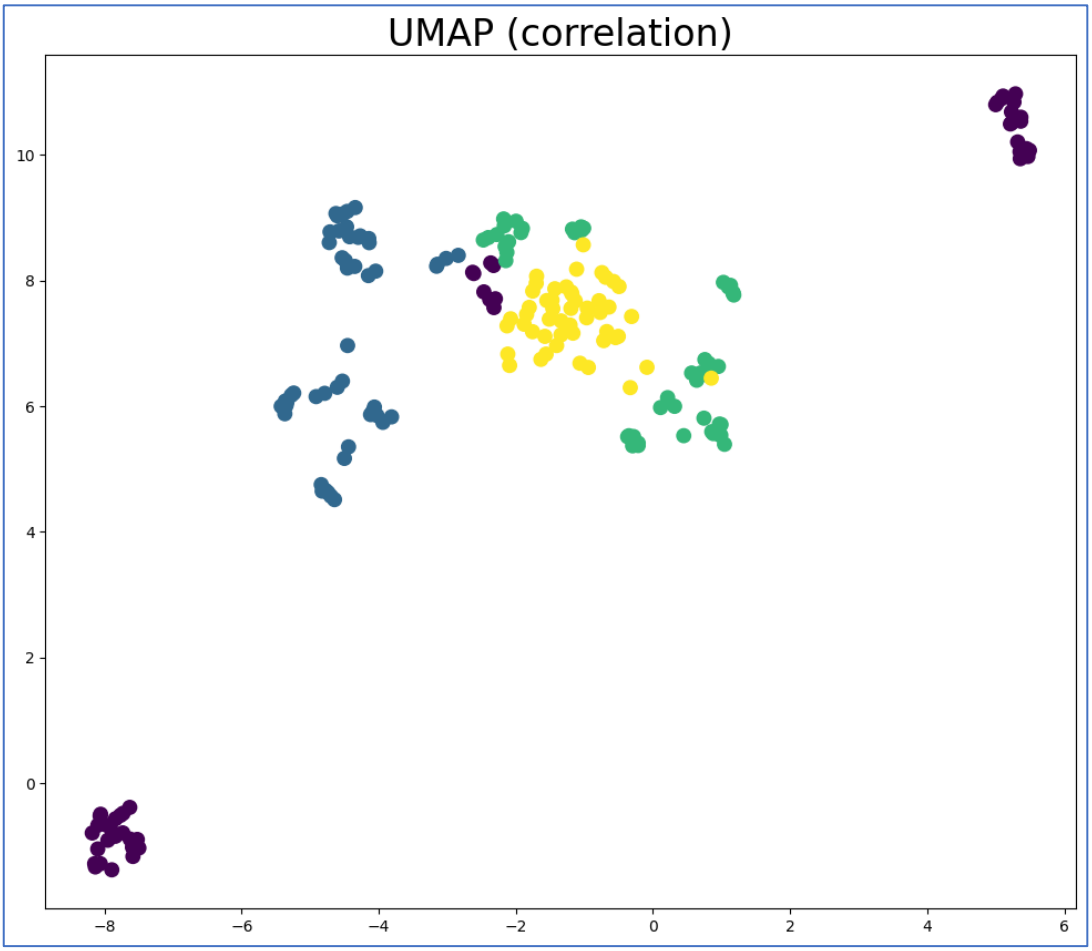


Рисунок 32 – Результаты работы алгоритма снижения размерности UMAP для данных, обработанных дискретной производной (корреляционная метрика)

Проверим результаты работы алгоритма на тех же данных, но с использованием метрики Канберры, показанные на рисунке 33. В этом случае классы АИ-80 (фиолетовый), АИ-92 (синий) и АИ-98 (желтый) сформировали очень плотные и удаленные друг от друга кластеры, а группа АИ-95 (бирюзовый) разбилась на 3 кластера, но отличимых от остальных. Лишь 2 образца АИ-98 и один АИ-92 попали в «чужие» кластеры. DBI в данном случае равен 0.372888, а SC 0.734253.

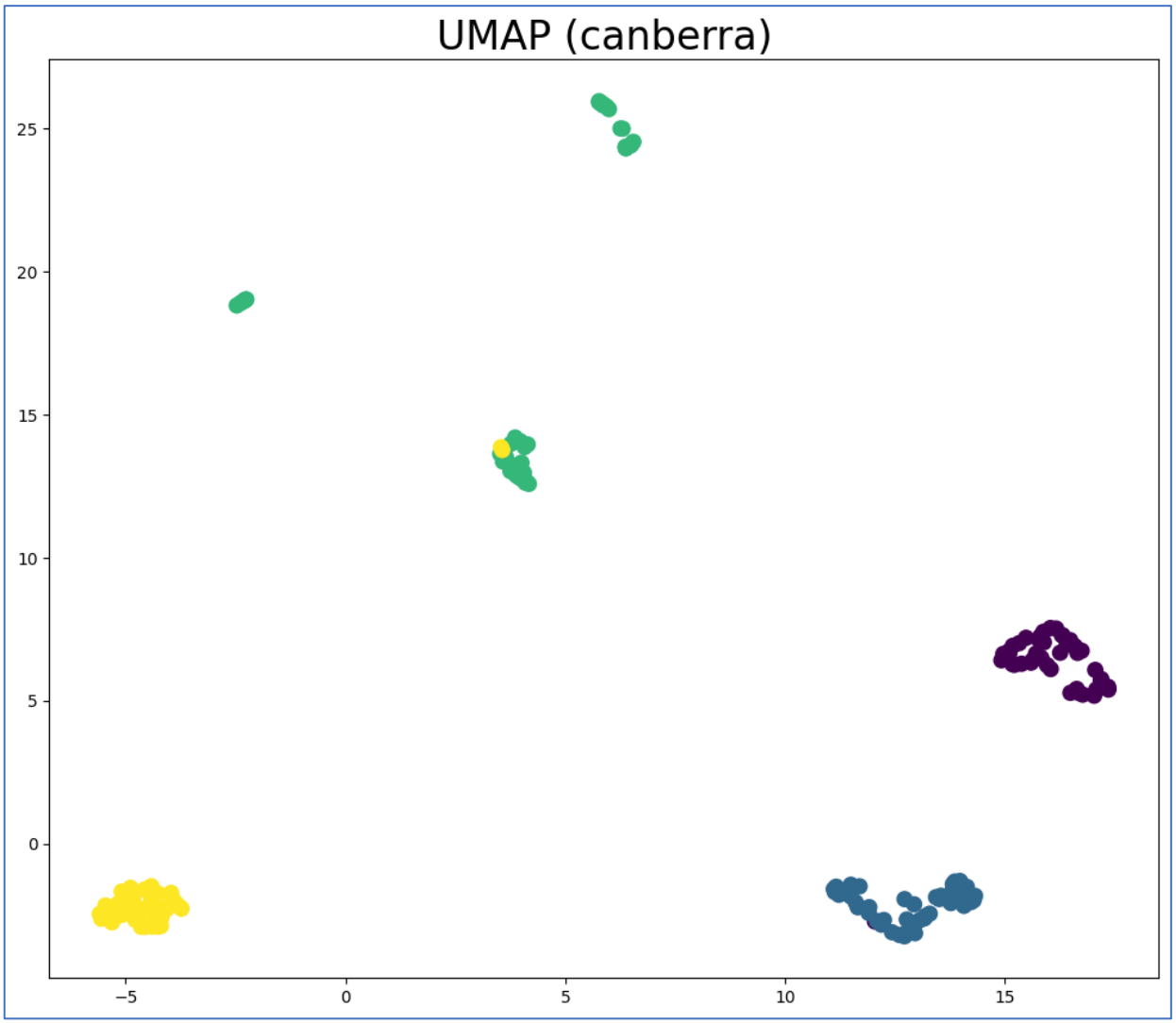


Рисунок 33 – Результаты работы алгоритма снижения размерности UMAP для данных, обработанных дискретной производной (метрика Канберры)

Рассмотрим также результат работы алгоритма с применением косинусной метрики на необработанных данных, аналогично рассмотренному с алгоритмом t-SNE. На рисунке 34 виден похожий результат, где класс АИ-98 (желтый) сформировал отдаленный сжатый кластер, но в отличие от результатов t-SNE, кластеры АИ-92 (синий) и АИ-95 (бирюзовый) получились более сжатыми и отдаленными друг от друга. В остальном результаты схожие. DBI в данном случае равен 0.792872, а SC 0.45263.

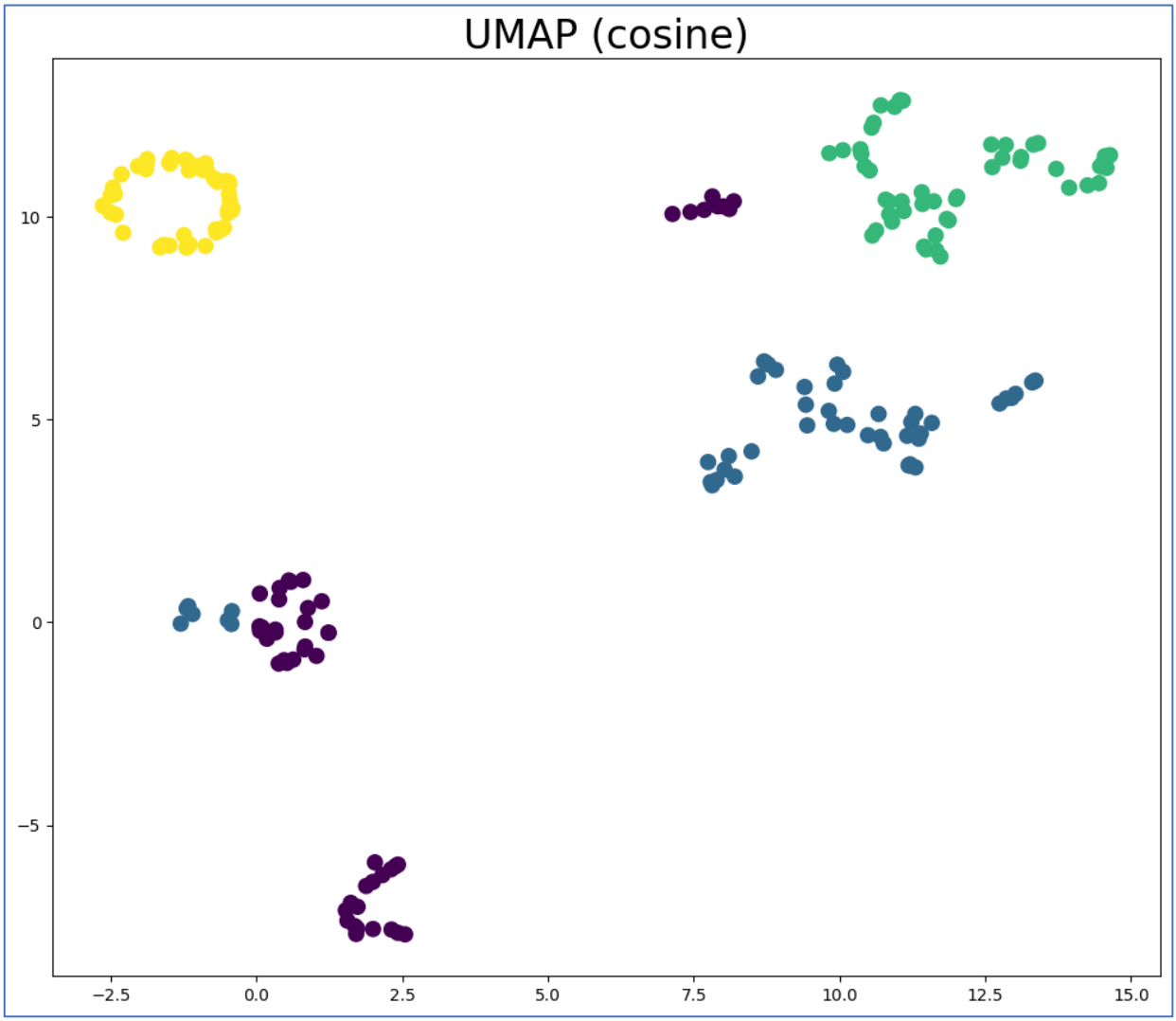


Рисунок 34 – Результаты работы алгоритма снижения размерности UMAP для необработанных данных (косинусная метрика)

Алгоритм UMAP по оценке SC, приведенной в таблице 15, является лучшим среди рассмотренных как в среднем, так и в лучшем случае. Наивысшую оценку также получила комбинация метрики Канберры и данных, обработанных дискретной производной, – 0.7342. Это самая высокая оценка, полученная при тестировании алгоритмов. Таким образом, идеальная оценка метрики SC, единица, является недостижимой.

Таблица 15 – Оценка SC UMAP

data\metric	braycurtis	canberra	chebyshev	correlation	cosine	euclidean	hamming	manhattan
convolve	0.352	0.333	0.362	0.667	0.602	0.422	-0.051	0.410
correlate	0.347	0.336	0.301	0.585	0.501	0.348	-0.007	0.304
cumsum	0.275	0.303	0.307	0.545	0.441	0.288	0.217	0.270
diff	0.628	0.734	0.131	0.311	0.341	0.281	0.338	0.426
normal	0.591	0.554	0.491	0.611	0.633	0.560	0.310	0.538
reciprocal	0.437	0.548	0.141	0.313	0.267	0.258	0.271	0.499
square	0.315	0.566	0.351	0.320	0.336	0.280	0.227	0.346

Оценка DBI, приведенная в таблице 16, также согласуется с оценкой SC. Применение алгоритма с метрикой Канберры на данных, обработанных дискретной производной, получило лучшую оценку в 0.3728. Эта оценка также является лучшей, за исключением нескольких аномальных результатов.

Таблица 16 – Оценка DBI UMAP

data\metric	braycurtis	canberra	chebyshev	correlation	cosine	euclidean	hamming	manhattan
convolve	2.069	1.497	2.171	0.469	0.553	2.262	4.563	2.375
correlate	2.807	2.179	3.078	0.657	0.785	2.684	4.769	3.186
cumsum	2.662	1.606	3.222	0.562	0.824	2.550	2.044	2.694
diff	0.804	0.373	1.385	1.519	1.327	0.988	1.041	0.851
normal	0.805	0.870	0.918	0.552	0.619	0.639	1.126	0.751
reciprocal	1.332	0.898	1.862	1.593	1.485	2.881	1.015	1.010
square	1.426	0.959	1.175	0.983	0.881	1.419	1.829	1.383

3.2.9. Анализ оценок алгоритмов снижения размерности

В большинстве случаев наблюдается корреляция между показателями SC и DBI, что подтверждает достоверность оценки качества снижения размерности. Однако следует учитывать, что некоторые методы дают противоречивые оценки, что требует комплексного анализа и возможного применения дополнительных метрик или методов проверки.

Среди рассмотренных методов UMAP и t-SNE показали наилучшие результаты по оценке SC, что указывает на их высокую применимость для задач визуализации высокоразмерных данных. Классические методы, такие как MDS и PCA, хоть и дают приемлемые результаты, часто страдают от слияния кластеров и недостаточной дистанции между кластерами. Лучшие результаты алгоритмов приведены в таблицах 17 и 18.

Таблица 17 – Оценка SC алгоритмов снижения размерности

\Алгоритм Данные	**ISOMAP**	**LLE**	**MDS**	**PCA**	**HM**	**SpectralEmbedding**	**TSNE**	**UMAP**
convolve24	0.610	0.623	0.253	0.329	0.313	0.323	0.627	0.667
correlate24	0.454	0.485	0.220	0.299	0.144	0.263	0.551	0.585
cumsum24	0.519	0.349	0.139	0.278	0.070	0.266	0.546	0.545
diff24	0.602	0.422	0.096	0.085	0.121	0.221	0.629	0.734
normal24	0.631	0.615	0.242	0.360	0.166	0.479	0.617	0.633
reciprocal24	0.510	0.295	0.031	-0.257	0.129	-0.017	0.489	0.548
square24	0.495	0.380	0.089	0.036	0.105	0.330	0.486	0.566

Таблица 18 – Оценка DBI алгоритмов снижения размерности

\Алгоритм Данные	**ISOMAP**	**LLE**	**MDS**	**PCA**	**HM**	**SpectralEmbedding**	**TSNE**	**UMAP**
convolve24	0.506	0.375	2.201	2.401	4.470	1.817	0.426	0.469
correlate24	0.899	0.862	2.612	2.541	4.481	1.921	0.652	0.657
cumsum24	0.717	1.579	3.139	3.137	2.864	1.960	0.601	0.562
diff24	0.591	0.716	5.388	1.743	6.004	1.738	0.572	0.373
normal24	0.603	0.490	1.998	1.699	1.784	0.914	0.571	0.552
reciprocal24	1.516	1.813	13.744	2.285	5.883	1.949	0.867	0.898
square24	1.094	1.033	5.584	5.463	2.547	1.420	0.778	0.881

Результаты показывают, что предварительная обработка данных существенно влияет на результаты снижения размерности. Необработанные данные и данные, полученные с использованием свёртки, зачастую дают лучшие результаты, тогда как квадратные преобразования и дискретная производная могут как улучшать, так и ухудшать показатели в зависимости от алгоритма.

При этом выбор метрики расстояния самого алгоритма снижения является не менее важным параметром. При использовании таких метрик, как корреляционная, косинусная или Канберры, оценка качества потенциальной кластеризации заметно улучшалась. Особенно ярко эта тенденция прослеживалась для алгоритмов t-SNE и UMAP, где сочетание определённых метрик с конкретными типами предварительной обработки (например, Канберры с данными, обработанными дискретной производной) давало лучшие показатели как по SC, так и по DBI. Это подчёркивает необходимость комплексного подбора параметров для достижения оптимальной визуализации и разделения данных [60].

Дальнейшие исследования могут быть направлены на разработку адаптивных методов предварительной обработки и автоматизированного подбора метрик расстояния для оптимизации работы алгоритмов снижения размерности. Кроме того, интеграция нескольких методов в гибридные подходы может способствовать получению более устойчивых и интерпретируемых результатов.

3.3. Анализ кластеризации

Далее будет проведена оценка результатов работы алгоритмов кластеризации на данных, обработанных с использованием алгоритмов снижения размерности. Оценка

результатов будет проведена с помощью скорректированного индекса Рэнда. Скорректированный индекс Рэнда (Adjusted Rand Index, ARI) – это статистическая мера, используемая для оценки сходства между двумя кластеризациями данных, в данном случае, между результатами алгоритма кластеризации и истинными метками [89; 114]. В отличие от стандартного индекса Рэнда, который измеряет процент согласованных пар объектов (как попавших в один кластер в обоих разбиениях, так и разделенных в обоих), ARI вводит поправку на случайное совпадение кластеров. Эта корректировка делает метрику более надежной, устраняя систематическую ошибку, возникающую при сравнении разбиений с разным количеством кластеров. ARI особенно ценен в задачах валидации кластеризации, где требуется объективно оценить качество алгоритма без априорных предположений о структуре данных.

Метод расчета ARI основывается на анализе таблицы сопряженности, элементы которой отражают количество объектов, общих для конкретных пар кластеров в двух разбиениях. Индекс вычисляется путем сравнения наблюдаемого количества согласованных пар с математическим ожиданием при условии случайного распределения кластеров, нормализуя результат в диапазоне от −0.5 до 1. Значение 1 указывает на полное совпадение назначенных меток с истинными, 0 – на уровень согласия, ожидаемый при случайном распределении назначений, а отрицательные значения – на худшее качество по сравнению со случайным угадыванием. ARI широко применяется в биоинформатике, компьютерном зрении и социальных науках благодаря своей устойчивости к дисбалансу кластеров и способности корректно интерпретировать результаты даже при отсутствии эталонных меток. Его независимость от абсолютных значений и акцент на относительное сходство делают ARI универсальным инструментом для сравнения алгоритмов кластеризации в разнородных условиях.

3.3.1. Оценка сдвига среднего значения

Алгоритм MeanShift показал не очень хорошие результаты при кластеризации данных, обработанных алгоритмами снижения размерности ISOMAP, Locally Linear Embedding, t-SNE и UMAP с оценкой ARI от 0.6854 до 0.7442.

Лучший результат алгоритма, полученный на данных, обработанных дискретной производной, к которым был применен алгоритм снижения размерности UMAP с метрикой Брея-Кёртиса изображен на рисунке 35. Видно, как часть точек, принадлежащих кластеру АИ-95 (бирюзовый), которые расположились близко к кластеру АИ-98 (желтый) по результатам работы алгоритма примкнули к последним. То же произошло и с точками АИ-98, попавшими внутрь большого кластера АИ-95.

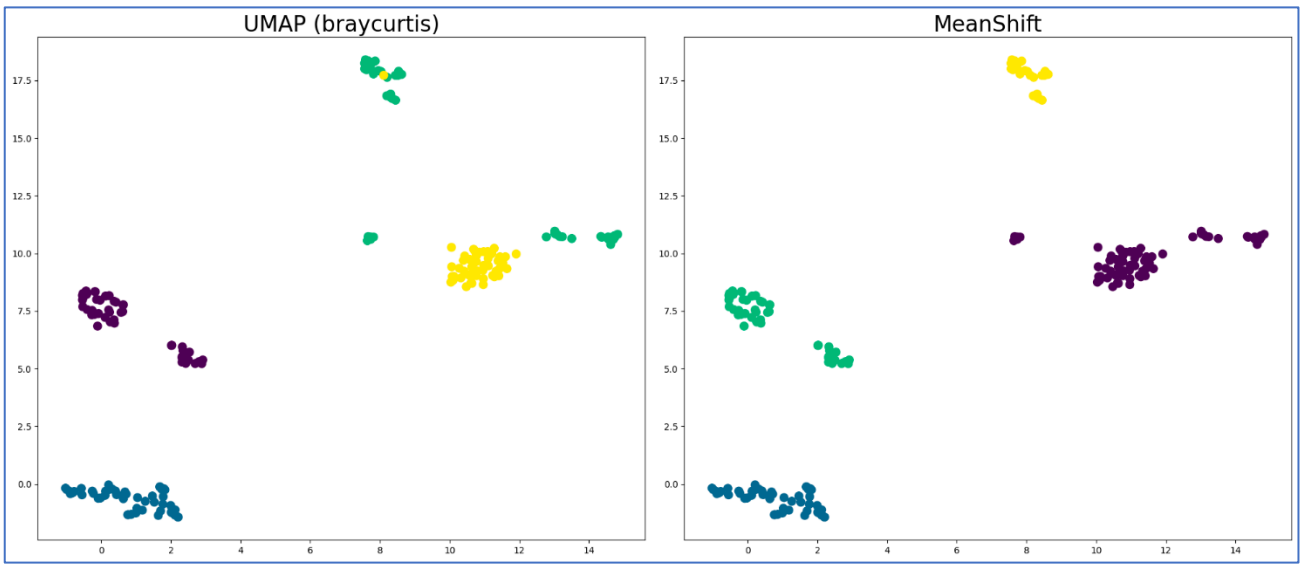


Рисунок 35 – Результаты работы алгоритма кластеризации MeanShift

На данных, обработанных MDS, PCA и нейросетевым методом результаты совершенно ошибочные – оценка от 0.2372 до 0.3579. Алгоритм также не смог выполнить кластеризацию на данных, к которым не был применен алгоритм снижения

размерности – лучшая оценка была получена на данных кумулятивной суммы (0.1955). Лучшие оценки работы алгоритма в зависимости от примененного алгоритма снижения размерности приведены в таблице 19.

Таблица 19 – Оценка ARI для MeanShift

Алгоритм СР	**MeanShift**
ISOMAP	0.695
LLE	0.685
MDS	0.237
No reduction	0.196
PCA	0.331
HM	0.358
SpectralEmbedding	0.559
TSNE	0.711
UMAP	0.744

3.3.2. Оценка основанной на плотности пространственной кластеризации приложений с шумами

Алгоритм DBSCAN в некоторых случаях совершенно провалился, совершенно не сумев выделить кластеры, но при правильной комбинации алгоритма снижения размерности и метрики расстояния самого алгоритма кластеризации удалось достигнуть отличных результатов (до оценки 0.9733).

Лучший результат был достигнут на данных свёртки, обработанных алгоритмом снижения размерности UMAP с использованием корреляционной метрики и использованием метрики Чебышёва самим алгоритмом, результат показан на рисунке 36. Видно, что алгоритм превосходно справился с различением трех кластеров, расположенных близко друг к другу. Проблемы возникли только с парой точек, которые были явно ошибочно расположены алгоритмом снижения размерности.

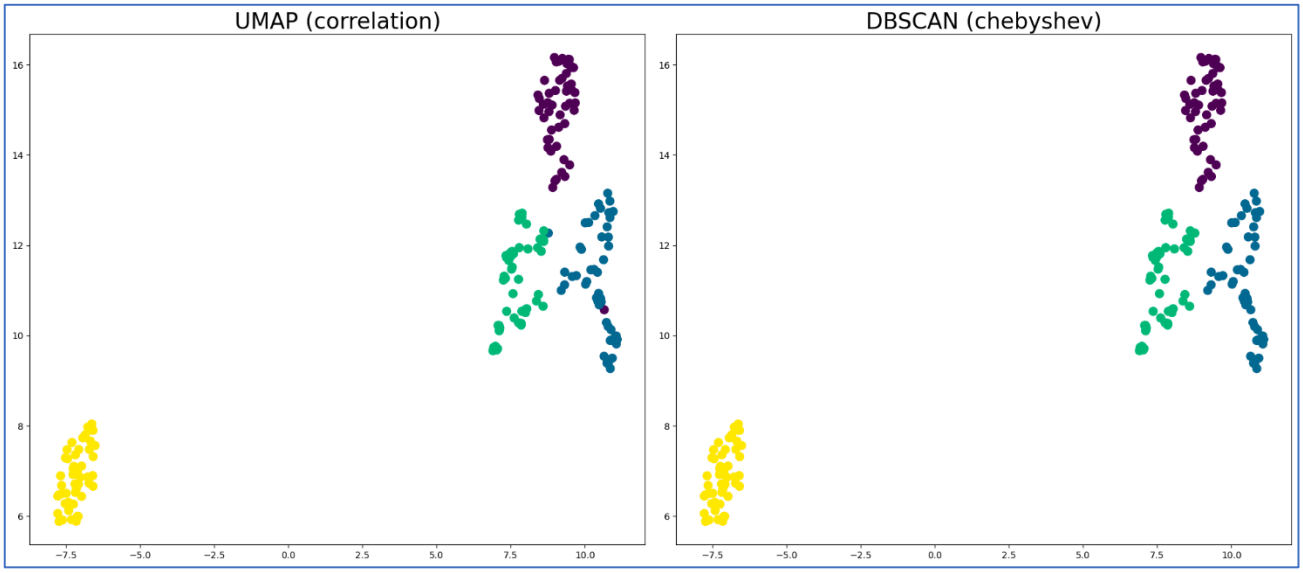


Рисунок 36 – Результаты работы алгоритма кластеризации DBSCAN

Оценки ARI лучших результатов работы алгоритма DBSCAN показаны в таблице 20. Алгоритм также не смог показать хороших результатов на данных, без применения алгоритма снижения размерности.

Таблица 20 – Оценка ARI для DBSCAN

\Метрика Алгоритм	**braycurtis**	**canberra**	**chebyshev**	**correlation**	**c o s i n e**	**euclidean**	**hamming**	**manhattan**
ISOMAP	0.671	0.848	0.463	0.467	0.330	0.610	0.129	0.732
LLE	0.683	0.855	0.000	0.496	0.685	0.001	0.330	0.006
MDS	0.002	0.555	0.103	0.328	0.000	0.053	0.000	0.044
No reduction	0.391	0.000	0.013	0.502	0.502	0.000	0.000	0.000
PCA	0.001	0.600	0.354	0.326	0.001	0.316	0.000	0.260
HM	0.000	0.389	0.330	0.000	0.000	0.330	0.000	0.330
SpectralEmbedding	0.461	0.692	0.152	0.357	0.391	0.152	0.000	0.152
TSNE	0.704	0.938	0.817	0.496	0.705	0.751	0.329	0.646
UMAP	0.794	0.960	0.973	0.496	0.711	0.924	0.000	0.843

3.3.3. Оценка упорядочения точек для обнаружения кластерной структуры

Алгоритм кластеризации OPTICS оказался самым неудачным среди рассматриваемых, и не смог показать хороших результатов. Результаты оценок представлены в таблице 21.

Таблица 21 – Оценка ARI для OPTICS

\Метрика Алгоритм	braycurtis	canberra	chebyshev	correlation	c o s i n e	euclidean	hamming	manhattan	minkowski
ISOMAP	0.217	0.247	0.238	0.467	0.244	0.222	0.110	0.254	0.222
LLE	0.364	0.422	0.398	0.496	0.685	0.477	0.330	0.398	0.398
MDS	0.147	0.136	0.135	0.328	0.128	0.164	0.000	0.147	0.160
No reduction	0.425	0.412	0.379	0.379	0.379	0.436	0.002	0.460	0.500
PCA	0.131	0.129	0.149	0.326	0.091	0.121	0.000	0.136	0.121
HM	0.175	0.124	0.158	0.071	0.101	0.112	0.000	0.190	0.150
Spectral Embedding	0.207	0.233	0.209	0.357	0.420	0.172	0.000	0.220	0.204
TSNE	0.472	0.481	0.489	0.496	0.412	0.540	0.329	0.470	0.520
UMAP	0.487	0.440	0.442	0.496	0.300	0.438	0.000	0.342	0.438

3.3.4. Оценка иерархической пространственной кластеризации приложений с шумами на основе плотности

Алгоритм HDBSCAN в среднем показал результаты на 70% лучше, чем DBSCAN, но в пике не смог достигнуть таких же отличных результатов.

Лучший результат работы алгоритма получен на данных, обработанных дискретной производной, с применением алгоритма UMAP (метрика Брея-Кёртиса), с применением самим алгоритмом метрики Канберры. Результат в данном случае оценен ARI в 0.8673. На рисунке 37 видно, что алгоритм хорошо справился с разделением кластеров, но одна из точек АИ-98 (желтый), расположенная близко к кластеру АИ-95 (бирюзовый) была присвоена алгоритмам к последнему. А вторая

половина точек АИ-95, разделенная от первой кластером АИ-98, была выделена алгоритмом как отдельный, пятый класс точек с выделением некоторых точек как выбросы (серый). Эти ошибки также скорее можно считать просчетом алгоритма снижения размерности, а не кластеризации.

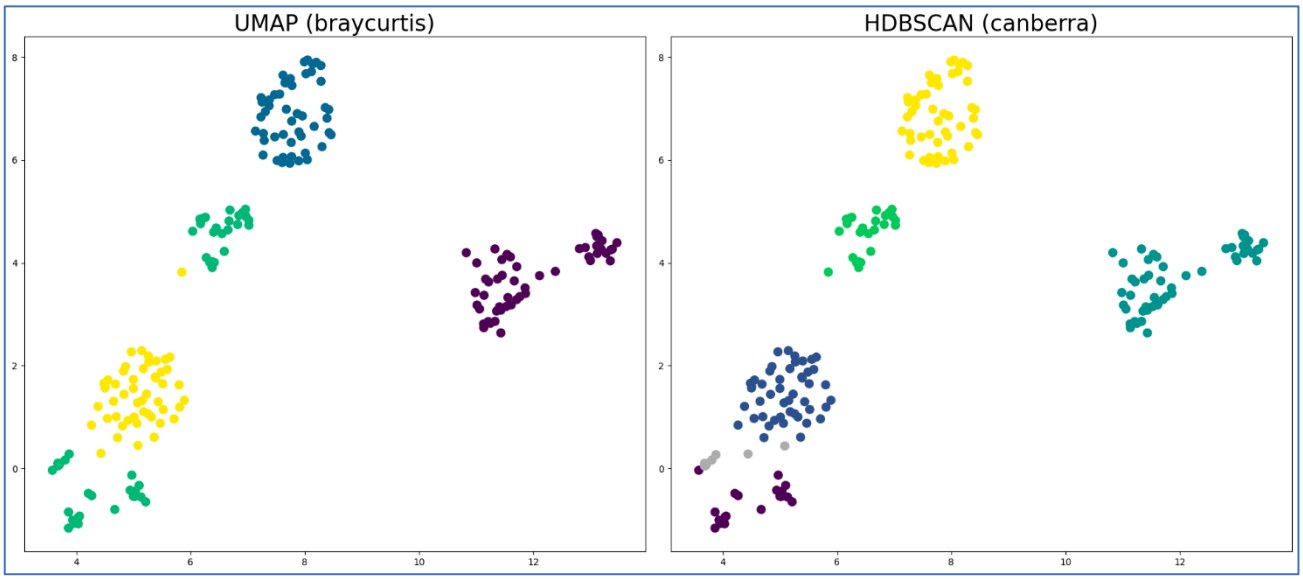


Рисунок 37 – Результаты работы алгоритма кластеризации HDBSCAN

Результаты работы алгоритма по оценке ARI показаны в таблице 22. Алгоритм также смог показать хорошие результаты на данных, обработанных дискретной производной, но без применения алгоритма снижения размерности (0.8175).

Таблица 22 – Оценка ARI для HDBSCAN

\Метрика Алгоритм	**braycurtis**	**canberra**	**chebyshev**	**correlation**	**cosine**	**euclidean**	**hamming**	**manhattan**
ISOMAP	0.802	0.851	0.663	0.467	0.612	0.794	0.254	0.651
LLE	0.739	0.733	0.728	0.496	0.726	0.754	0.419	0.805
MDS	0.442	0.549	0.423	0.328	0.467	0.536	0.000	0.528
No reduction	0.818	0.467	0.486	0.549	0.492	0.452	0.561	0.495
PCA	0.451	0.576	0.421	0.326	0.198	0.430	0.000	0.412
HM	0.417	0.415	0.402	0.313	0.432	0.381	0.000	0.395
Spectral Embedding	0.559	0.677	0.548	0.357	0.422	0.646	0.000	0.654
TSNE	0.846	0.840	0.859	0.496	0.577	0.853	0.562	0.859
UMAP	0.859	0.867	0.860	0.496	0.553	0.865	0.000	0.858

3.3.5. Оценка спектральной кластеризации

Алгоритм Spectral Clustering также смог достичь отличных результатов на данных, обработанных t-SNE и UMAP (0.9733), и также хороших результатов на данных, обработанных ISOMAP и Locally Linear Embedding с оценками 0.8851 и 0.8971 соответственно, что лучше, чем результаты алгоритмов DBSCAN и HDBSCAN.

Лучший результат работы алгоритма показан на рисунке 38. Результат получен в результате применения алгоритма к данным, обработанным t-SNE с корреляционной метрикой на данных свёртки. Видно, что у алгоритма возникли проблемы только с точками, которые алгоритм снижения размерности ошибочно расположил вблизи «чужих» кластеров.

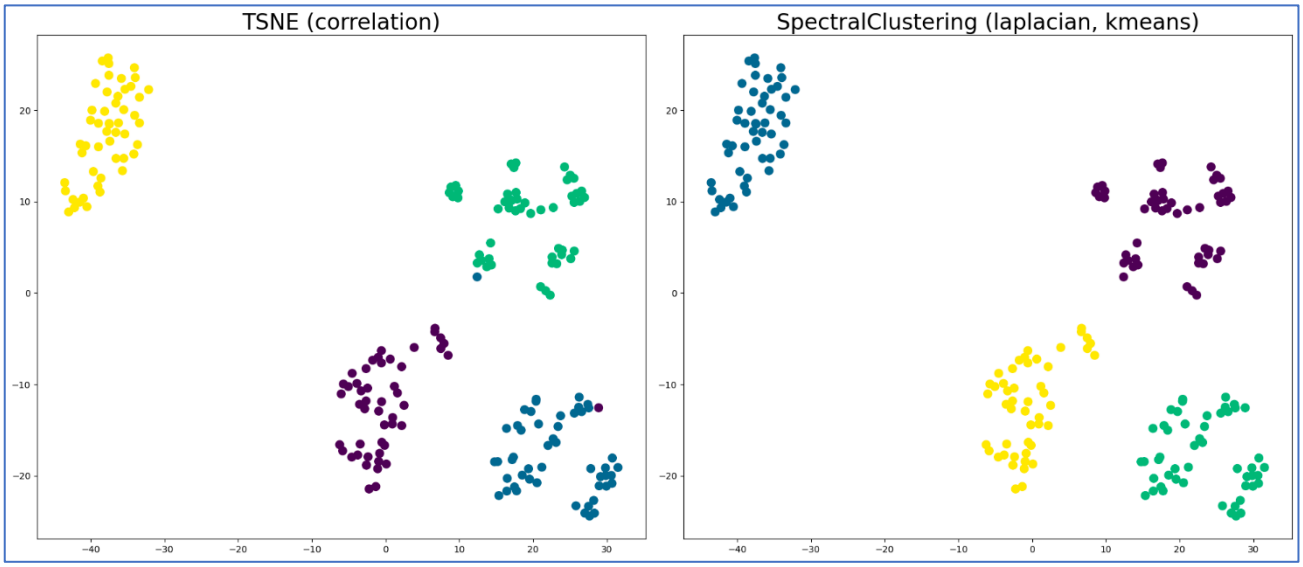


Рисунок 38 – Результаты работы алгоритма кластеризации Spectral Clustering

Результаты работы алгоритма сведены в таблицу 23. Алгоритм не смог достичь хороших результатов на данных, необработанных алгоритмом снижения размерности.

Таблица 23 – Оценка ARI для Spectral Clustering

Алгоритм СР	Spectral Clustering
ISOMAP	0.885
LLE	0.897
MDS	0.620
No reduction	0.354
PCA	0.461
HM	0.364
SpectralEmbedding	0.684
TSNE	0.973
UMAP	0.973

3.3.6. Оценка K-средних

Алгоритм K-means в среднем показал лучшие результаты среди прочих, сохранив ту же тенденцию отличных результатов при использовании t-SNE и UMAP и хороших результатов при использовании ISOMAP и LLE.

Лучший результат алгоритма показан на рисунке 39. Здесь, как и в предыдущих случаях, ошибочно определены были лишь пара точек, которые были некорректно расположены алгоритмом снижения размерности, в остальном алгоритм отлично справился с разделением даже близко расположенных и не совсем сжатых и целостных кластеров. Оценка ARI в данном случае равна 0.9733.

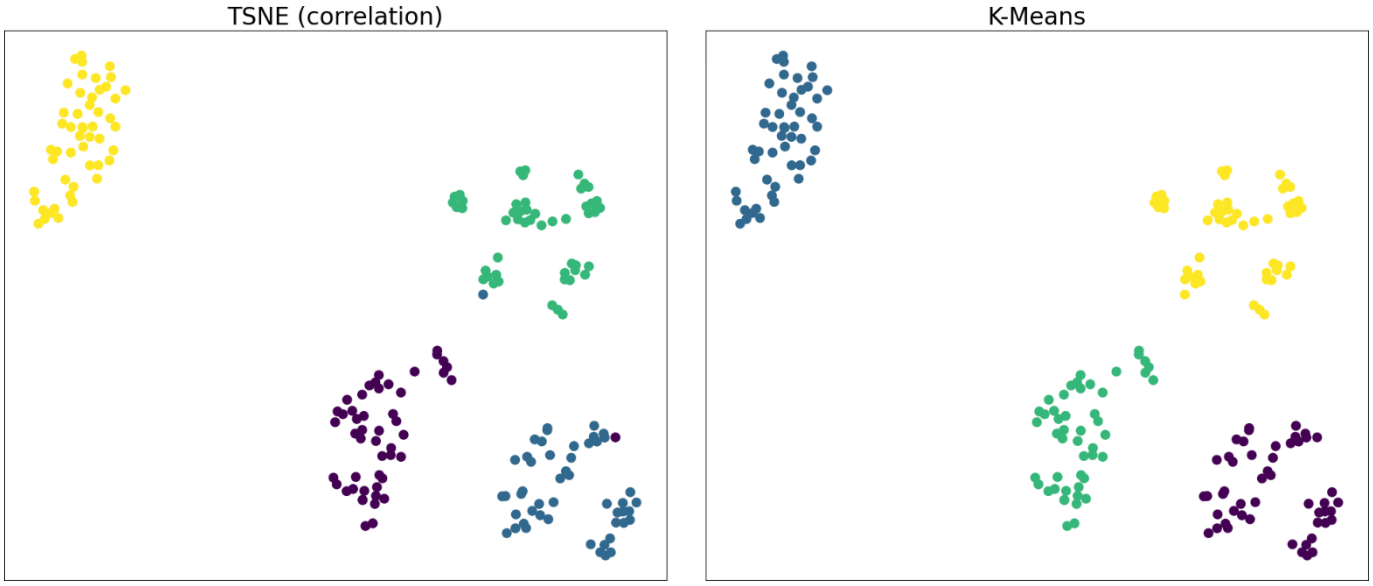


Рисунок 39 – Результаты работы алгоритма кластеризации K-means

Оценки лучших результатов работы алгоритма показаны в таблице 24. Алгоритм также не справился с кластеризацией необработанных данных.

Таблица 24 – Оценка ARI для K-means

Алгоритм СР	**K-Means**
ISOMAP	0.921
LLE	0.897
MDS	0.541
No reduction	0.555
PCA	0.472
HM	0.357
SpectralEmbedding	0.638
TSNE	0.973
UMAP	0.973

3.3.7. Анализ оценок алгоритмов кластеризации

Сравнительный анализ проводился для алгоритмов кластеризации MeanShift, DBSCAN, OPTICS, HDBSCAN, Spectral Clustering и K-means. Каждый из них показал свою специфику при работе с данными, предварительно обработанными различными методами снижения размерности. Общая тенденция, наблюдаемая в исследовании,

такова: качественное снижение размерности создаёт основу для построения корректных кластеров, тогда как работа алгоритмов с необработанными данными зачастую приводит к ошибкам, вызванным высокой размерностью и шумом в данных. Лучшие результаты работы алгоритмов представлены в таблице 25.

Таблица 25 – Результаты алгоритмов кластеризации

\Кластеризация Снижение разм.	DBSCAN	HDBSCAN	K-Means	MeanShift	OPTICS	SpectralClustering
ISOMAP	0.848	0.851	0.921	0.695	0.467	0.961
LLE	0.855	0.805	0.897	0.685	0.685	0.936
MDS	0.555	0.549	0.541	0.237	0.328	0.620
No reduction	0.502	0.818	0.555	0.196	0.500	0.618
PCA	0.600	0.576	0.472	0.331	0.326	0.581
HM	0.389	0.432	0.357	0.358	0.190	0.422
SpectralEmbedding	0.692	0.677	0.638	0.559	0.420	0.708
TSNE	0.938	0.859	0.973	0.711	0.540	0.973
UMAP	0.973	0.867	0.973	0.744	0.496	0.973

3.4. Обзор результатов анализа сочетаний алгоритмов снижения размерности и кластеризации

Проведенное исследование позволило оценить эффективность различных алгоритмов кластеризации в сочетании с методами снижения размерности. Исследование показало, что предварительное снижение размерности данных существенно влияет на качество кластеризации. Применение алгоритмов вроде t-SNE и UMAP демонстрирует заметное улучшение результатов по сравнению с исходными данными, обработанными методами, такими как MDS, PCA или вовсе без снижения размерности.

Среди алгоритмов снижения размерности UMAP и t-SNE оказались наиболее эффективными методами предобработки. Их способность сохранять локальную и глобальную структуру данных позволила улучшить кластеризацию для всех алгоритмов, особенно для K-means, Spectral Clustering и DBSCAN.

ISOMAP и LLE показали умеренную эффективность, обеспечивая ARI в диапазоне 0.68–0.92. Они подходят для данных с нелинейной структурой, но уступают UMAP и t-SNE в сложных сценариях.

Линейные методы (PCA, MDS) оказались наименее полезными. Их применение часто приводило к значительному падению качества кластеризации (ARI ≤ 0.6), что связано с неспособностью учитывать нелинейные зависимости.

Отказ от снижения размерности в большинстве случаев ухудшал результаты, показывая ARI менее 0.5, за исключением HDBSCAN, который частично компенсировал высокую размерность за счет анализа плотности.

Среди алгоритмов кластеризации MeanShift показал ограниченную эффективность. Наилучшие результаты (ARI: 0.6854-0.7442) были достигнуты с нелинейными методами снижения размерности (UMAP, t-SNE), тогда как линейные методы (PCA, MDS) и отсутствие предобработки данных привели к низким оценкам (ARI ≤ 0.3579). Основная проблема алгоритма – чувствительность к перекрытию кластеров и шуму.

DBSCAN продемонстрировал хорошие результаты: в худших случаях (например, с MDS или без снижения размерности) алгоритм полностью провалился (ARI ≈ 0), но при использовании UMAP с метрикой Чебышёва достиг пиковой точности с оценкой ARI равной 0.9733. Это подчеркивает критическую зависимость DBSCAN от выбора метрики расстояния и качества предобработки данных.

OPTICS оказался наименее эффективным алгоритмом. Максимальный ARI (0.5195) был получен с t-SNE, но даже этот результат значительно уступает другим методам.

HDBSCAN улучшил результаты DBSCAN в среднем на 70%, достигнув ARI в 0.8673 с UMAP и метрикой Канберры. Однако алгоритм допустил ошибки в зонах

перекрытия кластеров, что связано с артефактами снижения размерности. Интересно, что HDBSCAN показал достойные результаты даже на необработанных данных (ARI=0.8175), что выделяет его среди аналогов.

Spectral Clustering и K-means продемонстрировали наивысшую точность. Оба алгоритма достигли максимального ARI (0.9733) с t-SNE и UMAP, что связано с их способностью работать с нелинейными структурами. K-means также показал стабильно высокие результаты с ISOMAP и LLE (ARI ≈ 0.9), подтвердив свою универсальность. Стоит отметить, что последним шагом Spectral Clustering является применение алгоритма K-means, так что при отсутствии лучших результатов в пике, а также худших результатов в среднем можно сказать, что применение этого алгоритма не оправдано по крайней мере по соображениям вычислительной сложности.

На основе проведённого исследования можно сделать следующие рекомендации:

Этап предобработки данных критически важен. Без снижения размерности даже продвинутые алгоритмы (кроме HDBSCAN) демонстрируют низкую эффективность. Для задач кластеризации спектральных данных рекомендуется применять методы снижения размерности t-SNE или UMAP. Они позволяют сохранить важные геометрические и топологические особенности исходного пространства.

Каждому алгоритму необходимо подбирать индивидуальные параметры (например, метрику расстояния для DBSCAN или пороговые значения для HDBSCAN), так как оптимальное сочетание параметров может существенно повысить качество кластеризации. Результаты показывают, что даже небольшие изменения в выборе метрики могут привести к резкому улучшению или ухудшению результатов.

Неверное позиционирование отдельных точек, наблюдаемое в ряде случаев (например, ошибки при распределении точек между соседними кластерами), зачастую обусловлено погрешностями методов снижения размерности. Поэтому при

интерпретации результатов важно учитывать, что ошибки могут возникать не из-за некорректности алгоритма кластеризации, а вследствие особенностей предварительной обработки данных. Это также говорит о необходимости разработки новых, более совершенных методов предварительной обработки и снижения размерности, а не кластеризации.

Алгоритмы Spectral Clustering и K-means продемонстрировали наилучшие результаты в сочетании с t-SNE и UMAP. Однако для более сложных, нечетко разделённых данных, а также при отсутствии знаний о количестве кластеров, могут оказаться предпочтительными методы, учитывающие плотность распределения точек, такие как DBSCAN или HDBSCAN [57].

3.5. Описание разработанной методики формирования цифровых образов и метода снижения размерности

Проведённое исследование подчёркивает важность комплексного подхода при решении задач кластеризации. Ключевым моментом является не только выбор самого алгоритма кластеризации, но и предварительная обработка данных, включая снижение размерности. Правильное сочетание методов позволяет добиться высокой точности и стабильности результатов, что особенно важно в областях, требующих объективной валидации кластеров, таких как биоинформатика, компьютерное зрение и социальные науки.

В качестве результата была разработан метод кластеризации спектральных данных, описанный на рисунке 40, ключевой особенностью которой является акцент на этапе предварительной математической обработки (трансформации спектров) и последующей визуализации за счёт алгоритма снижения размерности. В

предложенном методе предварительная обработка выступает не просто фильтром шумов, а инструментом формирования информативного представления, оптимизированного под выбранную метрику и алгоритм встраивания. Применение многоэтапной процедуры: адаптивной предобработки и снижения размерности с подбором метрики размерности и последующей кластеризации – обеспечивает улучшение разделимости классов, повышение устойчивости к матричным искажениям, удобство визуального контроля качества разбиения и упрощение интерпретации результатов специалистом. В совокупности с адаптивным подбором параметров предварительной обработки и выбором числа кластеров для K-Means на основе внутренних метрик кластеризации формулируется новый метод кластерного анализа спектральных данных химических веществ, который позволяет добиться большей автоматизации, масштабируемости и интерпретируемости, что делает метод эффективным инструментом в современном системном анализе спектральных данных [52].

Рисунок 40 – Предложенный алгоритм кластеризации

На основе полученных результатов также предлагается следующая методика формирования цифровых образов веществ по их спектральным данным, представляющая собой комплексный подход, включающий несколько последовательных этапов обработки, описанных на рисунке 41. На первом этапе спектры подвергаются предварительной обработке, которая реализуется в двух вариантах: либо с использованием дискретной свёртки, позволяющей сглаживать

данные и выделять скрытые закономерности, либо с применением дискретной производной, обеспечивающей акцентирование локальных изменений интенсивностей и повышение контрастности информативных признаков. Далее для уменьшения размерности и перехода к компактным цифровым представлениям используется один из современных нелинейных методов визуализации многомерных данных – t-SNE или UMAP. При этом выбор метрики расстояния адаптирован к характеру предварительной обработки: для данных, обработанных свёрткой, применяется корреляционная метрика, которая отражает степень сходства спектров с точки зрения формы сигналов, тогда как для данных, полученных через дискретную производную, используется метрика Канберры, более чувствительная к относительным изменениям компонент спектра. Такое сочетание методов позволяет формировать устойчивые и информативные цифровые образы веществ, что обеспечивает более точную последующую кластеризацию методом K-Means.

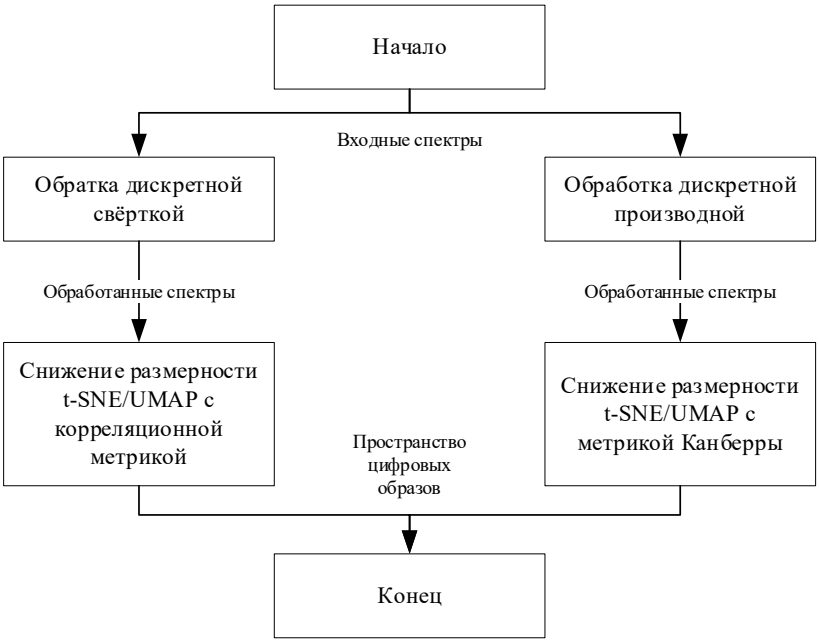


Рисунок 41 – Разработанная методика формирования цифровых образов спектральных данных

3.6. Валидация разработанного метода на больших данных

Во время разработки и отладки метода эксперименты проводились на образцах различных бензинов, собранных с помощью ИК-Фурье спектрометра АФ-3 [57]. Эти данные содержат 4 вида бензинов по 50 образцов каждого. Теперь апробация точности разработанного метода будет проведена на открытом наборе рамановских спектров химических соединений активных фармацевтических ингредиентов (API) [104].

В основе набора лежат 3510 отдельных рамановских спектров для 32 коммерческих соединений (растворителей и реагентов), полученных с помощью анализатора Endress+Hauser Raman Rxn2 с возбуждением лазером длиной волны 785 нм и разрешением 1 $см^{-1}$ в диапазоне 150…3425 $см^{-1}$ (всего 3276 отсчётов данных на спектр). В каждый эксперимент включена автоматическая предварительная обработка (вычитание темнового тока, фильтрация космических лучей и коррекция интенсивности) без какого-либо дальнейшего ручного сглаживания или фильтрации. При этом образцы хранятся и измеряются в лабораторных условиях с учётом рекомендаций по хранению, чтобы свести к минимуму влияние примесей [103].

Таблица 26 содержит сводные показатели точности кластеризации для шести алгоритмов кластеризации (в столбцах: DBSCAN, HDBSCAN, K-Means, MeanShift, OPTICS, SpectralClustering) и восьми методов снижения размерности (в строках: ISOMAP, Locally Linear Embedding, MDS, без снижения, PCA, Spectral Embedding, t-SNE, UMAP), применённых к исходным спектрам и к векторам, полученным с помощью дискретной производной в соответствии с первым шагом предложенной методики.

По результатам видно, что предложенная методика обеспечивает высокое значение точности, близкое к 99,7%. Эти результаты свидетельствуют о том, что

нелинейные алгоритмы, сохраняющие локальную структуру данных (особенно с применением метрики Канберры), эффективно выделяют грань между классами спектров веществ.

Таблица 26 – Результаты точности предложенной методики на рамановских спектрах

\Кластеризация Обработка, СР	DBSCAN	HDBSCAN	K-Means	MeanShift	OPTICS	SpectralClustering
raman	**0.7**	**0.794**	**0.924**	**0.095**	**0.082**	**0.913**
ISOMAP	0.064	0.131	0.183	0.049	0.024	0.323
LLE	0.122	0.229	0.361	0.071	0.082	0.575
MDS	0.06	0.073	0.195	0.008	0.03	0.27
-	0	0.543	0.281	0.01	0.047	0.753
PCA	0.06	0.051	0.187	0.01	0.041	0.228
SE	0.098	0.182	0.343	0.037	0.026	0.386
TSNE	0.7	0.794	0.924	0.058	0.06	0.913
UMAP	0.523	0.185	0.615	0.095	0.054	0.623
raman_diff	**0.96**	**0.958**	**0.997**	**0.073**	**0.819**	**0.967**
ISOMAP	0.106	0.081	0.227	0.022	0.057	0.375
LLE	0.424	0.822	0.768	0.073	0.819	0.836
MDS	0.111	0.218	0.283	0.013	0.044	0.496
-	0.783	0.958	0.29	0.019	0.819	0.714
PCA	0.062	0.19	0.223	0.019	0.035	0.465
SE	0.371	0.563	0.404	0.013	0.055	0.749
TSNE	0.938	0.932	0.997	0.062	0.19	0.967
UMAP	0.96	0.487	0.986	0.062	0.062	0.967

В таблице 27 приведены результаты оценки точности использования популярных алгоритмов для идентификации тех же рамановских спектров. Модели этих методов обучались на 80% данных. По результатам видно, что методы PLS-DA и SVM не справились с задачей, показав точность около 70%, а методы LDA и SIMCA показали хорошие результаты в 94% и 93% соответственно.

Таблица 27 – Результаты популярных методов на рамановских спектрах

Метод	Точность
PLS-DA	0.701
LDA	0.946
SIMCA	0.931
SVM	0.732

Рисунок 42 иллюстрирует двумерную проекцию обработанных в соответствии с предложенной методикой данных с помощью дискретной производной и применения алгоритма t-SNE с метрикой Канберры. Ярко выраженные удалённые друг от друга компактные кластеры подчёркивают эффективность преобразования, позволяющего выполнить кластеризацию. Единственными почти слившимися классами являются «4-метилпентан-2-он» и «метилизобутилкетон», которые по сути являются одним и тем же веществом.

Полученные результаты указывают на то, что предложенная методика превосходит классические методы машинного обучения (PLS-DA, LDA, SIMCA, SVM) по точности кластеризации и при этом существенно проще в настройке и применении, в том числе не требует предварительного обучения.

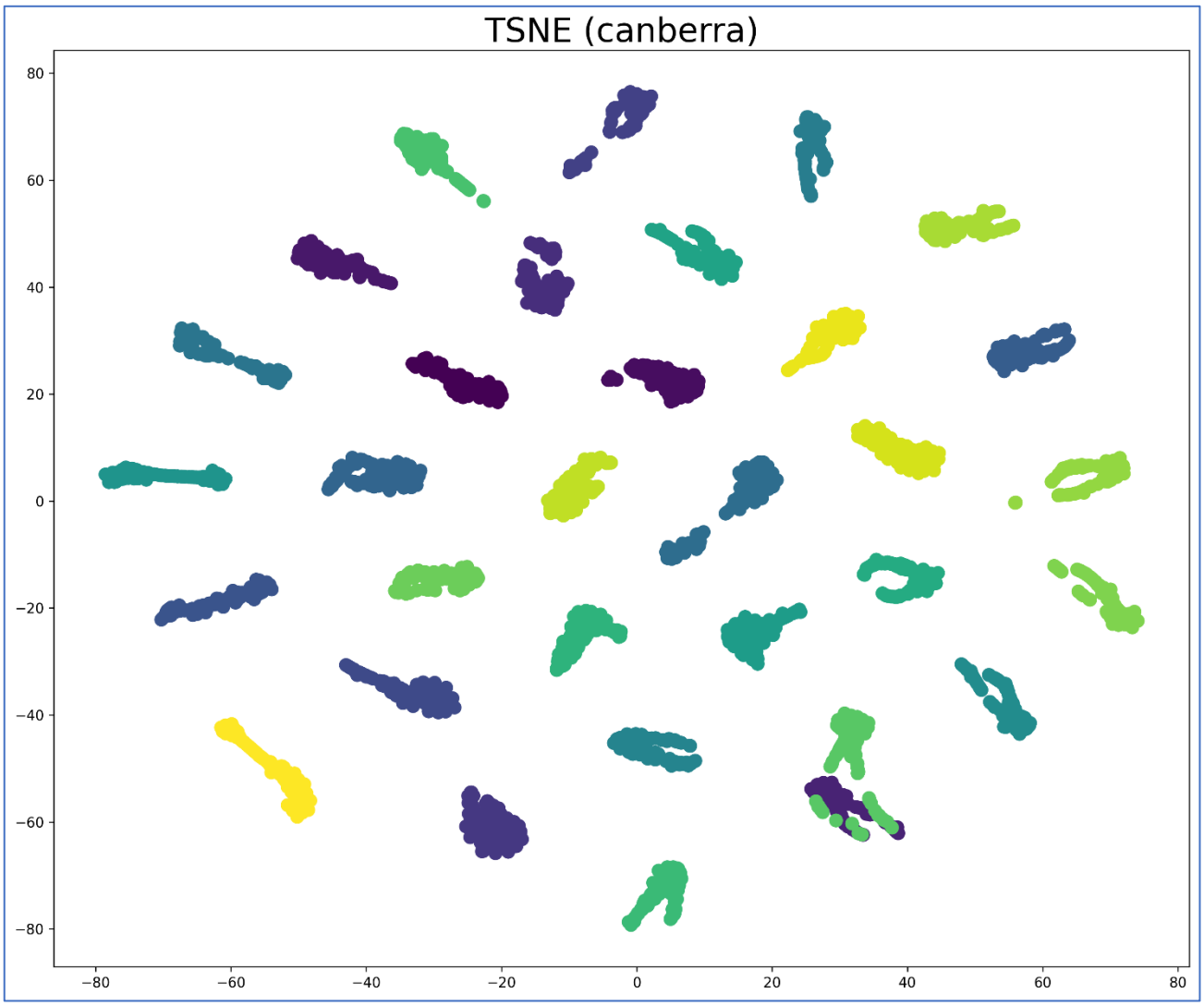


Рисунок 42 – Визуализация разделения спектров

Разработанный метод кластеризации спектральных данных в совокупности с методикой формирования их цифровых образов сочетает простоту реализации, высокую точность и универсальность применения к различным типам спектров. Его внедрение в аналитические лаборатории позволяет существенно ускорять и упрощать задачи качественной оценки веществ: от контроля качества продуктов и фармацевтики до биомедицинской диагностики и экологического мониторинга.

3.7. Выводы по третьей главе

Выводы по третьей главе диссертационной работы сформулированы следующим образом. Выполненный в третьей главе комплексный анализ этапов предварительной обработки спектров, выбора метрик расстояния, методов нелинейного снижения размерности и алгоритмов кластеризации подтвердил ключевую гипотезу о том, что качество автоматической классификации спектральных данных определяется не отдельным компонентом процедуры, а их согласованным сочетанием. Эксперименты показали, что корректно подобранная цепочка преобразований – вычисление дискретной производной, последующее нелинейное вложение (в частности t-SNE или UMAP) с метрикой Канберры и далее кластеризация методом K-Means – обеспечивает устойчивое выделение информативных признаков и высокую разделимость классов в наблюдаемых пространствах. Эти результаты формализованы в главе и подтверждены визуализациями и количественными метриками качества кластеризации.

Добавленные в третью главу дополнительные исследования, включающие валидацию разработанного алгоритма на крупном открытом наборе рамановских спектров (3510 спектров для 32 коммерческих соединений), продемонстрировали высокую переносимость предложенной методики. При применении первого шага – дискретной производной – в сочетании с нелинейными алгоритмами снижения размерности и последующим K-Means получены значения точности кластеризации, близкие к верхней границе возможного (99.7%), что свидетельствует о практической применимости метода к разнотипным спектральным данным. Эти выводы иллюстрируются сводными таблицами точности и проекциями, приведёнными в главе.

Оценка производительности алгоритма выявила важные особенности масштабирования. На небольших наборах бензинов время построения модели предложенной методикой сопоставимо с классическими методами и находится в одном порядке величин. При переходе к большому набору рамановских спектров предложенный алгоритм оказывается существенно быстрее PLS-DA и остаётся вполне приемлемым по сравнению с LDA, что указывает благоприятную масштабируемость при обработке высокоразмерных и объёмных спектральных массивов. Таким образом, с точки зрения соотношения «точность/время» разработанный подход оказывается конкурентоспособным.

Исследование ограничений метода показало, что ключевыми факторами, влияющими на качество, остаются выбор способа предварительной обработки и метрики расстояния. В ряде конфигураций некорректно подобранная предобработка или неадаптированная метрика приводили к ухудшению разделимости и снижению показателей; пара случаев почти неразличимых веществ («4-метилпентан-2-он» и «метилизобутилкетон») иллюстрирует пределы разрешающей способности чисто спектральных признаков при отсутствии дополнительной предметной информации. Это подчёркивает необходимость внимательной валидации комбинаций методов и корректной настройки гиперпараметров в зависимости от конкретной предметной области. С другой стороны, разработанный метод может выявить ошибки во входных данных в то время, как обучаемые методы будут обучаться ошибкам.

Предложенная методика обладает преимуществом простоты реализации и отсутствием в потребности в предварительном обучении, что облегчает её интеграцию в аналитические рабочие процессы. В то же время дальнейшее повышение автоматизации возможно за счёт внедрения модулей автоматического подбора гиперпараметров, адаптивного выбора метрик и гибридных схем, объединяющих

преимущества плотностных и центроидных алгоритмов кластеризации для обработки сильно неравномерных или перекрывающихся классов.

Третья глава даёт обоснование выбранного подхода: сочетание целевых процедур предобработки, нелинейного снижения размерности и последующей кластеризации обеспечивает эффективную, воспроизводимую и практически применимую систему для классификации спектральных данных. При этом достигнутые результаты на бензиновых и рамановских наборах подтверждают универсальность метода и указывают на конкретные направления дальнейшей оптимизации и адаптации под прикладные задачи аналитической химии и контроля качества.

Глава 4. Описание разработанного программного обеспечения

4.1. Структура базы данных спектрального анализа

Реляционная база данных реализована на PostgreSQL. База данных хранит как сами спектры и их классы и подклассы, так и результаты снижения размерности и кластеризации [62]. На рисунке 43 показана схема таблиц, предназначенных для хранения иерархически структурированных данных о спектрах с возможностью их классификации и анализа.

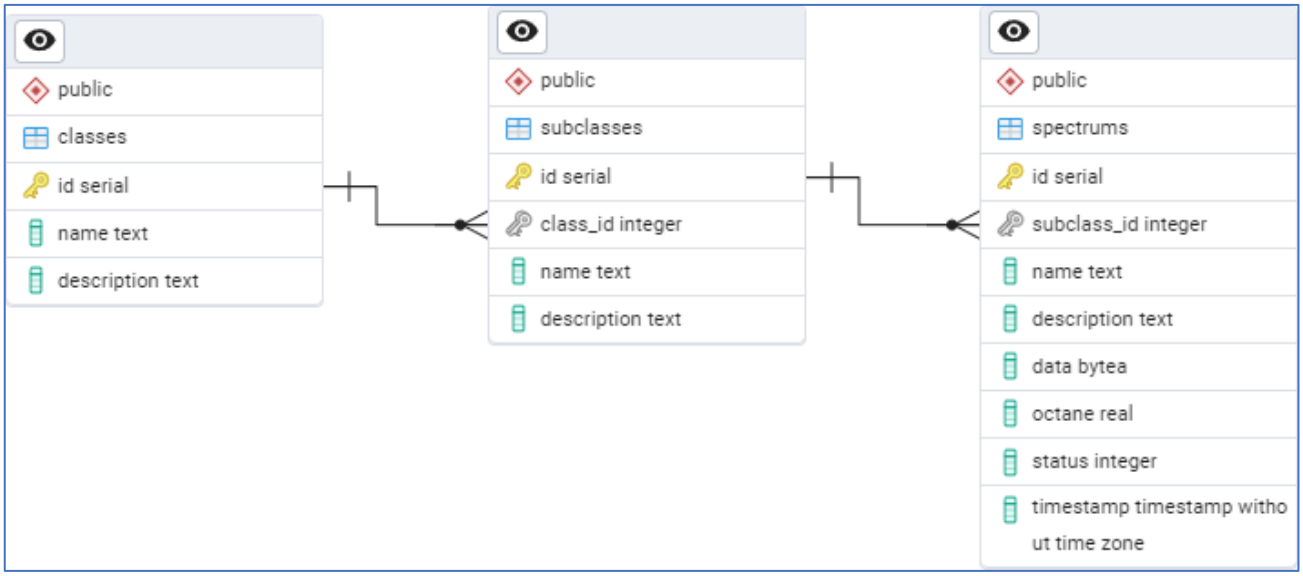


Рисунок 43 – Схема базы данных (спектры)

Схема состоит из трёх основных таблиц и поддерживает связи с дополнительными таблицами, связанными с кластеризацией данных.

1. Таблица classes (классы) хранит категории для классификации спектров. Поля:
   - id (serial) – уникальный идентификатор класса.
   - name (text) – название класса.
   - description (text) – описание (по умолчанию пустая строка).

2. Таблица subclasses (подклассы) хранит дочерние элементы классов. Каждый подкласс принадлежит одному классу. Поля:
   - id (serial) – уникальный идентификатор подкласса.
   - class_id (integer) – ссылка на classes.id (внешний ключ).
   - name (text) – название подкласса.
   - description (text) – описание (по умолчанию пустая строка).

3. Таблица spectrums (спектры) хранит данные о спектрах, включая двоичные данные измерений. Поля:
   - id (serial) – уникальный идентификатор спектра.
   - subclass_id (integer) – ссылка на subclasses.id (внешний ключ).
   - name (text) – название спектра.
   - description (text) – описание (по умолчанию пустая строка).
   - data (bytea) – бинарные данные спектра.
   - octane (real) – октановое число (для спектров, отображающих бензины).
   - status (integer) – статус обработки (по умолчанию 0).
   - timestamp (timestamp) – время добавления записи (автоматически заполняется текущим временем).

База данных имеет следующие особенности:

Поле spectrums.data (тип bytea) позволяет хранить любые файлы спектров (например, изображения, измерения или структурированные данные).

Поле spectrums.status (по умолчанию 0) может использоваться для отслеживания этапов обработки данных (например, 0 – не обработан, 1 – в процессе, 2 – завершен, 3 – эталонный, 4 – устаревший).

Поле spectrums.timestamp автоматически фиксирует время добавления записи.

На рисунке 44 представлена вторая часть схемы базы данных, добавляющая структуры для хранения и анализа моделей машинного обучения, связанных с кластеризацией и снижением размерности данных из первой части базы.

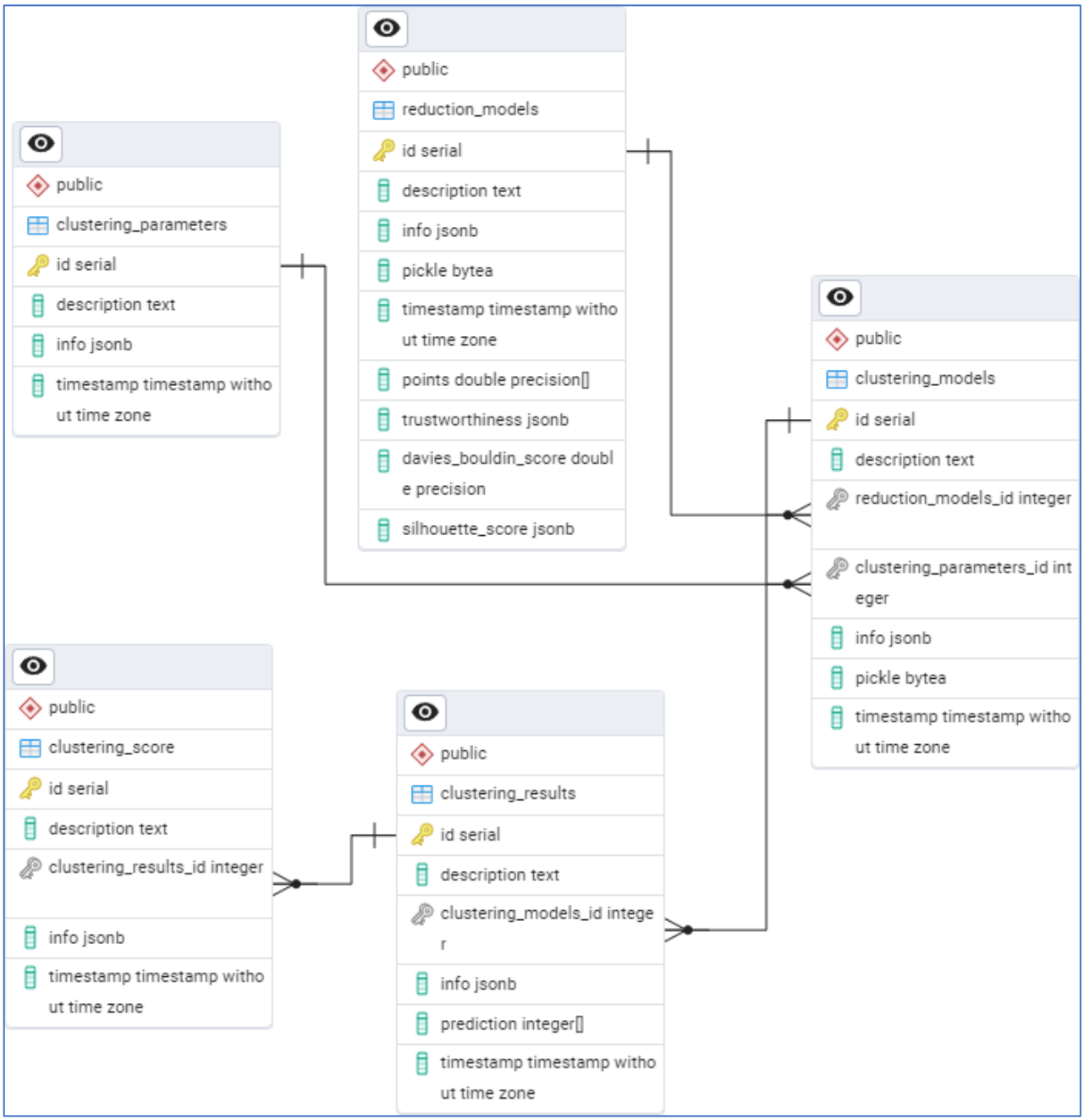


Рисунок 44 – Схема базы данных (обработка)

Основные таблицы:

1. Таблица reduction_models (модели снижения размерности) хранит информацию о моделях, уменьшающих размерность данных (например, UMAP, PCA, t-SNE). Поля:
   - id (serial) – уникальный идентификатор модели.
   - info (jsonb) – параметры модели (metric, n_components, random_state и др.).
   - pickle (bytea) – сериализованная модель (бинарные данные, например, через pickle).
   - points (double precision[]) – массив точек в уменьшенной размерности (результат преобразования данных).
   - trustworthiness, davies_bouldin_score, silhouette_score – метрики качества снижения размерности.
   - timestamp – время создания модели.

2. Таблица clustering_parameters (параметры кластеризации) Содержит настройки алгоритмов кластеризации (например, K-Means, DBSCAN). Поля:
   - info (jsonb) – параметры алгоритма (n_clusters, random_state, название метода).
   - timestamp – время создания параметров.

3. Таблица clustering_models (модели кластеризации) объединяет модели снижения размерности и параметры кластеризации для создания итоговой модели кластеризации результатов работы моделей снижения размерности. Поля:
   - reduction_models_id – ссылка на модель снижения размерности.

- clustering_parameters_id – ссылка на параметры кластеризации.
- pickle (bytea) – сериализованная модель кластеризации.
- info (jsonb) – дополнительные данные (например, имя набора данных).

4. Таблица clustering_results (результаты кластеризации) хранит предсказанные кластеры для данных. Поля:
   - clustering_models_id – ссылка на модель, использованную для кластеризации.
   - prediction (integer[]) – массив меток кластеров для каждого объекта (спектра).
   - info (jsonb) – метаданные (например, имя набора данных).

5. Таблица clustering_score (оценки кластеризации) содержит метрики качества кластеризации (например, силуэт, индекс Дэвиса-Боулдина). Поля:
   - clustering_results_id – ссылка на результат кластеризации.
   - info (jsonb) – метод оценки (например, silhouette_score).

База данных имеет следующие особенности:

Поля info хранят параметры и метаданные в гибком формате JSON, что позволяет адаптировать схемы под разные алгоритмы (например, для UMAP: min_dist, n_neighbors; для K-Means: n_clusters).

Поля pickle (в reduction_models и clustering_models) позволяют сохранять обученные модели для повторного использования, что положительно сказывается на скорости проведения исследований.

Данные из таблицы spectrums могут быть преобразованы в низкоразмерное представление через reduction_models.

Результаты кластеризации (clustering_results.prediction) могут быть связаны с исходными спектрами через внешние ключи. Например, массив prediction может соответствовать выборке спектров, обработанных через конкретную модель.

4.2. Описание программы для проведения исследования

Программный модуль для проведения исследования обеспечивает взаимодействие с базой данных спектрального анализа, включая создание моделей машинного обучения для снижения размерности и кластеризации, сохранение результатов и их визуализацию. Код написан на Python с использованием асинхронных операций для работы с PostgreSQL и интеграцией библиотек машинного обучения. Ниже описаны основные компоненты программы и их функции.

1. Инициализация базы данных
    - init(). Создает таблицы classes, subclasses, spectrums.
    - create_tables(). Создает таблицы для моделей (reduction_models, clustering_parameters и др.).
    - defineclasses(). Заполняет базовые классы бензинов (АИ-80, АИ-92 и др.).
2. Работа со спектрами
    - get_spectrums(ids). Возвращает спектры по их ID, декодируя данные из CSV.
    - save_reduction_picture(). Визуализирует точки после снижения размерности и сохраняет в изображение (2D/3D).

3. Модели снижения размерности
    - create_reduction_model(). Создает модель снижения размерности на основе параметров.
    - save_reduction_model(). Сохраняет модель и её метрики (точки, оценки SC, DBI) в БД.
    - update_reduction_model(), update_reduction_points(). Обновляют существующие модели.
    - get_reduction_model(). Ищет модель в БД по параметрам.
4. Модели кластеризации
    - create_clustering_model(). Обучает модель (K-Means, DBSCAN, SpectralClustering и др.).
    - save_clustering_params(), save_clustering_model(). Сохраняют параметры и модели в БД. Модели сериализуются через pickle и сохраняются как bytea.
    - get_clustering_parameters(), get_clustering_model(). Поиск параметров и моделей.
5. Результаты кластеризации
    - save_clustering_results(). Сохраняет предсказанные метки кластеров.
    - save_score_results(). Рассчитывает и сохраняет метрики качества (например, ARI).
    - get_max_score(). Возвращает лучшие результаты кластеризации по заданным критериям.
6. Анализ и оптимизация
    - get_clustering_result_info(). Агрегирует данные из БД (суммы, средние, максимумы).
    - shuffle_data(). Перемешивает данные для кросс-валидации.

Особенности реализации:

- Настройки подключения к БД загружаются из внешнего файла settings.json.
- Уровень детализации логирования управляется через DEBUG_INFO.
- Все операции с БД используют async/await для неблокирующих запросов.

4.3. Описание модуля экспертно-нейросетевой системы

На рисунке 45 представлена общая схема функционирования ЭНС. На схеме изображен процесс работы системы, интегрированной в производство для контроля качества продуктов нефтеперерабатывающей промышленности (автомобильных бензинов) по их спектральным данным [15]. В текущей реализации спектральные данные описываются центроидами и дисперсиями, а для классификации используются нейронные сети, работающие по методу определения ближайших соседей. Используемая однослойная нейронная сеть определяет ближайшего соседа с уверенностью в 43%, что позволяет выполнить задачу на рассматриваемых четырех классах бензинов, но создает риск провалить эту задачу на более разнообразном наборе данных. Разработанные модуль решает эти проблемы, заменяя используемые решения, используемые для цифровых представлений веществ и их кластеризации.

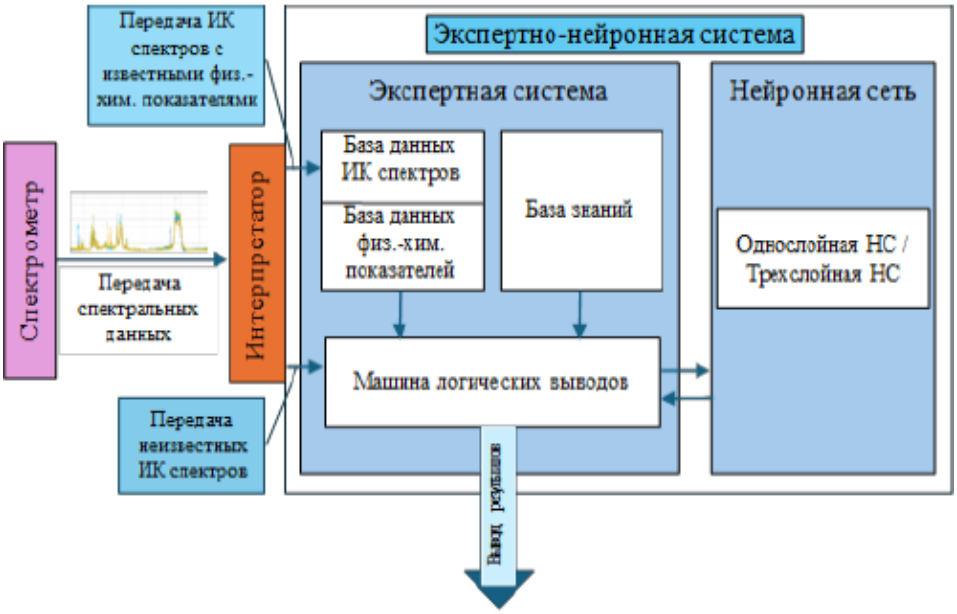


Рисунок 45 – Схема функционирования ЭНС

4.3.1. Программная реализация модуля

В этом разделе описаны программные модификации, внедренные в ЭНС для обеспечения работы модуля. Разработанный модуль, как и ЭНС, реализован на Node.js с использованием Express. Интегрированы библиотеки для работы с PostgreSQL, аутентификация Clerk, D3.js и Python-скрипты. Интегрированный модуль включает в себя следующие функциональные компоненты [63].

Модуль для работы с базой данных PostgreSQL.

Подключение осуществляется через пул соединений (Pool) для управления запросами. Настройки БД читаются из внешнего файла settings.json. При старте для проверки подключения выполняется тестовый запрос SELECT NOW().

Функция reset() удаляет необходимые таблицы и создает их заново. Затем заполняет демо-данными (классы, подклассы, спектры) из CSV-файлов.

Основные операции для работы с БД:

- Загрузка спектров из CSV в БД.
- Конвертация CSV в бинарные данные (bytea).
- Связка данных с классами и подклассами.
- Получение спектра из БД по ID.
- Извлечение массива спектров из БД по списку ID.
- Получает подклассы из БД для анализа схожести.

Модуль обработки данных.

Функции для преобразования форматов:

- Чтение CSV-файла в массив отсчетов спектра.
- Конвертация бинарных данных из БД в массив float.
- Преобразование массива float в бинарный формат для сохранения в БД.

Вычисление меры схожести двух спектров с учетом фильтров (например, игнорирование артефактов в диапазоне 0-450 нм).

Нейросетевой расчет, сравнивающий спектр с эталонными из БД и рассчитывающий октановые числа на основе схожести. Возвращает результаты в формате, пригодном для визуализации.

Вызов Python-скриптов через spawn для ML-расчетов.

Модуль для работы с REST API.

Конечные точки:

- GET /oiltable возвращает список всех спектров с метаданными (класс, подкласс, октановое число).
- POST /check анализирует спектр по id (уже существующий в БД) или загруженному файлу (загрузка пользовательских данных).
- POST /oilupload загружает CSV-файл в БД.
- GET /reset сбрасывает БД и заполняет её демо-данными.
- GET /getfilters и GET /getmeasures возвращают списки фильтров и метрик для клиента.

Обработка файлов:

- POST /checkfromfile анализирует спектр из CSV-файла.
- POST /extractspectrumfromfile извлекает данные из CSV для предпросмотра.

Серверная часть имеет следующие технические особенности реализации:

- Асинхронность. Все функции используют async/await для работы с БД и файлами.
- Обработка ошибок. Проверка подключения к БД при старте, обработка ошибок JWT (просроченные токены, неверная подпись) и логирование ошибок в консоль.
- Производительность. Ресурсоемкие расчеты схожести спектров и обработка больших CSV-файлов вынесены в отдельную потоковую обработку.

4.3.2. Работа с модулем в интерфейсе ЭНС

Интерфейс ЭНС при работе с интегрированным модулем представляет собой одностраничное приложение (SPA), разработанное на React с интеграцией

современных библиотек и инструментов. Базовая архитектура строится на компонентном подходе, обеспечивающем модульность и повторное использование кода.

Структура интерфейса включает заголовок с логотипом системы, навигационными элементами и кнопкой управления пользовательским профилем. Основная область контента разделена на маршруты с использованием react-router-dom, где корневой маршрут (/) отображает таблицу химических продуктов (OilTable), а маршрут /90 – специализированный интерфейс для анализа октановых чисел (App90) [55].

Аутентификация реализована через сервис Clerk, обеспечивающий авторизацию через JWT-токены. Состояние SignedIn определяет доступ к функционалу системы: аутентифицированным пользователям предоставляется полный доступ к данным, тогда как SignedOut автоматически перенаправляет на страницу входа (RedirectToSignIn). Локализация интерфейса на русский язык повышает удобство использования.

Визуализация данных осуществляется через компонент OilTable, взаимодействующий с серверным API для загрузки спектральной информации. Таблица поддерживает динамическое обновление, сортировку и фильтрацию записей, что продемонстрировано на рисунке 46.

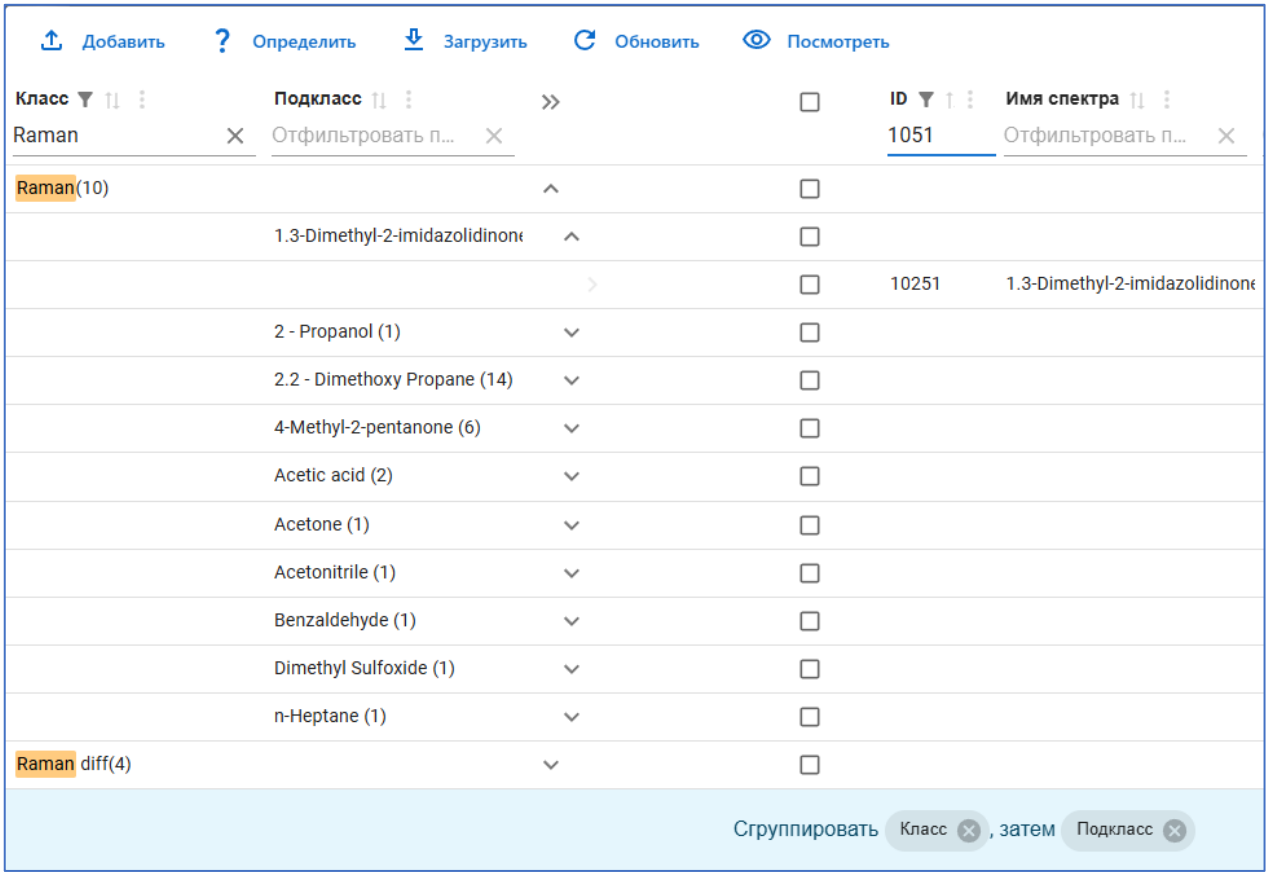


Рисунок 46 – Интерфейс таблицы химических продуктов

Для отображения графиков спектров используется библиотека, интегрированная через REST-узлы (/getchartdata), что позволяет отображать исследуемые спектральные данные, как показано на рисунке 47.

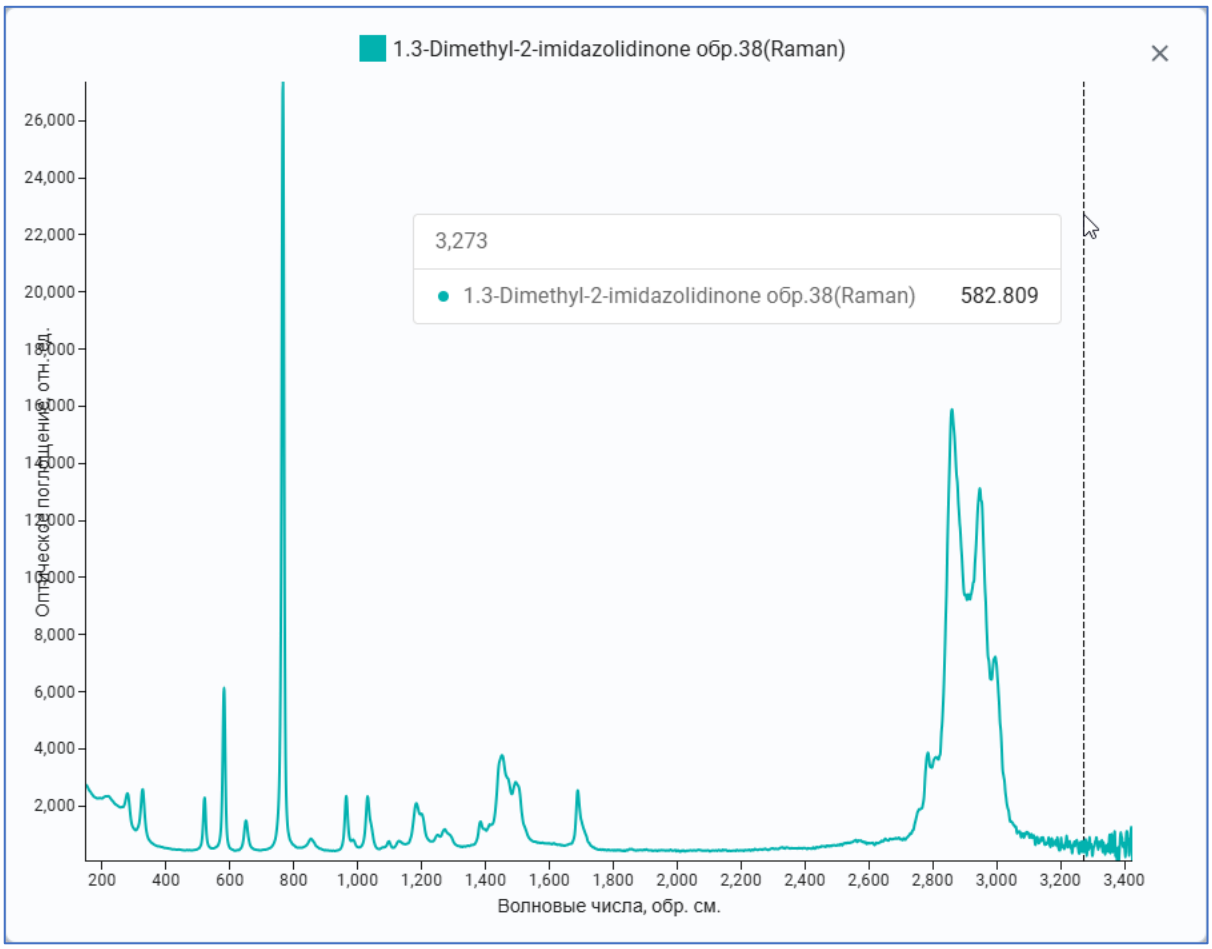


Рисунок 47 – Интерфейс просмотра спектров

Система уведомлений на базе notistack (SnackbarProvider) обеспечивает обратную связь с пользователем: успешная загрузка файлов, ошибки валидации или сетевые сбои отображаются в виде тостов с возможностью ручного закрытия. Адаптивный дизайн, реализованный через CSS-модули и компоненты Material-UI Joy, гарантирует корректное отображение на различных устройствах, включая мобильные платформы.

Взаимодействие с сервером построено на асинхронных HTTP-запросах. Например, загрузка CSV-файлов через форму (/oilupload) включает парсинг данных на клиенте с последующей отправкой бинарного представления на сервер. Анализ

спектров (/checkfromfile) выполняется через отправку файла и получение результатов в формате JSON для визуализации в интерфейсе.

Обработка состояний реализована через хуки React (useState, useEffect), обеспечивающие реактивность интерфейса при изменении данных. Референсы (useRef) используются для управления DOM-элементами, такими как таблицы или поля ввода. Оптимизация производительности достигается за счёт кэширования вычислений и минимизации перерисовок компонентов.

Безопасность обеспечивается валидацией токенов через middleware Clerk и ограничением доступа к критическим операциям (например, сброс БД через /reset). Все запросы к API сопровождаются проверкой JWT, что предотвращает несанкционированный доступ.

Таким образом, интерфейс сочетает современные подходы к разработке интерфейсов, строгую типизацию данных и глубокую интеграцию с серверными сервисами, обеспечивая эффективное решение задач анализа спектров химических продуктов.

Пример полного цикла пользовательского взаимодействия с программой:

- Пользователь загружает CSV-файл с новым спектром.
- Клиент отправляет файл на /oilupload.
- Сервер сохраняет данные в БД и возвращает ID новой записи.
- Пользователь выбирает спектр для анализа, клиент вызывает /check с его ID.
- Сервер вычисляет схожесть с эталонами, возвращает JSON с результатами.
- Клиент отрисовывает график через D3.js и обновляет таблицу.
- При попытке несанкционированного доступа к /reset middleware блокирует запрос и перенаправляет на вход.

Такая архитектура обеспечивает высокую отказоустойчивость, минимальные задержки и соответствие требованиям безопасности.

4.4. Информационно-функциональная модель системы

На представленной на рисунке 48 схеме изображена информационно-функциональная модель основного процесса кластеризации спектральных данных в нотации IDEF0. Функция кластеризации спектральных данных (блок A0) отражает ключевой этап обработки и анализа спектральной информации.

Входами процесса служат собственно спектральные данные, включая спектральные данные неизвестного спектра, которые должны быть интегрированы в модель для последующей классификации или сравнения. Эти данные поступают на обработку, где подвергаются трансформациям и кластерному разбиению.

Управляющие воздействия представлены методами сбора данных, алгоритмами предварительной обработки, методами сжатия информации, стандартами обучения и методами мониторинга выполнения кластеризации. Они задают правила, процедуры и ограничения, в рамках которых осуществляется обработка: обеспечивают корректность измерений, формируют требования к качеству спектров, определяют алгоритмические подходы и контроль качества работы системы.

Механизмы реализации процесса включают аналитика спектральных данных, программное обеспечение, алгоритмы обработки спектральных данных, методы снижения размерности и методы кластеризации. Эти ресурсы обеспечивают функционирование системы, позволяя применять соответствующие математические и программные инструменты, а также экспертную оценку при необходимости.

Выходом процесса являются результаты кластеризации, то есть структурированное представление спектральных данных в виде групп и классов, что облегчает их дальнейшую интерпретацию и использование для идентификации веществ, обнаружения аномалий или построения диагностических моделей.

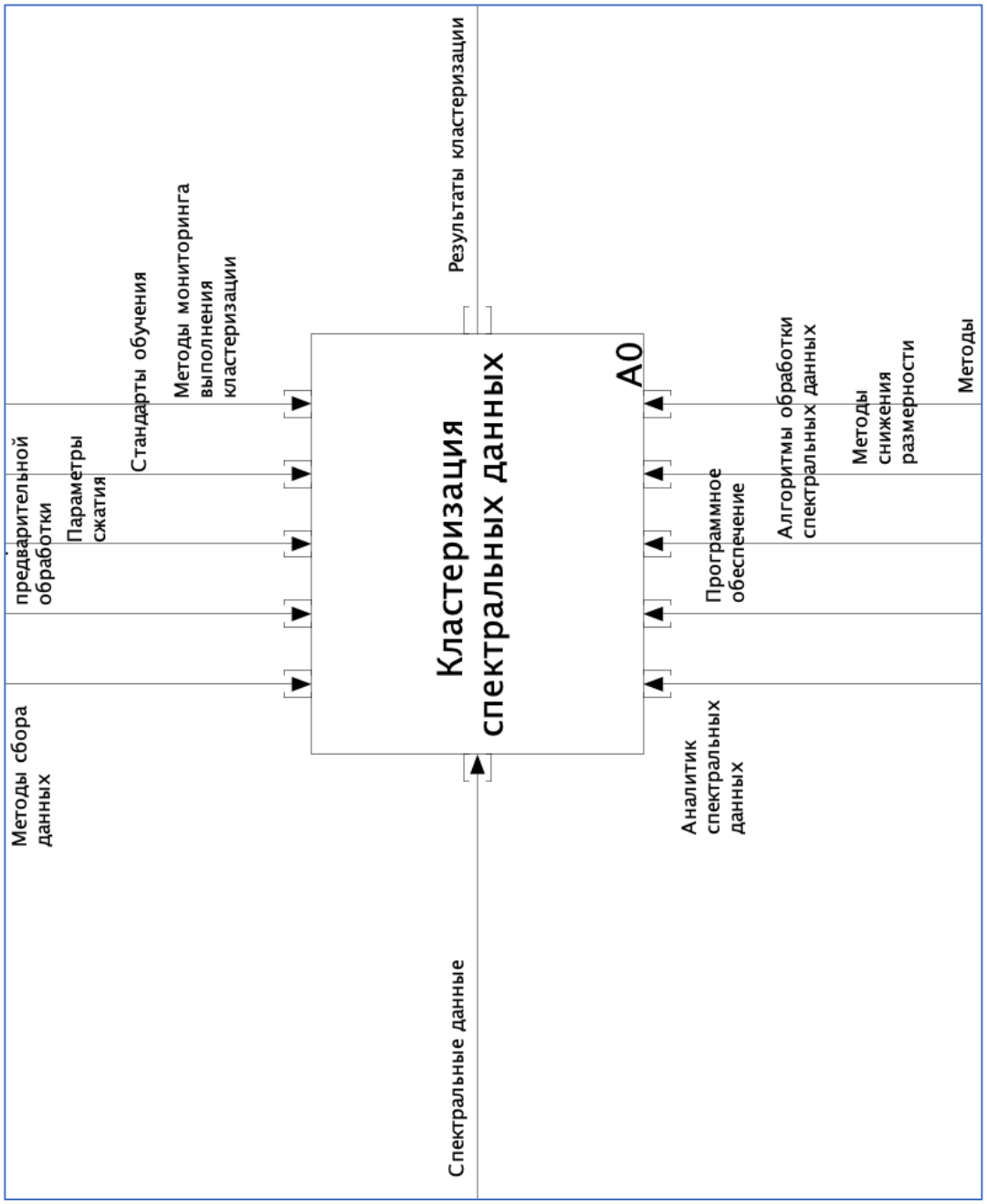


Рисунок 48 – Верхний уровень модели

На рисунке 49 показана декомпозиция основного процесса кластеризации спектральных данных в нотации IDEF0. Центральная функция разложена на шесть подпроцессов (A1-A6), которые последовательно отражают этапы обработки, подготовки и анализа данных.

На первом этапе осуществляется получение исходных спектральных данных (блок A1). Входом являются спектры, полученные из экспериментальных установок, а управление задаётся методами сбора данных, которые определяют протоколы и процедуры регистрации. Выходом служит собранный набор спектральных данных, готовых к дальнейшей обработке. Важную роль на этом этапе играет аналитик, который контролирует корректность выполнения измерений.

Полученные спектральные данные подвергаются нормализации, фильтрации и другим видам предобработки (блок A2). Алгоритмы предварительной обработки служат управляющим воздействием, определяющим используемые методы коррекции шумов, сглаживания или центрирования данных. На выходе формируются обработанные спектральные данные, которые уже содержат выделенные и очищенные признаки, пригодные для дальнейших шагов.

Следующий этап связан с уменьшением размерности данных (блок A3). Входом являются обработанные спектры, а управляющим фактором выступают параметры сжатия. На выходе формируются компактные представления спектров, в которых сохранена структурная информация, необходимая для эффективной кластеризации. На этом этапе также используются метрики расстояния, которые задают способ измерения близости объектов. Механизмами процесса являются методы снижения размерности и аналитик.

Подпроцесс обучения эксперта (блок A4) ориентирован на формирование у специалиста навыков и знаний для корректной интерпретации результатов кластеризации. На вход поступают обработанные данные и необученный эксперт.

Процесс выполняется с помощью программного обеспечения и аналитика, а контролируется стандартами обучения. В результате необученный эксперт превращается в обученного специалиста, способного контролировать и корректировать процесс, а также оценивать качество полученных кластеров.

Проведение кластеризации – это центральный шаг, на котором применяются выбранные методы кластеризации к сжатым данным (блок А5). Управляющие воздействия задаются методами мониторинга выполнения кластеризации. Входом являются сжатые данные. Выходом – результаты кластеризации, то есть распределение спектров по группам в соответствии с выявленными закономерностями. Процесс выполняется с помощью обученного эксперта и методов кластеризации.

На завершающем этапе осуществляется проверка качества и интерпретация полученных результатов (блок А6). Входом служат результаты кластеризации, а управляющими воздействиями – критерии оценки кластеризации. Процесс так же выполняется с помощью обученного эксперта и методов кластеризации. Выходом являются отчёты, выводы об анализируемых объектах и структурированные данные, которые могут быть использованы в научных исследованиях или прикладных задачах.

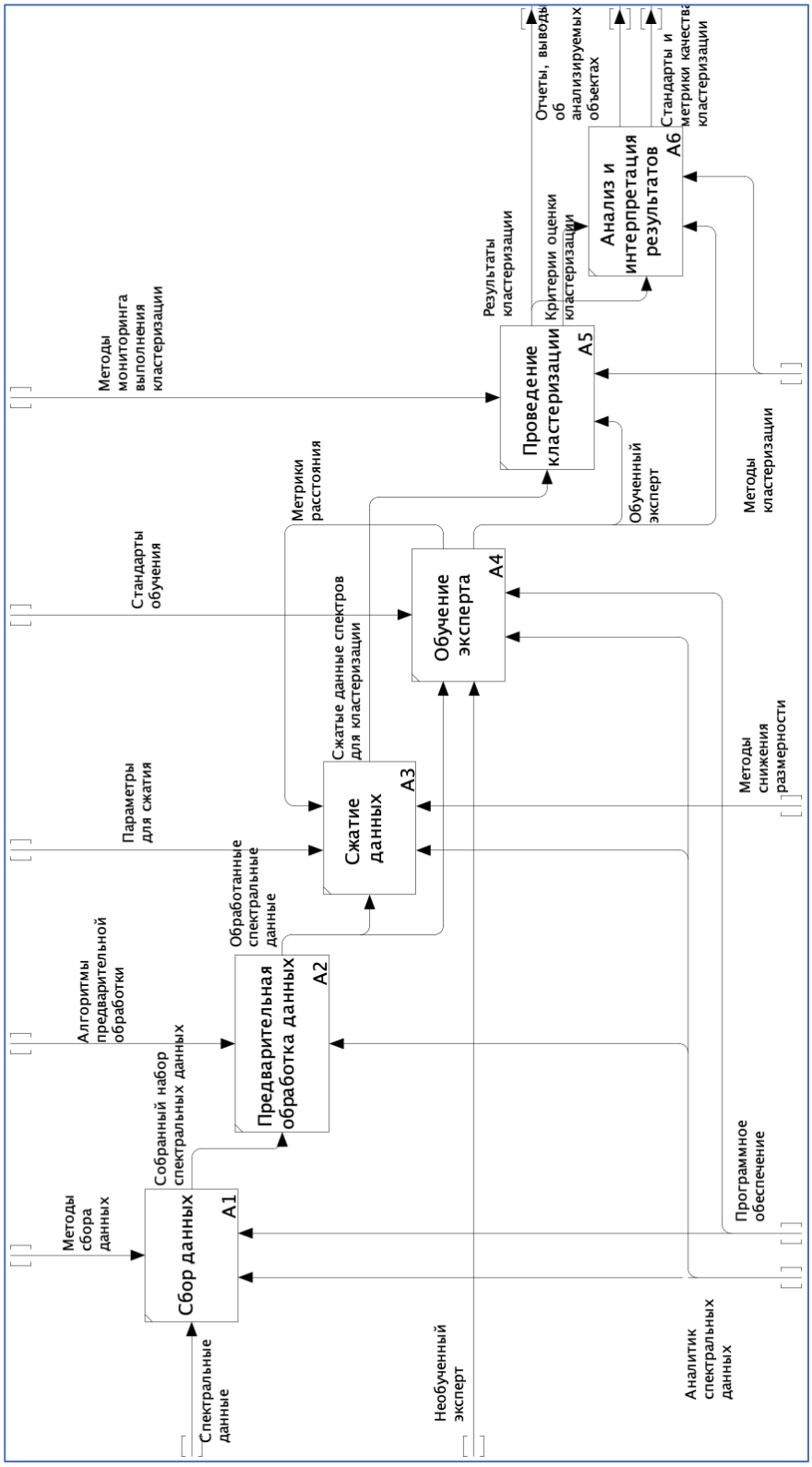


Рисунок 49 – Второй уровень модели

4.5. Описание обучающей программы спектрального анализа

В ходе работы была разработана программная система для анализа спектральных данных, включающая механизмы генерации синтетических данных, их обработки и визуализации с использованием методов кластеризации и уменьшения размерности. В основе системы лежит модульный подход, позволяющий пользователю выбирать различные рассмотренные методы обработки, алгоритмы снижения размерности, а также методы кластеризации. Программа разработана с графическим интерфейсом на основе PyQt6, что делает её удобной для образовательных целей и научных исследований [64; 102].

Система позволяет пользователю как загружать спектральные данные, так и генерировать их синтетически с помощью встроенных алгоритмов. Это дает возможность тестировать различные методы анализа на контролируемых примерах. Важной составляющей программы является модуль предобработки данных, который позволяет выделять ключевые особенности спектральных данных перед применением методов машинного обучения.

Для анализа многомерных данных предусмотрены описанные выше алгоритмы уменьшения размерности. Эти методы позволяют представить исходные данные в двух- или трёхмерном пространстве, что облегчает их интерпретацию и последующую кластеризацию. Процесс кластеризации реализован с использованием описанных выше алгоритмов, каждый из которых имеет свои особенности и области применения.

Результаты обработки и анализа представлены в удобной визуальной форме, продемонстрированной на рисунках ниже. Графический интерфейс обеспечивает интуитивное управление параметрами алгоритмов, а также позволяет пользователю в

реальном времени анализировать влияние различных подходов на конечный результат. Метрики оценки качества кластеризации помогают количественно оценивать эффективность разбиения данных на группы.

На рисунке 46 показан интерфейс этапа генерации и преобразования данных. Здесь пользователь выбрал открытую базу данных рамановских спектров химических соединений активных фармацевтических ингредиентов (API), снижение размерности этих данных алгоритмом UMAP с выбранной метрикой Канберры и определенными другими параметрами алгоритма. Ниже описана справка об алгоритме и его параметрах, а еще ниже результаты оценки полученного пространства с помощью индекса Дэвиса-Боулдина, коэффициента силуэта кластера и список точек, отображающих спектры с указанием индекса их класса и координат для более точного исследования. Справа отображены полученные точки во вложенном пространстве.

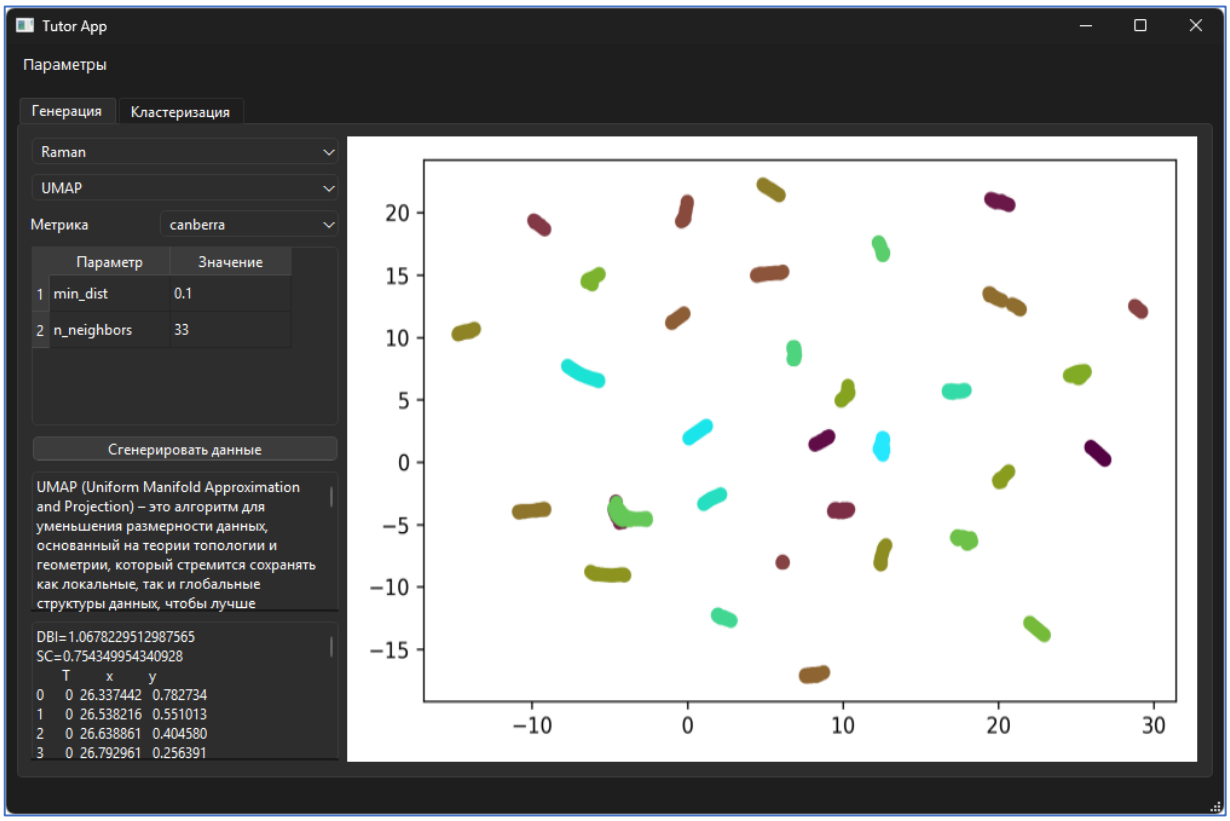


Рисунок 50 – Интерфейс обучающей программы, генерация данных

На рисунке 51 показан интерфейс этапа кластеризации данных. Это окно идентично предыдущему, но в результате указана оценка кластеризации с помощью скорректированного индекса Рэнда, а к списку точек добавлен столбец класса, определенного алгоритмом кластеризации.

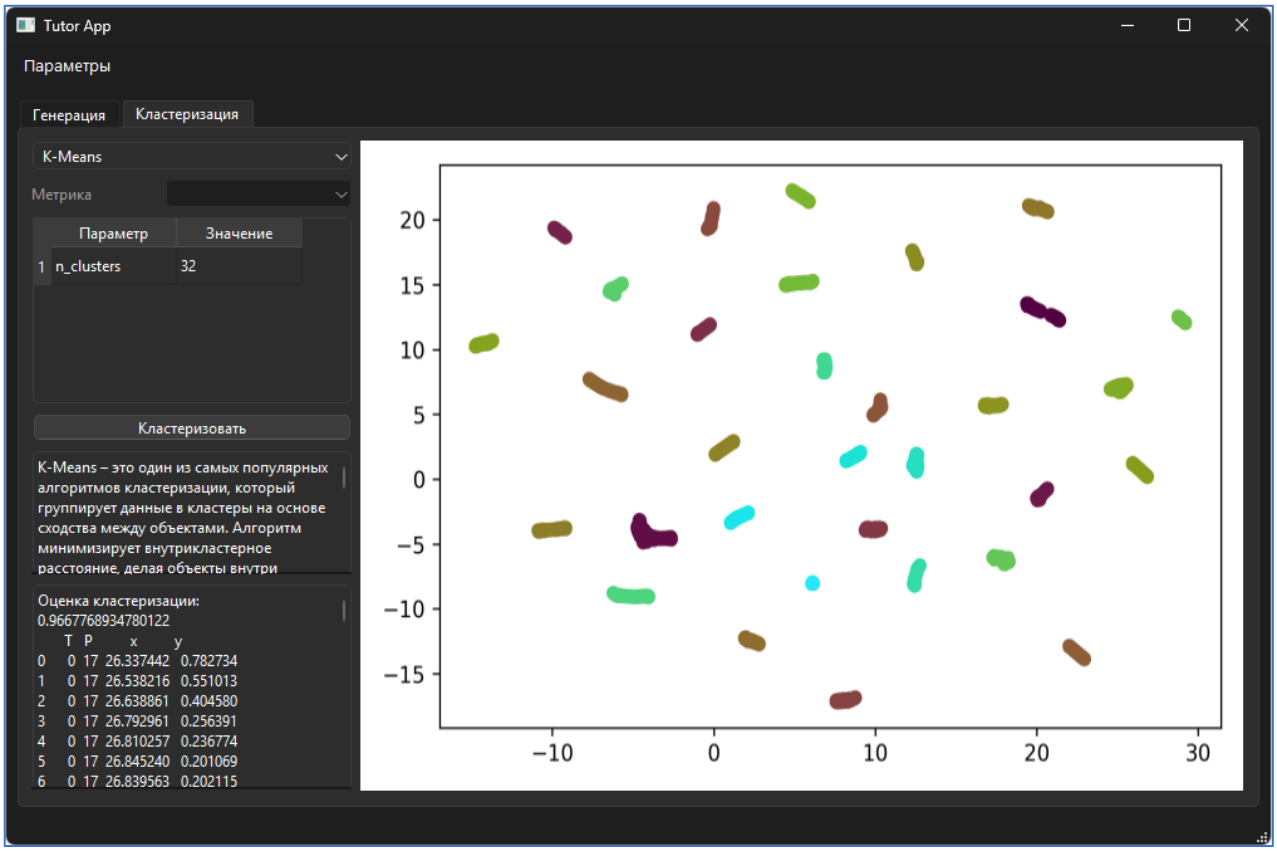


Рисунок 51 – Интерфейс обучающей программы, кластеризация данных

С точки зрения программной реализации, система написана на языке Python и использует библиотеки NumPy, SciPy, scikit-learn и PyQt6. Предобработка данных осуществляется в виде отдельных модулей, что позволяет легко изменять или добавлять новые методы обработки. Графический интерфейс построен на основе архитектуры Model-View-Controller (MVC). Отображение данных осуществляется с

помощью библиотеки Matplotlib, а интерактивные элементы управления параметрами алгоритмов интегрированы через виджеты PyQt6.

Процесс выполнения анализа данных организован в виде отдельных классов, отвечающих за генерацию данных, их обработку, кластеризацию и визуализацию. Важной особенностью является поддержка многопоточности, что обеспечивает плавное выполнение вычислительных задач даже при работе с большими наборами данных. Встроенная система логирования позволяет отслеживать ход выполнения операций и анализировать возникающие ошибки.

Программа разработана с целью предоставления удобного инструмента для студентов и исследователей в области машинного обучения. Она позволяет изучать принципы работы алгоритмов кластеризации и уменьшения размерности, а также применять их к реальным данным. За счёт интерактивности и визуализации результатов программа облегчает освоение ключевых концепций анализа данных.

4.6. Анализ производительности метода

Были также проведены измерения времени выполнения задачи с фиксированным параметром случайности. Измерения времени выполнения каждого метода повторялись по 10 раз, а относительное стандартное отклонение не превышало 0.03. На наборах данных бензинов построение модели предложенной методикой занимает в среднем 0.6535 секунд, в то время как метод PLS-DA выполняет классификацию за 0.2373 секунд, а LDA – за 0.0895 секунд. Это означает, что по абсолютным значениям предложенная методика медленнее упомянутых аналогов (он приблизительно в 2,75 раза медленнее PLS-DA и в 7.3 раза медленнее LDA), однако все три алгоритма работают в одном порядке величин времени выполнения на данных малого объёма,

что позволяет говорить о практической сопоставимости их быстродействия при анализе небольших выборок и о возможности использования в поточных анализаторах.

На больших наборах рамановских спектров распределение времён иное: предложенный алгоритм выполняется в среднем за 26.0698 секунд, PLS-DA – за 96.5153 секунд, а LDA – за 2.7267 секунды. В этой ситуации предложенный алгоритм оказывается существенно быстрее PLS-DA (примерно в 3.7 раза), но заметно медленнее LDA (примерно в 9.6 раза). Из этого следует, что при масштабировании на большие объёмы спектральных данных разработанный метод демонстрирует лучшую масштабируемость и производительность по сравнению с PLS-DA, сохраняя приемлемое время обработки, тогда как LDA остаётся наиболее быстрым по вычислениям, хоть и проигрывает в точности выполнения задачи. В практическом плане полученные результаты показывают, что предложенная методика является конкурентоспособным на небольших данных, что важно для задач непрерывного поточного анализа спектров, и обладает преимуществом по скорости перед PLS-DA при работе с большими объёмами.

4.7. Выводы по четвертой главе

В четвертой главе представлены результаты разработки и реализации программного обеспечения, обеспечивающего системную интеграцию методов обработки спектральных данных, их анализа и визуализации. Описано создание базы данных, архитектуры прикладного комплекса и пользовательского интерфейса, ориентированного как на производственные задачи, так и на образовательные и исследовательские применения.

Построенная база данных, реализованная на платформе PostgreSQL, обеспечивает целостное хранение исходных спектров, их классов и подклассов, а также связанных с ними моделей снижения размерности и кластеризации. Такое решение устраняет проблему разрозненности данных и формирует единое пространство для дальнейшего анализа.

Разработанный программный модуль на языке Python обеспечивает тесную интеграцию с базой данных и автоматизацию всего цикла обработки: от загрузки спектров и формирования моделей снижения размерности до сохранения результатов кластеризации и расчёта метрик качества. Внедрение асинхронных операций при работе с PostgreSQL и реализация визуализации позволяют повысить производительность и обеспечить наглядность анализа. Системный подход проявляется также в обеспечении возможности повторного использования обученных моделей, что снижает издержки на вычислительные ресурсы и повышает эффективность эксплуатации.

Серверная часть системы, реализованная на Node.js, дополнительно расширяет функциональность комплекса, обеспечивая взаимодействие с клиентской частью через REST API, управление аутентификацией и защиту данных. Использование асинхронных механизмов, кэширования и разгрузки ресурсоемких вычислений в отдельные процессы подтверждает соответствие решения современным принципам построения отказоустойчивых и масштабируемых информационных систем.

Клиентская часть, реализованная на React, обеспечивает удобный интерфейс для работы с базой данных и аналитическими модулями. Возможность загрузки файлов, динамической визуализации спектров и отображения результатов кластеризации в графической форме снимает одну из ключевых проблем, обозначенных ранее, а именно – недостаток инструментов визуализации для экспертной интерпретации результатов машинного обучения. Таким образом, система не только автоматизирует

обработку данных, но и формирует основу для доверия пользователей к результатам интеллектуального анализа.

Также была разработана обучающая программа, построенная на PyQt6 и предназначенная для генерации синтетических данных, их обработки и визуализации. Она позволяет студентам и исследователям осваивать принципы применения методов снижения размерности и кластеризации в условиях, приближенных к реальным. Такой инструмент решает задачу формирования компетенций и способствует широкому внедрению разработанных подходов в образовательный и исследовательский процесс.

Проведенное экономическое обоснование, представленное в приложении А, подтверждает высокую эффективность внедрения разработанного комплекса. Показано, что повышение точности идентификации спектральных данных и снижение числа ошибок классификации ведет к значительной экономии ресурсов и сокращению объёмов некондиционной продукции. Рассчитанные показатели окупаемости демонстрируют, что система обладает потенциалом быстрой адаптации в промышленности и приносит ощутимую выгоду.

Свидетельства о государственной регистрации базы данных, программного модуля ЭНС и обучающей программы представлены в приложениях Б, В и Г соответственно.

Основные выводы и результаты работы

В ходе выполненной работы достигнуты следующие результаты, соответствующие поставленным задачам:

1. Были исследованы существующие подходы к обработке и интерпретации спектральных данных, включая традиционные хемометрические методы и современные алгоритмы машинного обучения. Систематизация и сопоставление различных подходов позволили выявить их достоинства и ограничения, определить критерии применимости в зависимости от структуры и природы данных. В результате был сформирован обоснованный выбор наиболее перспективных подходов, которые обеспечивают достоверность и воспроизводимость распознавания жидких сред и могут быть положены в основу дальнейшей разработки методов кластеризации.

2. Разработана методика формирования цифровых образов жидких сред на основе их спектральных характеристик и метод кластеризации спектральных данных по этим образам. Проведён анализ полученных методик, который показал, что наибольшую эффективность обеспечивают алгоритмы, сочетающие предобработку сигналов дискретной свёрткой и дискретной разницей с последующим применением модифицированных нелинейных методов снижения размерности t-SNE и UMAP.

3. Доказано, что формирование цифровых образов позволяет выделить уникальные признаки жидких сред, повышающие точность их кластеризации. Разработанные решения прошли проверку на реальных наборах данных, показавшую повышение точности кластеризации с 75% до 99.7%. Показана устойчивость предложенной методики при работе с разнотипными спектрами: ИК и рамановскими. Обоснована достоверность полученных результатов с использованием количественных метрик качества кластеризации. Установлено, что предложенные

решения демонстрируют переносимость и высокую точность при анализе как небольших, так и крупных наборов спектров.

4. Разработано программное обеспечение, реализующее полный цикл обработки спектральных данных: от предобработки и снижения размерности до кластеризации и визуализации результатов. Программный комплекс обладает модульной архитектурой, поддерживает выбор различных методов и параметров, а также интегрируется с внешними системами анализа.

5. Сформирована база данных цифровых образов жидких сред, включающая результаты, полученные различными методами обработки и снижения размерности. База данных позволяет проводить сравнительный анализ подходов, а также служит ресурсом для последующего обучения алгоритмов машинного обучения.

6. Разработанные методы и программные решения интегрированы в существующую экспертно-нейросетевую систему. Показана их практическая применимость: обеспечено повышение точности и скорости автоматической классификации спектров, расширены возможности системы для работы с большими объёмами данных и повышена её универсальность за счёт поддержки различных типов жидких сред.

В совокупности результаты выполненной работы демонстрируют, что предложенная методика формирования цифровых образов жидких сред, разработанное программное обеспечение и созданная база данных этих образов образуют интегрированную систему, способную эффективно анализировать спектральные данные химических соединений. Достижения исследования способствуют повышению достоверности оценки качества свойств продукции, что имеет стратегическое значение для развития химической промышленности и повышения её конкурентоспособности на мировом рынке. Акты о внедрении результатов исследования приведены в приложениях Д, Е, Ж, И, К и Л.

Список сокращений и условных обозначений

1.	АИ	Автомобильный, исследовательский
2.	БД	База данных
3.	БИК	Ближний инфракрасный
4.	ДНК	Дезоксирибонуклеиновая кислота
5.	ИИ	Искусственный интеллект
6.	ИК	Инфракрасный
7.	ККЛ	Квантово-каскадный лазера
8.	ППР	Поверхностный плазмонный резонанс
9.	ППУ	Пенополиуретана
10.	РФ	Российская Федерация
11.	СР	Снижение размерности
12.	УФ	Ультрафиолет
13.	ЭВМ	Электронно-вычислительная машина
14.	ЭНС	Экспертно-нейросетевая система
15.	API	Application Programming Interface
16.	API	Active Pharmaceutical Ingredient
17.	ARI	Adjusted Rand index
18.	CNN	Convolutional neural network
19.	CSS	Cascading Style Sheets
20.	CSV	Comma-Separated Values
21.	DA	Discriminant analysis
22.	DBI	Davies-Bouldin score
23.	DBSCAN	Density-Based Spatial Clustering of Applications with Noise

24.	DNN	Deep Neural Network
25.	DOM	Document Object Model
26.	HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
27.	HTTP	Hypertext Transfer Protocol
28.	ISOMAP	Isometric Mapping
29.	JSON	JavaScript Object Notation
30.	JWT	JSON Web Token
31.	LDA	Latent Dirichlet allocation
32.	LLE	Locally Linear Embedding
33.	LTSA	Local Tangent Space Alignment
34.	MDS	Multi-dimensional Scaling
35.	ML	Machine Learning
36.	MLLE	Modified Locally Linear Embedding
37.	MSC	Multiplicative Scatter Correction
38.	MVC	Model-View-Controller
39.	NIR	Near-infrared spectroscopy
40.	OCC	One-class classification
41.	OPTICS	Ordering Points To Identify the Clustering Structure
42.	PCA	Principal component analysis
43.	PLS	Projection to Latent Structures
44.	PLS-DA	Partial Least Squares - Discriminant Analysis
45.	POPOP	1,4-bis(5-phenyloxazol-2-yl) benzene
46.	PPO	Polyphenylene Oxide
47.	QDA	Qualitative Data Analysis
48.	REST	Representational State Transfer

49.	RNN	Recurrent neural network
50.	ROC	Receiver operating characteristic
51.	SC	Silhouette Coefficient
52.	SE	Spectral Embedding
53.	SHAP	SHapley Additive exPlanations
54.	SIMCA	Soft Independent Modeling of Class Analogies
55.	SNV	Standard Normal Variate
56.	SPA	Single-page application
57.	SQL	Structured Query Language
58.	SVM	Support vector machine
59.	t-SNE	t-distributed stochastic neighbor embedding
60.	UI	User interface
61.	UMAP	Uniform Manifold Approximation and Projection
62.	VIP	Variable Importance in the Projection

Список литературы

1. Алгасов А. С. Количественный анализ рентгеноспектральных данных для смеси соединений методами машинного обучения / А. С. Алгасов, С. А. Гуда, А. А. Гуда [и др.] // Поверхность. Рентгеновские, синхротронные и нейтронные исследования. – 2021. – № 5. – С. 95-101.

2. Апяри, В. В. Аналитические возможности цифровых цветометрических технологий. Определение нитрит-ионов с использованием пенополиуретана / В. В. Апяри, С. Г. Дмитриенко, Ю. А. Золотов // Вестник Московского университета. Серия 2: Химия. – 2011. – Т. 52, № 1. – С. 36-42. – EDN OGJYMT. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=16925586 (дата обращения: 12.08.2025).

3. Апяри, В. В. Новые подходы в анализе методами оптической молекулярной абсорбционной спектроскопии с использованием гетерогенных аналитических систем : специальность 02.00.02 «Аналитическая химия» : автореферат диссертации на соискание ученой степени доктора химических наук / Апяри Владимир Владимирович. – Москва, 2016. – 22 с. – EDN ZQCVAP. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=30439538 (дата обращения: 12.08.2025).

4. Апяри, В. В. Применение цифрового фотоаппарата и компьютерной обработки данных для определения органических веществ с использованием диазотированного пенополиуретана / В. В. Апяри, С. Г. Дмитриенко // Журнал аналитической химии. – 2008. – Т. 63, № 6. – С. 581-588. – EDN IKPTON. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=9989763 (дата обращения: 12.08.2025).

5. Архипова, В. В. Определение полигексаметиленгуанидина гидрохлорида с использованием наночастиц золота и пенополиуретана / В. В. Архипова, В. В. Апяри, С. Г. Дмитриенко // Вестник Московского университета. Серия 2: Химия. – 2015. – Т.

56, № 1. – С. 34-40. – EDN TKDXDV. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=23038428 (дата обращения: 12.08.2025).

6. Битюков, В. К. Автоматизация контроля качества гомогенизации молока методами ультразвуковой спектроскопии / В. К. Битюков, А. А. Хвостов, Д. И. Ребриков, В. Е. Мерзликин // Вестник Воронежского государственного университета инженерных технологий. – 2015. – № 1(63). – С. 74-81. – EDN TTUFCT. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=23478366 (дата обращения: 12.08.2025).

7. Вагин, В. А. Быстрые методы снижения размерности спектральных данных для их образной визуализации / В. А. Вагин, А. Е. Краснов, Д. Н. Никольский // Журнал прикладной спектроскопии. – 2019. – Т. 86, № 1. – С. 116-121. – EDN MJUJLN. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=36761800 (дата обращения: 12.08.2025).

8. Вагин, В. А. Снижение размерности спектральных данных в фурье-спектроскопии / В. А. Вагин, А. Е. Краснов, Д. Н. Никольский // Акустооптические и радиолокационные методы измерений и обработки информации : Материалы 12-й Международной научно-технической конференции, Москва, 13–16 октября 2019 года. – Москва: Научно-технологический центр уникального приборостроения РАН, 2019. – С. 76-80. – EDN DBCEIS. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=41433088 (дата обращения: 12.08.2025).

9. Вагин, В. А. Современная фурье-спектроскопия и быстрый нейроподобный метод снижения размерности спектральных данных / В. А. Вагин, А. Е. Краснов // Физические основы приборостроения. – 2020. – Т. 9, № 3(37). – С. 86-91. – DOI 10.25210/jfop-2003-086091. – EDN OTBXOC. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=44126989 (дата обращения: 12.08.2025).

10. Васильев Н. С. Алгоритм идентификации веществ по конечному набору спектров вторичного излучения / Н. С. Васильев, Ил. С. Голяк, А. Н. Морозов // Оптика и спектроскопия. – 2015. – Т. 118. – № 1. – С. 157-162.

11. Влияние снижения выхода светлых на 1% на НПЗ: потери и оптимизация процессов. – URL: https://inner.su/articles/vliyanie-snizheniya-vykhoda-svetlykh-na-1-na-npz-poteri-i-optimizatsiya-protsessov/ (дата обращения: 14.09.2025). – Текст : электронный.

12. Гаврилов, Д. А. Алгоритм распознавания веществ в системах машинного зрения / Д. А. Гаврилов // Естественные и технические науки. – 2011. – № 2(52). – С. 376-379. – EDN MITEYY. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=16285513 (дата обращения: 12.08.2025).

13. Голяк И. С. Идентификация химических соединений по спектрам рассеянного излучения в диапазоне длин волн 5.3–12.8 мкм с применением перестраиваемого квантово-каскадного лазера / И. С. Голяк, А. Н. Морозов, С. И. Светличный [и др.] // Химическая физика. – 2019. – Т. 38. – № 7. – С. 3-10.

14. Горбунова, М. В. Сорбция наностержней золота на пенополиуретане как способ получения нанокомпозитного материала с поверхностным плазмонным резонансом для целей химического анализа / М. В. Горбунова, М. А. Матвеева, В. В. Апяри [и др.] // Российские нанотехнологии. – 2017. – Т. 12, № 3-4. – С. 57-63. – EDN QGVXUD. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=32204840 (дата обращения: 12.08.2025).

15. Гусев, К. В. Автоматизация контроля качества нефтепродуктов для обеспечения эффективного управления технологическим процессом: специальность 2.3.3 «Автоматизация и управление технологическими процессами и производствами» : диссертация на соискание ученой степени доктора технических наук / Гусев Кирилл Вячеславович. – Воронеж, 2025. – 152 с.

16. Данчук, А. И. Концентрирование и определение некоторых ионов тяжелых металлов с применением нетканых материалов и мицеллярно-насыщенных фаз ПАВ : специальность 02.00.02 «Аналитическая химия» : диссертация на соискание ученой степени кандидата химических наук / Данчук Александра Ильинична, 2018. – 133 с. – EDN NKWCEB. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=54457460 (дата обращения: 12.08.2025).

17. Дмитриенко, С. Г. Использование реакций диазотирования и азосочетания с участием пенополиуретана для определения нитрит-ионов с помощью спектроскопии диффузного отражения и цветометрических сканер-технологий / С. Г. Дмитриенко, В. В. Апяри, О. А. Свиридова [и др.] // Вестник Московского университета. Серия 2: Химия. – 2004. – Т. 45, № 2. – С. 131-138. – EDN HBOCUF. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=8387455 (дата обращения: 12.08.2025).

18. Емельянов, О. Э. Неразрушающий контроль нестероидных противовоспалительных средств методом ИК-спектроскопии в ближней области / О. Э. Емельянов, В. Г. Амелин, А. В. Третьяков // Известия Саратовского университета. Новая серия. Серия: Химия. Биология. Экология. – 2024. – Т. 24, № 2. – С. 135-143. – DOI 10.18500/1816-9775-2024-24-2-135-143. – EDN LIZBKL. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=67325582 (дата обращения: 12.08.2025).

19. Ерохин, С. Д. Анализ существующих методов снижения размерности входных данных / С. Д. Ерохин, Б. Б. Борисенко, И. Д. Мартишин, А. С. Фадеев // T-Comm: Телекоммуникации и транспорт. – 2022. – Т. 16, № 1. – С. 30-37. – DOI 10.36724/2072-8735-2022-16-1-30-37. – EDN LEHFTU. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=47809227 (дата обращения: 12.08.2025).

20. Журавлева, В. В. Упрощенный показатель силуэта для определения качества кластерных структур / В. В. Журавлева, А. С. Маничева // Известия Алтайского государственного университета. – 2022. – № 4(126). – С. 110-114. – DOI

10.14258/izvasu(2022)4-17. – EDN QKUSRP. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=49423339 (дата обращения: 12.08.2025).

21. Зайцева, В. В. Методы лазерной спектроскопии: краткая характеристика и области применения / В. В. Зайцева, В. С. Михайленко, М. А. Кича // Вестник МАНЭБ. – 2021. – Т. 26, № 3. – С. 61-67. – EDN MJDIWQ. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=47396363 (дата обращения: 12.08.2025).

22. Зайцева, В. В. Основы и современные аспекты спектроскопии комбинационного рассеяния / В. В. Зайцева, В. С. Михайленко, М. А. Кича // Вестник МАНЭБ. – 2021. – Т. 26, № 3. – С. 7-12. – EDN QNNGLF. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=47396354 (дата обращения: 12.08.2025).

23. Исаченко А. И. Определение цистеина методом спектроскопии диффузного отражения по его влиянию на формирование нанокомпозитов золота на основе пенополиуретана / А. И. Исаченко, В. В. Апяри, П. А. Волков [и др.] // Журнал аналитической химии. – 2020. – Т. 75. – № 7. – С. 629-635.

24. Кацков, Д. А. Одновременное определение элементов в атомно-абсорбционной спектрометрии с электротермической атомизацией / Д. А. Кацков // Заводская лаборатория. Диагностика материалов. – 2019. – Т. 85, № 1-1. – С. 5-17. – DOI 10.26896/1028-6861-2019-85-1-I-5-17. – EDN YWVSDJ. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=36948481 (дата обращения: 12.08.2025).

25. Косолапов, Ю. В. Численный метод различения спектральных данных в ИК-области для идентификации гретых пищевых жиров / Ю. В. Косолапов, С. А. Красников, Т. В. Шленская, Е. В. Грузинов // Хранение и переработка сельхозсырья. – 2007. – № 3. – С. 56-57. – EDN HZHGSV. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=9472161 (дата обращения: 12.08.2025).

26. Кочиков, И. В. Алгоритмы идентификации веществ по ИК-спектрам в базе данных гибридного типа по молекулярным спектральным постоянным (ИСМОЛ) / И.

В. Кочиков, Г. М. Курамшина, Л. М. Самков [и др.] // Вычислительные методы и программирование. – 2007. – Т. 8, № 4. – С. 70-73. – EDN IJUQPL. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=9952653 (дата обращения: 12.08.2025).

27. Красников, С. А. Визуализация больших данных в виде многомерных векторов на плоскость / С. А. Красников, М. А. Овчинникова, К. В. Гусев // Научно-технический вестник Поволжья. – 2024. – № 1. – С. 155-158. – EDN IQSAUL. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=60398122 (дата обращения: 12.08.2025).

28. Красников, С. А. Интеллектуальная компьютерная квалиметрия бензинов по инфракрасным спектрам / С. А. Красников, С. В. Николаева, А. Е. Краснов, А. С. Мясоедов // Естественные и технические науки. – 2019. – № 12(138). – С. 307-311. – EDN EOTNRQ. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=41854340 (дата обращения: 12.08.2025).

29. Красников, С. А. Метод сжатия и визуализации многомерных данных / С. А. Красников, Е. А. Чернов. – Москва : ООО "Издательство «Спутник+», 2022. – 115 с. – ISBN 978-5-9973-6391-8. – EDN GJBNLB. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=49312961 (дата обращения: 24.09.2025).

30. Красников, С. А. Метод спектральной компьютерной квалиметрии / С. А. Красников, С. В. Николаева, К. В. Гусев, М. А. Овчинников // Научно-технический вестник Поволжья. – 2023. – № 4. – С. 143-146. – EDN PXIAYD. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=50750223 (дата обращения: 24.09.2025).

31. Красников, С. А. Методология построения систем контроля качества жидких сред по спектральным характеристикам : специальность 05.13.01 «Системный анализ, управление и обработка информации (по отраслям)» : диссертация на соискание

ученой степени доктора технических наук / Красников Степан Альбертович. – Владимир, 2012. – 244 с. – EDN QFNDLV.

32. Краснов, А. Е. Информационные технологии пищевых производств в условиях неопределенности (системный анализ, управление и прогнозирование с элементами компьютерного моделирования) / А. Е. Краснов, О. Н. Красуля, О. В. Большаков, Т. В. Шленская. – Москва : Всероссийский научно-исследовательский институт мясной промышленности им. В.М. Горбатова РАСХН, 2001. – 496 с. – ISBN 5-901768-02-7. – EDN RSVKBH.

33. Краснов, А. Е. Нейросетевой метод снижения размерности спектральных данных / А. Е. Краснов, В. А. Вагин, Д. Н. Никольский // Современные технологии обработки сигналов : 2-я Всероссийская конференция: доклады конференции, Москва, 13 декабря 2019 года. – Москва: Московское НТО радиотехники, электроники и связи им. А.С. Попова, 2019. – С. 136-141. – EDN IYGDXQ. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=42308374 (дата обращения: 12.08.2025).

34. Лимановская, О. В. Основы машинного обучения : учебное пособие / О. В. Лимановская, Т. И. Алферьева ; науч. ред. И. . Обабков ; Уральский федеральный университет им. первого Президента России Б. Н. Ельцина. – Екатеринбург : Издательство Уральского университета, 2020. – 91 с. / Текст : электронный. – URL: https://biblioclub.ru/index.php?page=book&id=699059 (дата обращения: 12.08.2025).

35. Макеева, О. В. Анализ методов и средств обработки векторных массивов данных в нейроструктурах / О. В. Макеева, С. А. Красников, С. В. Николаева // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. – 2023. – № 5. – С. 84-87. – DOI 10.37882/2223-2966.2023.05.20. – EDN RRWAAI. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=54280363 (дата обращения: 24.09.2025).

36. Мелехин А. О. Новый дериватизирующий агент для определения метаболитов нитрофуранов в куриных яйцах методом высокоэффективной жидкостной хроматографии–тандемной масс-спектрометрии / А. О. Мелехин, В. В. Толмачева, Е. Г. Шубина [и др.] // Журнал аналитической химии. – 2021. – Т. 76. – № 11. – С. 1012-1021.

37. Митрофанов, А. Ю. Влияние рассеяния света в образце на эффективность фотоинициирования / А. Ю. Митрофанов, А. С. Зверев, Д. А. Мальцев // Бутлеровские сообщения. – 2012. – Т. 31, № 9. – С. 126-129. – EDN PWSSKL. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=18860292 (дата обращения: 12.08.2025).

38. Мищенкова, А. И. Модель экспертной системы для контроля качества жидких сред по их спектральным характеристикам / А. И. Мищенкова, С. А. Красников, С. В. Николаева // Системный анализ в проектировании и управлении : сборник научных трудов XXII Международной научно-практической конференции, Санкт-Петербург, 22–24 мая 2018 года. Том Часть 2. – Санкт-Петербург: ФГАОУ ВО СПбПУ, 2018. – С. 147-151. – EDN UUKPTU. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=35279065 (дата обращения: 12.08.2025).

39. Николаева, С. В. Использование мер сходства для анализа данных / С. В. Николаева, С. А. Красников, М. Л. Рысин // Естественные и технические науки. – 2022. – № 11(174). – С. 213-215. – EDN EIGENJ. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=50169993 (дата обращения: 24.09.2025).

40. Николаева, С. В. Математическое моделирование смесей жидких сред : Учебник для студентов / С. В. Николаева. – Москва : Спутник +, 2024. – 104 с. – ISBN 978-5-9973-6801-2. – EDN KLZPLJ. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=60748704 (дата обращения: 24.09.2025).

41. Носенко, Т. Н. Применение инфракрасной спектроскопии и мультивариантного анализа к исследованию сывороток крови пациентов, больных

эпилепсией / Т. Н. Носенко, В. Е. Ситникова, Р. О. Олехнович, М. В. Успенская // Научно-технический вестник информационных технологий, механики и оптики. – 2019. – Т. 19, № 3. – С. 402-409. – DOI 10.17586/2226-1494-2019-19-3-402-409. – EDN JFTIOY. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=38219845 (дата обращения: 18.08.2025).

42. Овчинникова, М. А. Применение нейронных сетей для анализа многомерных данных в виде спектра / М. А. Овчинникова // Современные информационные технологии в образовании, науке и промышленности : Сборник трудов XXVII Международная конференция; XXV Международный конкурс научных и научно-методических работ., Москва, 08–09 февраля 2024 года. – Москва: Общество с ограниченной ответственностью "Издательство «Экон-Информ», 2024. – С. 45-46. – EDN JIHLID. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=65082339 (дата обращения: 12.08.2025).

43. Палов, В. Н. Использование интеллектуальных методов обработки данных раман-спектроскопии для диагностики злокачественных опухолей / В. Н. Павлов, А. Р. Билялов, Р. Ф. Гильманова [и др.] // Медицинский вестник Башкортостана. – 2018. – Т. 13, № 3(75). – С. 43-47. – EDN SQRSQL. / Текст : электронный. – URL: https://www.elibrary.ru/item.asp?id=35554620 (дата обращения: 18.08.2025).

44. Подвальный, С. Л. Сравнение алгоритмов кластерного анализа на случайном наборе данных / С. Л. Подвальный, А. В. Плотников, А. М. Белянин // Вестник Воронежского государственного технического университета. – 2012. – Т. 8, № 5. – С. 4-6. – EDN OYHIFV. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=17743528 (дата обращения: 12.08.2025).

45. Саакян А. В. Программное обеспечение для обработки спектральных данных методами хемометрики и машинного обучения / А. В. Саакян, А. Д. Левин // ANALYTICS Russia. – 2024. – Т. 14. – № 2. – С. 154-160.

46. Старухин, А. С. Плазмонное усиление флуоресценции металлокомплексов фталоцианинов в водных растворах наночастиц серебра / А. С. Старухин, В. В. Апяри, А. В. Горский [и др.] // XIII международные чтения по квантовой оптике (IWQO - 2019) : Сборник тезисов, Владимир, 09–14 сентября 2019 года. – Владимир: Издательство «Тровант», 2019. – С. 376-379. – EDN ETSWEJ. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=41133619 (дата обращения: 12.08.2025).

47. Таныкова Н. Г. Методы ИК-Фурье-спектроскопии в комплексном анализе осадочных пород / Н. Г. Таныкова, Ю. Ю. Петрова, М. Ю. Спасенных [и др.] // Zhurnal Analiticheskoi Khimii. – 2024. – Т. 79. – № 1. – С. 12-23.

48. Терентьева, А. И. Новые варианты использования процессов формирования, разрушения и агрегации наночастиц серебра в целях химического анализа / Е. А. Терентьева, А. И. Исаченко, В. В. Апяри, С. Г. Дмитриенко // VI Всероссийская конференция по наноматериалам с элементами научной школы для молодежи : Сборник материалов, Москва, 22–25 ноября 2016 года. – Москва: ИМЕТ РАН, 2016. – С. 615-616. – EDN XHEZOL. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=27655806 (дата обращения: 12.08.2025).

49. Федоркина И. А. Многоуровневый мониторинг окружающей среды / И. А. Федоркина, В. А. Курбатов // Тенденции развития науки и образования. – 2024. – Т. 106. – № 9. – С. 114-129.

50. Федорова О. А. Методы оптической спектроскопии (методическое пособие к задачам спецпрактикума кафедры химии нефти и органического катализа) / О. А. Федорова, И. И. Кулакова, Ю. А. Сотникова, [и др.]. – 2015. – 117 с. – URL: https://istina.msu.ru/publications/book/9106754/ (дата обращения: 12.08.2025). – Текст : электронный.

51. Фешина, Е. В. Экономическая необходимость производства спектрометров как фактора развития научно-технического прогресса / Е. В. Фешина, Е. Е. Острожная,

Д. А. Омельченко, П. Е. Фиге // Естественно-гуманитарные исследования. – 2019. – № 25(3). – С. 162-166. – EDN IPAGGD. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=41162158 (дата обращения: 12.08.2025).

52. Филатов, А. С. Алгоритмический конвейер обработки спектров жидких сред для автоматизированной спектральной аналитики / А. С. Филатов, С. В. Николаева // Автоматизация в промышленности. – 2025. – № 8. – С. 36-40. – EDN VXUKSM. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=82709488 (дата обращения: 20.08.2025).

53. Филатов, А. С. Алгоритмы оптимального различения обобщённых спектральных данных / А. С. Филатов, С. В. Николаева, С. А. Красников // Информационно-аналитические и интеллектуальные системы для производства и социальной сферы : Сборник статей всероссийской межвузовской научно-практической конференции молодых учёных, Москва, 24 ноября 2022 года / Российский биотехнологический университет. – Курск: Закрытое акционерное общество «Университетская книга», 2022. – С. 5-12. – EDN YHRPNB. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=49877302 (дата обращения: 20.08.2025).

54. Филатов, А. С. Влияние выбора метрик расстояния на визуализацию данных / А. С. Филатов, С. В. Николаева, В. Н. Гельмиярова // Научно-технический вестник Поволжья. – 2024. – № 9. – С. 86-88. – EDN JXVUMQ. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=72654970 (дата обращения: 20.08.2025).

55. Филатов, А. С. Интерфейс экспертно-нейросетевой системы / А. С. Филатов // Современные информационные технологии в образовании, науке и промышленности : Сборник трудов XXVI Международной конференции, XXIV Международного конкурса научных и научно-методических работ, III Международного конкурса «Нейросетевой рисунок», Москва, 09–10 ноября 2023 года. – Москва: Общество с

ограниченной ответственностью "Издательство «Экон-Информ», 2023. – С. 112-115. – EDN VAJDYM. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=59932220 (дата обращения: 20.08.2025).

56. Филатов, А. С. Исследование влияния методов предварительной обработки информации на точность выполнения кластерного анализа / А. С. Филатов // Современные информационные технологии в образовании, науке и промышленности : XXIX Международная конференция, XXVII Международный конкурс научных и научно-методических работ, IV Международный конкурс «Нейросетевой рисунок» : сборник трудов, Москва, 07–08 ноября 2024 года. – Москва: Экон-Информ, 2024. – С. 87-92. – EDN JUNLNH. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=78772342 (дата обращения: 20.08.2025).

57. Филатов, А. С. Кластеризация многомерных спектральных данных с применением алгоритма уменьшения размерности / А.С. Филатов, С. В. Николаева, С. А. Красников, М. В. Сартаков [и др.] // Научно-технический вестник Поволжья. – 2023. – № 10. – С. 273-277. – EDN UQMBOE. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=54795427 (дата обращения: 20.08.2025).

58. Филатов, А. С. Методы предварительной обработки спектров для улучшения их последующей визуализации и кластеризации / А. С. Филатов, С. В. Николаева // Инженерные технологии. – 2024. – № 4(8). – С. 15-21. – EDN IHPFUA. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=78508676 (дата обращения: 20.08.2025).

59. Филатов, А. С. Метрики расстояний алгоритма снижения размерности UMAP / А. С. Филатов // Современные информационные технологии в образовании, науке и промышленности : Сборник трудов. XXVIII Международная конференция. XXVI Международный конкурс научных и научно-методических работ. Всероссийский конкурс проектов «Научное творческое сообщество», Москва, 25–26 апреля 2024 года.

– Москва: Общество с ограниченной ответственностью "Издательство «Экон-Информ», 2024. – С. 120-122. – EDN WRNNSA. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=67980270 (дата обращения: 20.08.2025).

60. Филатов, А. С., Николаева С.В. Оценка результатов алгоритма снижения размерности для предсказания эффективности кластеризации // А. С. Филатов, С. В. Николаева // Вестник МЭИ. – 2025. – №5. – С. 114-119.

61. Филатов, А. С. Различение спектральных данных / А. С. Филатов, А. Е. Краснов, С. В. Николаева, С. А. Красников // Современные информационные технологии в образовании, науке и промышленности : XXIII Международная конференция, XXI Международный конкурс научных и научно-методических работ, II Международный конкурс «Нейросетевой рисунок», Москва, 10–11 ноября 2022 года. – Москва: Общество с ограниченной ответственностью "Издательство «Экон-Информ», 2022. – С. 65-67. – EDN AEEMSY. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=50067599 (дата обращения: 20.08.2025).

62. Филатов, А. С. Свидетельство о государственной регистрации базы данных № 2025622177 Российская Федерация. Структурированная база данных спектроскопических сигнатур углеводородных соединений с поддержкой многомерного статистического анализа и машинного обучения, свидетельство о государственной регистрации базы данных : заявл. 25.04.2025 : опубл. 23.05.2025 / А. С. Филатов ; заявитель ФГБОУ ВО «МИРЭА – Российский технологический университет». / Текст : электронный. – URL: https://www.fips.ru/registers-doc-view/fips_servlet?DB=DB&DocNumber=2025622177&TypeFile=html (дата обращения: 20.08.2025).

63. Филатов, А. С. Свидетельство о государственной регистрации программы для ЭВМ № 2024665128 Российская Федерация. Модуль кластеризации и анализа спектральных данных : № 2024663347 : заявл. 13.06.2024 : опубл. 27.06.2024 / А. С.

Филатов ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет». – EDN KNRGLL. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=68596850 (дата обращения: 20.08.2025).

64. Филатов, А. С. Свидетельство о государственной регистрации программы для ЭВМ № 2025619420 Российская Федерация. Образовательный инструмент для анализа спектральных данных с применением методов уменьшения размерности и кластеризации : заявл. 03.04.2025 : опубл. 16.04.2025 / А. С. Филатов, А. Е. Краснов ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет». – EDN VTSHEX. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=80655775 (дата обращения: 20.08.2025).

65. Храмов, Е. С. Исследование полупроводниковых наносфер,наноэллипсоидов и наностержней в контекстеприменения в оптических плазмонных сенсорах / Е. С. Храмов, В. А. Астапенко, Ю. А. Кротов // Труды МФТИ. Труды Московского физико-технического института (национального исследовательского университета). – 2020. – Т. 12, № 1(45). – С. 67-73. – DOI 10.53815/20726759_2020_12_1_67. – EDN DECMZF. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=42402994 (дата обращения: 12.08.2025).

66. Яблонцева, А. Д. Индекс Дэвиса-Болдина для оценки кластеризации методом k-средних в Python / А. Д. Яблонцева // Modern Science. – 2021. – № 7. – С. 388-392. – EDN ZTQYTF. / Текст : электронный. – URL: https://elibrary.ru/item.asp?id=46442728 (дата обращения: 12.08.2025).

67. Ankerst M. OPTICS / M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander // ACM SIGMOD Record. – 1999. – Т. 28. – № 2. – С. 49-60.

68. Anzanello M. J. A genetic algorithm-based framework for wavelength selection on sample categorization / M. J. Anzanello, G. Yamashita, M. Marcelo [и др.] // Drug Testing and Analysis. – 2017. – Т. 9. – № 8. – С. 1172-1181.

69. Aubertin K. Mesoscopic characterization of prostate cancer using Raman spectroscopy: potential for diagnostics and therapeutics / K. Aubertin, V. Q. Trinh, M. Jermyn [и др.] // BJU International. – 2018. – Т. 122. – № 2. – С. 326-336.

70. Avohou T. H. Interpretable One-Class Classification of Raman Spectra Using Prediction Bands Estimated by Wavelet Regression / T. H. Avohou, P.-Y. Sacré, P. Hubert, E. Ziemons // Analytical Chemistry. – 2022. – Т. 94. – № 10. – С. 4183-4191.

71. Baker M. J. Developing and understanding biofluid vibrational spectroscopy: a critical review / M. J. Baker, S. R. Hussain, L. Lovergne [и др.] // Chemical Society Reviews. – 2016. – Т. 45. – № 7. – С. 1803-1818.

72. Balabin R. M. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data / R. M. Balabin, S. V. Smirnov // Analytica Chimica Acta. – 2011. – Т. 692. – № 1-2. – С. 63-72.

73. Baria E. Supervised learning methods for the recognition of melanoma cell lines through the analysis of their Raman spectra / E. Baria, R. Cicchi, F. Malentacchi [и др.] // Journal of Biophotonics. – 2021. – Т. 14. – № 3.

74. Barnes R. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra / R. J. Barnes, M. S. Dhanoa, S. J. Lister // Applied Spectroscopy. – 1989. – Т. 43. – № 5. – С. 772-777.

75. Beaver C. Model Optimization for the Prediction of Red Wine Phenolic Compounds Using Ultraviolet–Visible Spectra / C. Beaver, T. S. Collins, J. Harbertson // Molecules. – 2020. – Т. 25. – № 7. – С. 1576.

76. Beć K. B. Interpretability in near-infrared (NIR) spectroscopy: Current pathways to the long-standing challenge / K. B. Beć, J. Grabska, C. W. Huck // TrAC Trends in Analytical Chemistry. – 2025. – Т. 189.

77. Beć K. B. Near-Infrared Spectroscopy in Bio-Applications / K. B. Beć, J. Grabska, C. W. Huck // Molecules. – 2020. – Т. 25. – № 12. – С. 2948.

78. Beck A. G. Recent Developments in Machine Learning for Mass Spectrometry / A. G. Beck, M. Muhoberac, C. E. Randolph [и др.] // ACS Measurement Science Au. – 2024. – Т. 4. – № 3. – С. 233-246.

79. Belkin M. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation / M. Belkin, P. Niyogi // Neural Computation. – 2003. – Т. 15. – № 6. – С. 1373-1396.

80. Blake N. Machine Learning of Raman Spectroscopy Data for Classifying Cancers: A Review of the Recent Literature / N. Blake, R. Gaifulina, L. D. Griffin [и др.] // Diagnostics. – 2022. – Т. 12. – № 6. – С. 1491.

81. Bookstein A. Generalized Hamming Distance / A. Bookstein, V. A. Kulyukin, T. Raita // Information Retrieval. – 2002. – Т. 5. – № 4. – С. 353-375.

82. Borg I. Modern Multidimensional Scaling: Theory and Applications / I. Borg, P. Groenen // Journal of Educational Measurement. – 2003. – Т. 40. – № 3. – С. 277-280.

83. Branden K. Vanden. Robust classification in high dimensions based on the SIMCA Method / K. Vanden Branden, M. Hubert // Chemometrics and Intelligent Laboratory Systems. – 2005. – Т. 79. – № 1-2. – С. 10-21.

84. Bray J. R. An Ordination of the Upland Forest Communities of Southern Wisconsin / J. R. Bray, J. T. Curtis // Ecological Monographs. – 1957. – Т. 27. – № 4. – С. 325-349.

85. Bro R. Principal component analysis / R. Bro, A. K. Smilde // Anal. Methods. – 2014. – Т. 6. – № 9. – С. 2812-2831.

86. Bury D. Phenotyping Metastatic Brain Tumors Applying Spectrochemical Analyses: Segregation of Different Cancer Types / D. Bury, G. Faust, M. Paraskevaidi [и др.] // Analytical Letters. – 2019. – Т. 52. – № 4. – С. 575-587.

87. Campello R. J. G. B. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection / R. J. G. B. Campello, D. Moulavi, A. Zimek, J. Sander // ACM Transactions on Knowledge Discovery from Data. – 2015. – Т. 10. – № 1. – С. 1-51.

88. Capecchi A. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome / A. Capecchi, D. Probst, J.-L. Reymond // Journal of Cheminformatics. – 2020. – Т. 12. – № 1. – С. 43.

89. Chacón J. E. Minimum adjusted Rand index for two clusterings of a given size / J. E. Chacón, A. I. Rastrojo // Advances in Data Analysis and Classification. – 2023. – Т. 17. – № 1. – С. 125-133.

90. Chen C. Rapid diagnosis of lung cancer and glioma based on serum Raman spectroscopy combined with deep learning / C. Chen, W. Wu, C. Chen [и др.] // Journal of Raman Spectroscopy. – 2021. – Т. 52. – № 11. – С. 1798-1809.

91. Chen F. Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing / F. Chen, C. Sun, Z. Yue [и др.] // Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. – 2022. – Т. 265.

92. Chen H. A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources / H. Chen, A. Chen, L. Xu [и др.] // Agricultural Water Management. – 2020. – Т. 240.

93. Comaniciu D. Mean shift: a robust approach toward feature space analysis / D. Comaniciu, P. Meer // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Т. 24. – № 5. – С. 603-619.

94. Damle A. Simple, direct and efficient multi-way spectral clustering / A. Damle, V. Minden, L. Ying // Information and Inference: A Journal of the IMA. – 2019. – Т. 8. – № 1. – С. 181-203.

95. Davies D. L. A Cluster Separation Measure / D. L. Davies, D. W. Bouldin // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1979. – Т. PAMI-1. – № 2. – С. 224-227.

96. Debus B. Deep learning in analytical chemistry / B. Debus, H. Parastar, P. Harrington, D. Kirsanov // TrAC Trends in Analytical Chemistry. – 2021. – Т. 145.

97. Deconinck E. Chemometrics and infrared spectroscopy – A winning team for the analysis of illicit drug products / E. Deconinck, C. Duchateau, M. Balcaen [и др.] // Reviews in Analytical Chemistry. – 2022. – Т. 41. – № 1. – С. 228-255.

98. Deconinck E. Chemometrics and chromatographic fingerprints to discriminate and classify counterfeit medicines containing PDE-5 inhibitors / E. Deconinck, P. Y. Sacré, P. Courselle, J. O. De Beer // Talanta. – 2012. – Т. 100. – С. 123-133.

99. Donoho D. L. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data / D. L. Donoho, C. Grimes // Proceedings of the National Academy of Sciences. – 2003. – Т. 100. – № 10. – С. 5591-5596.

100. Ester M. A density-based algorithm for discovering clusters in large spatial databases with noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining : KDD’96. – AAAI Press, 1996. – С. 226-231.

101. Ezenarro J. How Are Chemometric Models Validated? A Systematic Review of Linear Regression Models for NIRS Data in Food Analysis / J. Ezenarro, D. Schorn-García // Journal of Chemometrics. – 2025. – Т. 39. – № 6.

102. Filatov A. S. Interactive system for spectral data analysis using dimensionality reduction and clustering methods / A. S. Filatov. – Текст : электронный // Frontiers in

Cybersecurity, Artificial Intelligence, and Data-Driven Technologies: Research Anthology and Selected Contributions. – 2025. – С. 134-137. – URL: https://drive.google.com/file/d/1FYvYL5lorVx2-L2e7THR_yBX29lHmZV_/view (дата обращения: 20.08.2025).

103. Flanagan A. R. Open-source Raman spectra of chemical compounds for active pharmaceutical ingredient development / A. R. Flanagan, F. G. Glavin. – Текст : электронный // Scientific Data. – 2025. – Т. 12. – № 1. – С. 498. – URL: https://www.nature.com/articles/s41597-025-04848-6 (дата обращения: 12.08.2025).

104. Flanagan A. R. Open-source Raman spectra of chemical compounds for active pharmaceutical ingredient development. figshare. Dataset / A. R. Flanagan, F. G. Glavin. – 2025.

105. Friedman S. L. Chebyshev constant and Chebyshev points / S. L. Friedman // Transactions of the American Mathematical Society. – 1973. – Т. 186. – С. 129-129.

106. Gautam R. Review of multidimensional data processing approaches for Raman and infrared spectroscopy / R. Gautam, S. Vanga, F. Ariese, S. Umapathy // EPJ Techniques and Instrumentation. – 2015. – Т. 2. – № 1. – С. 8.

107. Geladi P. Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat / P. Geladi, D. MacDougall, H. Martens // Applied Spectroscopy. – 1985. – Т. 39. – № 3. – С. 491-500.

108. Georgiev D. RamanSPy: An Open-Source Python Package for Integrative Raman Spectroscopy Data Analysis / D. Georgiev, S. V. Pedersen, R. Xie [и др.] // Analytical Chemistry. – 2024. – Т. 96. – № 21. – С. 8492-8500.

109. Guo S. Optimization of Raman-spectrum baseline correction in biological application / S. Guo, T. Bocklitz, J. Popp // The Analyst. – 2016. – Т. 141. – № 8. – С. 2396-2404.

110. Guo S. Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications / S. Guo, P. Rösch, J. Popp, T. Bocklitz // Journal of Chemometrics. – 2020. – Т. 34. – № 4.

111. Han J. Data, measurements, and data preprocessing / J. Han, J. Pei, H. Tong. – Текст : электронный // Data Mining. – 2023. – С. 23-84. – URL: (дата обращения: 29.09.2025).

112. Herman Edwin. Calculus / Edwin. Herman, Gilbert. Strang. – OpenStax, Rice University, 2018.

113. Houhou R. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data / R. Houhou, T. Bocklitz. – Текст : электронный // Analytical Science Advances. – 2021. – Т. 2. – № 3-4. – С. 128-141. – URL: /doi/pdf/10.1002/ansa.202000162 (дата обращения: 16.08.2025).

114. Hubert L. Comparing partitions / L. Hubert, P. Arabie // Journal of Classification. – 1985. – Т. 2. – № 1. – С. 193-218.

115. Jolliffe I. T. . Principal Component Analysis / I. T. . Jolliffe. – Springer New York, NY, 2006.

116. Juan A. de. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem / A. de Juan, J. Jaumot, R. Tauler // Anal. Methods. – 2014. – Т. 6. – № 14. – С. 4964-4976.

117. Jumper J. Highly accurate protein structure prediction with AlphaFold / J. Jumper, R. Evans, A. Pritzel [и др.] // Nature. – 2021. – Т. 596. – № 7873. – С. 583-589.

118. Kennard R. W. Computer Aided Design of Experiments / R. W. Kennard, L. A. Stone // Technometrics. – 1969. – Т. 11. – № 1. – С. 137-148.

119. Klein N. Hyperspectral Target Identification Using Physics-Guided Neural Networks with Explainability and Feature Attribution / N. Klein, A. Carr, Z. Hampel-Arias

[и др.] // IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium. – IEEE, 2023. – С. 946-949.

120. Krause E. F. . Taxicab Geometry : an adventure in non-Euclidean geometry / E. F. . Krause. – Текст : электронный. – 1987. – С. 88. – URL: https://books.google.com/books/about/Taxicab_Geometry.html?id=IW7ICV0QXWwC (дата обращения: 29.09.2025).

121. Lance G. N. Computer Programs for Hierarchical Polythetic Classification («Similarity Analyses») / G. N. Lance, W. T. Williams // The Computer Journal. – 1966. – Т. 9. – № 1. – С. 60-64.

122. Leger M. N. Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures / M. N. Leger, A. G. Ryder // Applied Spectroscopy. – 2006. – Т. 60. – № 2. – С. 182-193.

123. Liberti L. Euclidean distance geometry and applications / L. Liberti, C. Lavor, N. Maculan, A. Mucherino. – 2012..

124. Lieber C. A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra / C. A. Lieber, A. Mahadevan-Jansen // Applied Spectroscopy. – 2003. – Т. 57. – № 11. – С. 1363-1367.

125. Lilek D. Machine Learning of Raman Spectroscopic Data: Comparison of Different Validation Strategies / D. Lilek, D. Zimmermann, L. Steininger [и др.] // Journal of Raman Spectroscopy. – 2025.

126. Luo R. Deep Learning for Raman Spectroscopy: A Review / R. Luo, J. Popp, T. Bocklitz // Analytica. – 2022. – Т. 3. – № 3. – С. 287-301.

127. Ma L. Systematic discovery about NIR spectral assignment from chemical structural property to natural chemical compounds / L. Ma, Y. Peng, Y. Pei [и др.] // Scientific Reports. – 2019. – Т. 9. – № 1. – С. 9503.

128. Maaten L. van der. Visualizing Data using t-SNE / L. van der Maaten, G. Hinton // Journal of Machine Learning Research. – 2008. – Т. 9. – № 86. – С. 2579-2605.

129. Maesschalck R. De. Decision criteria for soft independent modelling of class analogy applied to near infrared data / R. De Maesschalck, A. Candolfi, D. L. Massart, S. Heuerding // Chemometrics and Intelligent Laboratory Systems. – 1999. – Т. 47. – № 1. – С. 65-77.

130. Mainali D. Automated Fast Screening Method for Cocaine Identification in Seized Drug Samples Using a Portable Fourier Transform Infrared (FT-IR) Instrument / D. Mainali, J. Seelenbinder // Applied Spectroscopy. – 2016. – Т. 70. – № 5. – С. 916-922.

131. Martyna A. Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components / A. Martyna, A. Menżyk, A. Damin [и др.] // Chemometrics and Intelligent Laboratory Systems. – 2020. – Т. 202.

132. McInnes L. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction / L. McInnes, J. Healy, J. Melville. – 2020.

133. Mevik B.-H. The pls Package: Principal Component and Partial Least Squares Regression in R / B.-H. Mevik, R. Wehrens // Journal of Statistical Software. – 2007. – Т. 18. – № 2.

134. Milman B. L. General principles of identification by mass spectrometry / B. L. Milman // TrAC Trends in Analytical Chemistry. – 2015. – Т. 69. – С. 24-33.

135. Naes T. Incorporating interactions in multi-block sequential and orthogonalised partial least squares regression / T. Naes, I. Måge, V. H. Segtnan // Journal of Chemometrics. – 2011. – Т. 25. – № 11. – С. 601-609.

136. Nørgaard L. Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy / L. Nørgaard, A. Saudland, J. Wagner [и др.] // Applied Spectroscopy. – 2000. – Т. 54. – № 3. – С. 413-419.

137. Pearson K. Notes on regression and inheritance in the case of two parents / K. Pearson. – Текст : электронный // Proceedings of the Royal Society of London. – 1895. – T. 58. – C. 240-242. – URL: https://books.google.com/books?id=60aL0zlT-90C&pg=PA240 (дата обращения: 29.09.2025).

138. Penido C. A. F. O. Identification of Different Forms of Cocaine and Substances Used in Adulteration Using Near-infrared Raman Spectroscopy and Infrared Absorption Spectroscopy / C. A. F. O. Penido, M. T. T. Pacheco, R. A. Zângaro, L. Silveira // Journal of Forensic Sciences. – 2015. – T. 60. – № 1. – C. 171-178.

139. Ralbovsky N. M. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning / N. M. Ralbovsky, I. K. Lednev // Chemical Society Reviews. – 2020. – T. 49. – № 20. – C. 7428-7453.

140. Rebiere H. Raman chemical imaging for spectroscopic screening and direct quantification of falsified drugs / H. Rebiere, M. Martin, C. Ghyselinck [и др.] // Journal of Pharmaceutical and Biomedical Analysis. – 2018. – T. 148. – C. 316-323.

141. Ríos-Reina R. How Chemometrics Revives the UV-Vis Spectroscopy Applications as an Analytical Sensor for Spectralprint (Nontargeted) Analysis / R. Ríos-Reina, S. M. Azcarate // Chemosensors. – 2022. – T. 11. – № 1. – C. 8.

142. Ríos-Reina R. Spectralprint techniques for wine and vinegar characterization, authentication and quality control: Advances and projections / R. Ríos-Reina, J. M. Camiña, R. M. Callejón, S. M. Azcarate // TrAC Trends in Analytical Chemistry. – 2021. – T. 134.

143. Roger J.-M. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy / J.-M. Roger, A. Biancolillo, F. Marini // Chemometrics and Intelligent Laboratory Systems. – 2020. – T. 199.

144. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis / P. J. Rousseeuw // Journal of Computational and Applied Mathematics. – 1987. – T. 20. – C. 53-65.

145. Roweis S. T. Nonlinear Dimensionality Reduction by Locally Linear Embedding / S. T. Roweis, L. K. Saul // Science. – 2000. – Т. 290. – № 5500. – С. 2323-2326.

146. Ruff L. Deep One-Class Classification / L. Ruff, R. Vandermeulen, N. Goernitz [и др.] // Proceedings of the 35th International Conference on Machine Learning : Proceedings of Machine Learning Research / ред. J. Dy, A. Krause. – PMLR, 2018. – Т. 80. – С. 4393-4402.

147. Savitzky Abraham. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. / Abraham. Savitzky, M. J. E. Golay // Analytical Chemistry. – 1964. – Т. 36. – № 8. – С. 1627-1639.

148. Schölkopf B. Estimating the Support of a High-Dimensional Distribution / B. Schölkopf, J. C. Platt, J. Shawe-Taylor [и др.] // Neural Computation. – 2001. – Т. 13. – № 7. – С. 1443-1471.

149. Sculley D. Web-scale k-means clustering / D. Sculley // Proceedings of the 19th international conference on World wide web. – New York, NY, USA : ACM, 2010. – С. 1177-1178.

150. Sheehy G. Open-sourced Raman spectroscopy data processing package implementing a baseline removal algorithm validated from multiple datasets acquired in human tissue and biofluids / G. Sheehy, F. Picot, F. Dallaire [и др.] // Journal of Biomedical Optics. – 2023. – Т. 28. – № 02.

151. Singh K. P. Support vector machines in water quality management / K. P. Singh, N. Basant, S. Gupta // Analytica Chimica Acta. – 2011. – Т. 703. – № 2. – С. 152-162.

152. Strani L. One class classification (class modelling): State of the art and perspectives / L. Strani, M. Cocchi, D. Tanzilli [и др.] // TrAC Trends in Analytical Chemistry. – 2025. – Т. 183.

153. Tax D. M. J. Support Vector Data Description / D. M. J. Tax, R. P. W. Duin // Machine Learning. – 2004. – Т. 54. – № 1. – С. 45-66.

154. Tenenbaum J. B. A Global Geometric Framework for Nonlinear Dimensionality Reduction / J. B. Tenenbaum, V. de Silva, J. C. Langford // Science. – 2000. – Т. 290. – № 5500. – С. 2319-2323.

155. Tiwary S. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis / S. Tiwary, R. Levy, P. Gutenbrunner [и др.] // Nature Methods. – 2019. – Т. 16. – № 6. – С. 519-525.

156. Torniainen J. Open-source python module for automated preprocessing of near infrared spectroscopic data / J. Torniainen, I. O. Afara, M. Prakash [и др.] // Analytica Chimica Acta. – 2020. – Т. 1108. – С. 1-9.

157. Tran N. H. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry / N. H. Tran, R. Qiao, L. Xin [и др.] // Nature Methods. – 2019. – Т. 16. – № 1. – С. 63-66.

158. Ukil A. Improved Calibration of Near-Infrared Spectra by Using Ensembles of Neural Network Models / A. Ukil, J. Bernasconi, H. Braendle [и др.]. – 2015.

159. Verbeeck N. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry / N. Verbeeck, R. M. Caprioli, R. Van de Plas // Mass Spectrometry Reviews. – 2020. – Т. 39. – № 3. – С. 245-291.

160. Wahl J. Single-Step Preprocessing of Raman Spectra Using Convolutional Neural Networks / J. Wahl, M. Sjödahl, K. Ramser // Applied Spectroscopy. – 2020. – Т. 74. – № 4. – С. 427-438.

161. Wold S. Pattern recognition by means of disjoint principal components models / S. Wold // Pattern Recognition. – 1976. – Т. 8. – № 3. – С. 127-139.

162. Yan C. A review on spectral data preprocessing techniques for machine learning and quantitative analysis / C. Yan // iScience. – 2025. – Т. 28. – № 7.

163. Zhang R. Transfer-learning-based Raman spectra identification / R. Zhang, H. Xie, S. Cai [и др.] // Journal of Raman Spectroscopy. – 2020. – Т. 51. – № 1. – С. 176-186.

164. Zhang W. A Review of Machine Learning for Near-Infrared Spectroscopy / W. Zhang, L. C. Kasun, Q. J. Wang [и др.] // Sensors. – 2022. – Т. 22. – № 24. – С. 9764.

165. Zhang Z. MLLE: Modified Locally Linear Embedding Using Multiple Weights / Z. Zhang, J. Wang // Advances in Neural Information Processing Systems / ред. B. Schölkopf [и др.]. – MIT Press, 2006. – Т. 19.

166. Zhang Z. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment / Z. Zhang, H. Zha // Journal of Shanghai University (English Edition). – 2004. – Т. 8. – № 4. – С. 406-424.

167. Zhao J. Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy / J. Zhao, H. Lui, D. I. McLean, H. Zeng // Applied Spectroscopy. – 2007. – Т. 61. – № 11. – С. 1225-1232.

Приложение А
(справочное)
Экономическое обоснование разработанной системы

Ниже приведено подробное экономическое обоснование внедрения в производство, разработанного метода автоматической обработки и классификации спектральных данных. Расчёты сделаны на основании эмпирических характеристик метода, приведённых в третьей главе.

Для модели экономического обоснования принят сценарий лаборатории контроля качества нефтепродуктов, обрабатывающей 10'000 спектральных измерений в год. За «базовую» технологию принят текущий рабочий процесс с использованием метода LDA, показавшего наивысшую точность классификации в 94.6%, в то время как предложенная методика на больших наборах демонстрирует точность 99.7%. Стоимость одного события ошибочной классификации, включающая повторный анализ, утилизацию партии, логистику и косвенные потери, принята равной 50'000 рублей. Стоимость оператора – 400 рублей в час; среднее время на обработку и выпуск одного результата в базовом процессе 0.15 часа (9 минут), в процессе с предложенной методикой – 0.20 часа (12 минут). Первоначальные затраты на внедрение метода (разработка ПО, интеграция и дополнительное программное и аппаратное обеспечение) приняты равными 6'000'000 рублей (4'000'000 рублей – разработка, 2'000'000 рублей – модернизация оборудования). Ежегодное обслуживание оборудования оценивается в 10% от стоимости дополнительного оборудования – 200'000 рублей в год.

Количество ошибок в год при базовой технологии равно 10'000·(1-0.946)=540 ошибок в год. При предложенной методике количество ошибок равно 10'000·(1-

0.997)=30 ошибок в год. Стоимость ошибок: базовая – 540·50'000=27'000'000 рублей в год, при предложенной методике – 30·50'000=1'500'000 рублей в год. Годовая экономия на ошибках оценивается в 25'500'000 рублей.

Годовые затраты на оплату оператора при базовом процессе равны 10'000·0.15·400=600'000; при предложенной методике – 10'000·0.20·400=800'000 рублей; годовые потери по зарплате сотрудника лаборатории составляет 200'000 рублей.

Чистая годовая экономия вычисляется как суммарная годовая экономия минус ежегодные дополнительные расходы: 25'500'000-200'000-200'000=25'100'000 рублей в год. Для одноразовой первичной инвестиции 6'000'000 рублей это даёт срок окупаемости 6'000'000/25'100'000≈0.239 года, то есть приблизительно 2.9 месяца.

Внутренняя норма доходности для денежного потока при первоначальной инвестиции в 6'000'000 рублей и ежегодном притоке 25'100'000 рублей равна 418%, что указывает на чрезвычайно высокую доходность проекта в приведённом сценарии.

При меньших объёмах, например, 5'000 образцов в год, и допущении о падении стоимости ошибки до 20'000 рублей проект всё ещё остаётся экономически целесообразным: годовая чистая выгода порядка 4'700'000, срок окупаемости 1.27 года, а внутренняя норма доходности равна 78%.

Помимо прямой экономии на ошибках, внедрение метода даёт значимые косвенные преимущества: повышение качества продукции и репутации предприятия, уменьшение риска регуляторных штрафов и рекламаций, ускорение вывода продукции на рынок, улучшение прогнозируемости процессов и снижение операционного риска. Эти эффекты трудно оценить количественно, но в практических проектах их вклад часто сопоставим с экономией на исправлениях и потере сырья. Кроме того, улучшение точности идентификации с 94.6% до 99.7% снижает

волатильность операционных затрат и повышает надёжность автоматизации, что упрощает масштабирование и переход к онлайн-мониторингу.

При внедрении предложенной методики в поточных анализаторах в системах управления в реальном времени, при увеличении точности анализатора с 94.6% до 99.7%, объемы некачественной продукции уменьшатся в (1-0.946)/(1-0.997)=18 раз. Недавнее исследование проводит подробную оценку влияния увеличения выхода качественной продукции в нефтяной отрасли с большими объемами производства [11].

Для экономической оценки использования предложенной методики в качестве программного обеспечения для портативного анализатора, следует считать, что стоимость единовременной лицензии на пакет коммерческого программного обеспечения, как правило, составляет треть от цены комплекта устройства. Однако важная часть совокупной стоимости владения – это сервисные контракты, калибровки, расходники, удалённая поддержка и обновления ПО. Для аналитического оборудования годовые сервисные планы обычно оценивают приблизительно в 8-15 % от цены прибора в год.

Приложение Б

(справочное)

Свидетельство о регистрации базы данных

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации базы данных

№ 2025622177

Структурированная база данных спектроскопических сигнатур углеводородных соединений с поддержкой многомерного статистического анализа и машинного обучения

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет» (RU)***

Автор(ы): ***Филатов Александр Сергеевич (RU)***

Заявка № **2025621597**
Дата поступления **25 апреля 2025 г.**
Дата государственной регистрации
в Реестре баз данных ***23 мая 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1ca6300bf54f240f670bcca2026
Владелец Зубов Юрий Сергеевич
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов

Приложение В

(справочное)

Свидетельство о регистрации модуля ЭНС

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024665128

Модуль кластеризации и анализа спектральных данных

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет» (RU)***

Автор(ы): ***Филатов Александр Сергеевич (RU)***

Заявка № **2024663347**
Дата поступления **13 июня 2024 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***27 июня 2024 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Ю.С. Зубов

Приложение Г

(справочное)

Свидетельство о регистрации обучающей программы

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025619420

Образовательный инструмент для анализа спектральных данных с применением методов уменьшения размерности и кластеризации

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет» (RU)***

Авторы: ***Филатов Александр Сергеевич (RU), Краснов Андрей Евгеньевич (RU)***

Заявка № **2025617326**
Дата поступления **03 апреля 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***16 апреля 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

Приложение Д

(справочное)

Акт о внедрении в НТЦ УП РАН

Утверждаю
И.о. директора НТЦ УП РАН

М.С. Афанасьев

Акт об использовании
результатов диссертационной работы
«Обработка и кластеризация спектральных данных жидких сред»
Филатова Александра Сергеевича

Настоящим актом подтверждается использование результатов диссертационной работы Филатова А.С. «Обработка и кластеризация спектральных данных жидких сред», связанных с разработкой методики формирования цифровых образов жидких сред по их спектральным данным и созданием программно-алгоритмического комплекса для автоматизированного анализа и классификации многокомпонентных смесей в научно-исследовательской деятельности Научно-технологического центра уникального приборостроения Российской академии наук.

Разработанное программное обеспечение, интегрированное с фурье-спектрометром АФ-1 и многоканальным (много-зондовым) фурье-спектрометром (построенным на основе лабораторного ИК фурье-спектрометра ФСМ-2211 и работающим в непрерывном потоке), применяется при создании и модернизации экспериментальных методик, а также в научных исследованиях многокомпонентных смесей.

Внедренные решения расширили функциональные возможности оборудования, разрабатываемого в Научно-технологическом центре уникального приборостроения Российской академии наук, позволив автоматизировать обработку больших массивов экспериментальных спектральных данных. Новая методика обеспечивает высокую точность выделения ключевых признаков и формирование компактных образов для последующего машинного обучения. Это открывает путь к разработке современных средств непрерывного контроля качества на этапах производства и повышению эффективности управления технологическим процессом.

Главный научный сотрудник,
д.т.н. В.А. Вагин

Старший научный сотрудник, к.ф.-м.н. Д.В. Чуриков

Приложение Е

(справочное)

Акт о внедрении в АО МЭМП

УТВЕРЖДАЮ
Генеральный директор АО
Можайского экспериментально-
механического предприятия
В.М. Бергер
_____
«____» _____ 2025 г.

АКТ
о проверки результатов диссертационной работы в производственных условиях

Разработанная Филатовым А.С. система анализа спектральных данных, включающая методику формирования образов жидких сред (в частности, нефтепродуктов), специализированную базу данных и программное обеспечение для обработки, снижения размерности и кластеризации, обладает высокой практической значимостью и представляет интерес для Можайского экспериментально-механического предприятия. Существенным преимуществом разработки является способность обеспечивать эффективное и оперативное принятие решений в задачах контроля качества, что позволяет существенно ускорить и автоматизировать процесс оценки физико-химических свойств жидких сред (в частности, нефтепродуктов), а также повысить достоверность и устойчивость результатов анализа в условиях реального производства.

АО МЭМП подтверждает факт и целесообразность использования результатов диссертационной работы Филатова А.С. «Обработка и кластеризация спектральных данных жидких сред» в решаемых АО Можайским экспериментально-механическим предприятием задачах.

Председатель комиссии:
Начальник экономического отдела Царенко О.И.
Члены комиссии:
Зам.главного инженера Царенко И.А.
Заместитель директора Бергер А.В.

Приложение Ж

(справочное)

Акт о внедрении в АО «ВНИИ НП»

АКТ

о внедрении (использовании) результатов диссертационной работы
«Обработка и кластеризация спектральных данных жидких сред»
Филатова Александра Сергеевича

«3» октября 2025 г. г. Москва

АО «Всероссийский научно-исследовательский институт по переработке нефти (АО «ВНИИ НП») – научно-исследовательская организация, выполняющая научные исследования в области разработки технологических процессов и катализаторов нефте- и газохимии, производства масел и смазок, инженерно-технологического сопровождения процессов на нефте- газоперерабатывающих заводах, разработки и стандартизации методов испытаний нефти и нефтепродуктов.

Разработанная Филатовым А.С. программно-алгоритмическая система, включающая методы формирования образов жидких сред по их спектральным характеристикам, оптимизированные алгоритмы снижения размерности, а также модули кластеризации данных, обладает высокой практической значимостью и представляет интерес для применения в АО «ВНИИ НП» для разработки методов испытаний. Существенным преимуществом данной системы является возможность интеграции с поточными спектрометрами, что позволяет в автоматическом режиме осуществлять обработку и интерпретацию спектральной информации, обеспечивая тем самым надёжный и оперативный контроль качества нефтепродуктов в технологических процессах на нефтеперерабатывающих предприятиях.

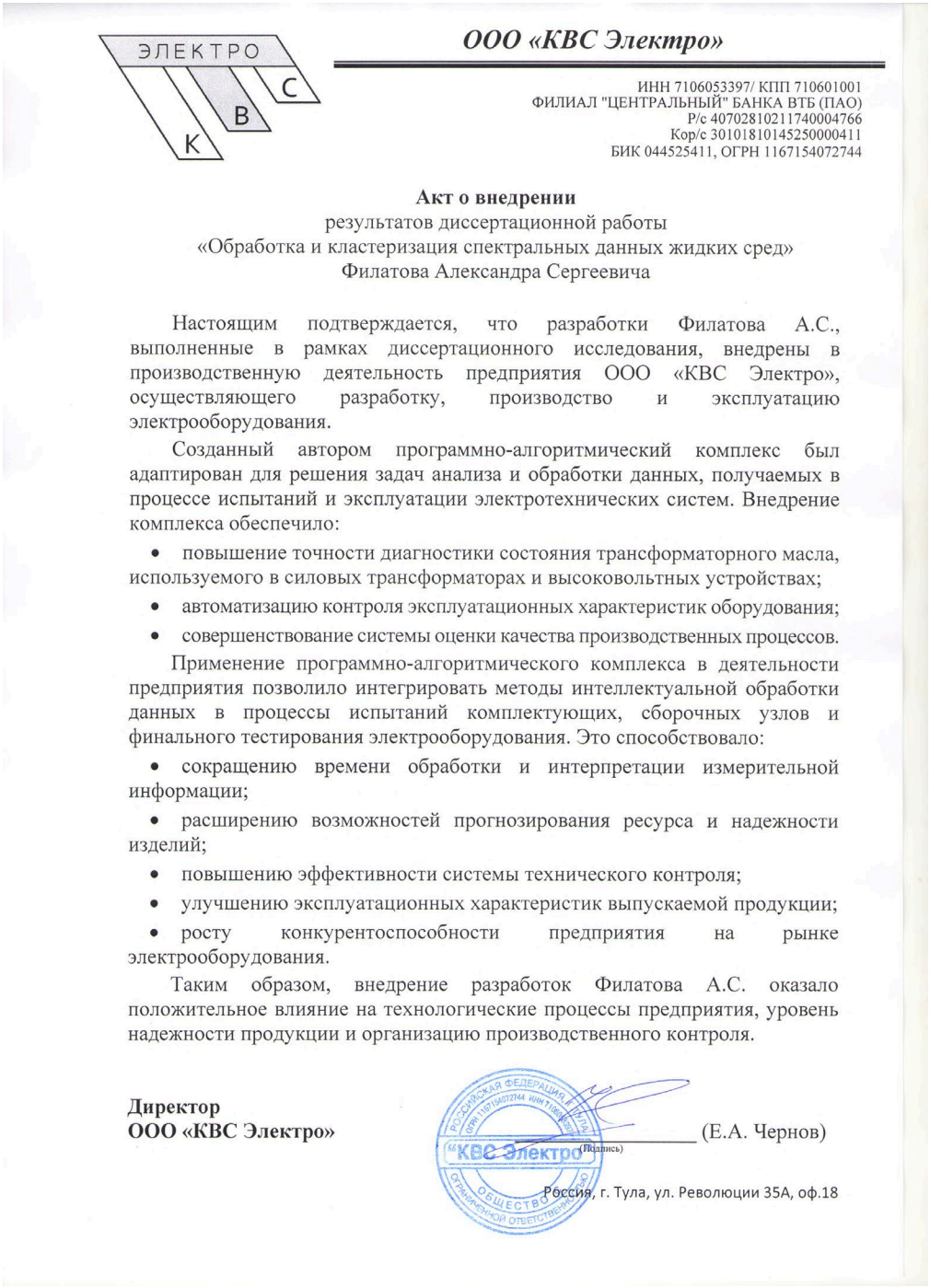
АО «ВНИИ НП» подтверждает факт и целесообразность внедрения результатов диссертационной работы Филатова А.С. по теме «Обработка и кластеризация спектральных данных жидких сред» в решаемых АО «ВНИИ НП» задачах.

Заместитель генерального директора АО «ВНИИ НП» по инженерно-технологическому сопровождению и внедрению М.Ю. Дубинский

Приложение И

(справочное)

Акт о внедрении в ООО «КВС Электро»



ЭЛЕКТРО К В С

ООО «КВС Электро»

ИНН 7106053397/ КПП 710601001
ФИЛИАЛ "ЦЕНТРАЛЬНЫЙ" БАНКА ВТБ (ПАО)
Р/с 40702810211740004766
Кор/с 30101810145250000411
БИК 044525411, ОГРН 1167154072744

Акт о внедрении
результатов диссертационной работы
«Обработка и кластеризация спектральных данных жидких сред»
Филатова Александра Сергеевича

Настоящим подтверждается, что разработки Филатова А.С., выполненные в рамках диссертационного исследования, внедрены в производственную деятельность предприятия ООО «КВС Электро», осуществляющего разработку, производство и эксплуатацию электрооборудования.

Созданный автором программно-алгоритмический комплекс был адаптирован для решения задач анализа и обработки данных, получаемых в процессе испытаний и эксплуатации электротехнических систем. Внедрение комплекса обеспечило:

- повышение точности диагностики состояния трансформаторного масла, используемого в силовых трансформаторах и высоковольтных устройствах;
- автоматизацию контроля эксплуатационных характеристик оборудования;
- совершенствование системы оценки качества производственных процессов.

Применение программно-алгоритмического комплекса в деятельности предприятия позволило интегрировать методы интеллектуальной обработки данных в процессы испытаний комплектующих, сборочных узлов и финального тестирования электрооборудования. Это способствовало:

- сокращению времени обработки и интерпретации измерительной информации;
- расширению возможностей прогнозирования ресурса и надежности изделий;
- повышению эффективности системы технического контроля;
- улучшению эксплуатационных характеристик выпускаемой продукции;
- росту конкурентоспособности предприятия на рынке электрооборудования.

Таким образом, внедрение разработок Филатова А.С. оказало положительное влияние на технологические процессы предприятия, уровень надежности продукции и организацию производственного контроля.

Директор
ООО «КВС Электро» _____ (Е.А. Чернов)
(Подпись)

КВС Электро

Россия, г. Тула, ул. Революции 35А, оф.18

Приложение К

(справочное)

Акт о внедрении в ФГБОУ ВО «МГУТУ им. К.Г. Разумовского (ПКУ)»

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО
ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное
образовательное учреждение высшего образования
**«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ ТЕХНОЛОГИЙ И УПРАВЛЕНИЯ
ИМЕНИ К.Г. РАЗУМОВСКОГО
(Первый казачий университет)»
(ФГБОУ ВО «МГУТУ им. К.Г. Разумовского (ПКУ)»)**
ОКПО 02068812 ОГРН 1027700200494
ИНН 7709125605 КПП 770901001
109004, г. Москва, ул. Земляной вал, д. 73
Телефон: (495) 640-54-36
E-mail: rektorat@mgutm.ru

_____ № _____
на №_____ от _____

АКТ

использования в учебном процессе результатов кандидатской диссертации Филатова А.С. на тему «Обработка и кластеризация спектральных данных жидких сред»

Научные результаты диссертационного исследования Филатова А.С., представленного на соискание ученой степени кандидата технических наук, применяются в образовательной деятельности со студентами программ высшего образования уровня бакалавриата по направлениям 09.03.01 Информатика и вычислительная техника, 09.03.03 Прикладная информатика. Использование материалов осуществляется на кафедре «Информационных систем и цифровых технологий» факультета цифровых технологий ФГБОУ ВО МГУТУ им. К.Г. Разумовского (ПКУ). Разработки внедрены в учебный процесс при проведении лекций и практических занятий по дисциплинам «Структуры и алгоритмы обработки данных», «Разработка программных приложений» и «Системное программное обеспечение».

Заведующая кафедрой
информационных систем и цифровых технологий **Н.А. Теплая**

Декан факультета цифровых технологий **В.В. Никитин**

Приложение Л

(справочное)

Акт о внедрении в ФГБОУ ВО «НИУ «МЭИ»

УТВЕРЖДАЮ
Проректор по науке и инновациям
ФГБОУ ВО «НИУ «МЭИ»
_____ И.И. Комаров
01.10. 2025 г.

АКТ

использования в учебном процессе результатов кандидатской диссертации Филатова Александра Сергеевича на тему «Обработка и кластеризация спектральных данных жидких сред»

Результаты диссертационной работы Филатова А.С. на соискание учёной степени кандидата технических наук используются в учебном процессе для студентов по программам высшего образования бакалавриата по направлению подготовки 11.03.04 «Электроника и наноэлектроника» при чтении дисциплин «Цифровая фильтрация изображений» и «Компьютерная обработка изображений» на кафедре физики им. В.А. Фабриканта ФГБОУ ВО «Национальный исследовательский университет «МЭИ».

Зам. зав. кафедрой физики
им. В.А. Фабриканта,
к.т.н., доцент К.М. Лапицкий