TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРАНТА

Тема работы
Кластеризация отклика калориметров эксперимента NA64 (CERN, SPS)

УДК 621.384.6.082.64

Обучающийся

Группа	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		10.06.2023 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к. ф.-м. н.		10.06.2023 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Спицына Л. Ю.	к.экон.н		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП, должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Томск – 2023 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований

Код компетенции	Наименование компетенции
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Губин Е. И.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ1И	Крамойкин Иван Алексеевич

Тема работы:

Кластеризация отклика калориметров эксперимента NA64 (CERN, SPS)	
Утверждена приказом директора (дата, номер)	№ 146-39/с от 26.05.2023 г.

Срок сдачи обучающимся выполненной работы:	10.06.2023 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.)*	Объектом исследования является алгоритм кластеризации откликов калориметров эксперимента *NA64* (*CERN*, *SPS*)
Перечень разделов пояснительной записки, подлежащих исследованию, проектированию и разработке *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования,*	1.Обзор алгоритмов кластеризации 2. Описание исходных данных 3. Разработка конвейера обработки данных 4. Описание выбранной модели калориметрического отклика 5. Описание алгоритма кластеризации

конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе)	6. Анализ качества кластеризации 7. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения. 8. Работа над разделом по социальной ответственности.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Архитектура конвейера данных 2. Результат аппроксимации эталонного сигнала модельной функцией 3. Распределения параметров модельной функции 4. Конфузионная матрица оценки результатов кластеризации 5. Диаграмма Ганта 6. Матрица SWOT

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Основная часть	ассистент ИШФВП, Дусаев Р. Р.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	доцент ОСГН ШБИП, к.экон.н, Спицына Л. Ю.
Социальная ответственность	доцент ООД ШБИП, к.б.н., Антоневич О. А.
Раздел на английском языке	доцент ОИЯ ШБИП, к.филос.н., Уткина А. Н.
Названия разделов, которые должны быть написаны на иностранном языке:	
Раздел 4 — NA64 (CERN, SPS) experiment calorimeter response clusterization.	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2023 г.

Задание выдал руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к. ф.-м. н., доцент		1.03.2023 г.

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		1.03.2023 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа Информационных технологий и робототехники
Направление подготовки (ООП / ОПОП) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2022 /2023 учебного года

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ1И	Крамойкин Иван Алексеевич

Тема работы:

Кластеризация отклика калориметров эксперимента NA64 (CERN, SPS)

Срок сдачи обучающимся выполненной работы:	10.06.2023 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2023	Основная часть	70
10.06.2023	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2023	Социальная ответственность	10
10.06.2023	Раздел на английском языке	10

СОСТАВИЛ:
руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:
руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		1.03.2023 г.

Реферат

Работа содержит пояснительную записку на 95 страницах, 8 рисунков, 23 таблицы и 1 приложение.

Ключевые слова: кластеризация, Машинное Обучение, конвейер обработки данных, ионизирующее излучение, тёмная материя.

Объектом исследования являются отклики калориметров эксперимента *NA64* (*CERN*, *SPS*).

Цель работы — реализация алгоритма кластеризации для определения групп калориметрических откликов, соответствующих адронной и электронной природе инициирующего пучка.

Разделение откликов калориметров от электронных и адронных пучков является важным этапом реконструкции физических событий эксперимента *NA64* (*CERN*, *SPS*). В настоящий момент оно осуществляется путём простого порогового отбора сигналов. В предложенной методике разделение достигается посредством кластеризации откликов, предварительно аппроксимированных модельной функцией. Использование методики предположительно позволит улучшить качество реконструкции физических событий. При интерпретации полученных кластеров может быть извлечена важная физическая информация.

Область применения: Физика Высоких Энергий.

Экономическая эффективность/значимость работы — применение методики позволит повысить качество отбора событий, снизив долю событий, выпадающих из анализа. Таким образом статистически необходимое количество данных может быть исследовано за меньшее время работы экспериментальной установки.

Содержание

Введение

NA64 – это эксперимент с фиксированной мишенью на протонном суперсинхротроне (*SPS*), расположенном в Европейском центре ядерных исследований (*CERN*). Эксперимент выполнен в герметичной постановке, для которой детектирование экзотических частиц осуществляется по недостающей энергии.

Среди всего технического оборудования эксперимента выделяются калориметры – детекторы, предназначенные для измерения энерговыделения. Принцип работы калориметров основан на том, что энерговыделение пропорционально числу и энергии фотонов, образованных инициирующим пучком в материале детектора. Эти фотоны и формируют отклик детектора, попадая на Фотоэлектронные Умножители (ФЭУ).

Отклик детектора зависит от особенностей протекания электромагнитных ливней, а также от свойств пучка частиц, порождающих ливень. Выделение сигналов, соответствующих определенным сценариям представляет интерес при анализе результатов эксперимента.

В результате развития и популяризации Машинного Обучения появились техники, широко используемые в науке и инженерии. В области Физики Высоких энергий эти мощные инструменты успешно используются при решении ключевых задач анализа данных с ускорителей. Примеры использования Машинного Обучения для отбора событий представлены в работах: [1], [2]. В указанных статьях предлагается для отбора событий использовать свёрточные и графовые нейронные сети.

В настоящей работе предложен способ кластеризации откликов калориметра, представленных в виде набора параметров функции, аппроксимирующей сигнал детектора. Актуальность работы заключается в возможности улучшения реконструкции физических событий эксперимента NA64, в которой в настоящий момент разделение сигналов адронного и

электронного пучков производится путём простого порогового отбора сигналов. Цель работы заключается в реализации алгоритма кластеризации для определения групп калориметрических откликов, соответствующих адронной и электронной природе инициирующего пучка.

Задачи:

– Определение структуры работ в рамках научного исследования;

– Обзор литературы с целью определения наиболее подходящего метода кластеризации сигналов калориметров с учётом особенностей распределения параметров аппроксимирующей функции и критериев быстродействия;

– Формирование набора данных с использованием программных решений, разработанных в научной группе ТПУ/NA64;

– Проектирование и реализация конвейера подготовки данных для кластеризации;

– Осуществления кластеризации на подготовленных данных;

– Интерпретация полученных кластеров.

1.2 Алгоритмы кластеризации

1.2.1 K-Means

Алгоритм *K-Means* [3] кластеризует данные, разделяя их на *n* групп с одинаковой дисперсией, минимизируя критерий, известный как *инерция*. Для использования алгоритма необходимо указать предполагаемое число кластеров. Он хорошо масштабируется для большого количества данных и широко используется во многих задачах различных областей.

Алгоритм *K-Means* делит набор данных на выборки из *N* образцов, принадлежащих одному из K непересекающихся кластеров, каждый из которых характеризуется средним значением μ. Средние значения обычно называют «центроидами кластеров», но это неправильный термин, так как средние значения не являются точками из рассматриваемого набора данных, хоть и находятся в том же пространстве.

Алгоритм *K-Means* выбирает центроиды путём минимизации инерции или *внутрикластерной дисперсии*:

$$\sum_{i=0}^{n} \min_{\mu_i \in C}(\|x_i - \mu_i\|^2) \quad (1)$$

Инерция может рассматриваться как мера того, насколько кластеры внутренне однородны. Этот критерий обладает некоторыми недостатками:

– При использовании инерции делается предположение о том, что кластера выпуклы и изотропны, что не всегда соответствует действительности. Это плохо соответствует случаю вытянутых кластеров или кластеров с нерегулярной формой.

– Метрика инерции ненормализована, известно только что чем меньше её значение – тем лучше, нулевое значение – оптимальное. Но во многомерных пространствах Евклидово расстояние имеет тенденцию к завышенной оценке (по причине «проклятия размерности»). Применение алгоритмов уменьшения размерности к данным перед кластеризацией позволяет избежать этого и значительно повысить производительность.

В общих чертах алгоритм состоит из трех шагов. На первом шаге выбираются начальные центроиды, простейший способ – выбрать кандидатов среди точек данных. После инициализации алгоритм заключается в повторении следующих двух шагов: На первом из них каждый образец относится к ближайшему центроиду. На втором создаются новые центроиды, путем вычисления среднего значения среди всех образцов, отнесенных к предыдущему центроиду. Вычисляется разница между старыми и новыми центроидами и алгоритм повторяется до тех пор, пока эта разница не станет меньше порогового значения.

При достаточном числе итераций алгоритм гарантированно сходится, однако это может быть локальным минимумом. Это сильно зависит от инициализации центроидов. В результате вычисления зачастую повторяются несколько раз с различными инициализациями центроидов.

Средняя временна сложность алгоритма составляет $O(knT)$, k – число кластеров, n – размер набора данных, T – число итераций.

Сильные стороны алгоритма:

– Сравнительно высокая эффективность при простоте реализации;

– Высокое качество кластеризации;

– Возможность распараллеливания вычислений;

– Вариативность модификаций.

Слабые стороны алгоритма:

– Необходимость задавать число кластеров;

– Чувствительность к инициализации центроидов;

– Чувствительность к выбросам и зашумлению;

– Тенденция алгоритма сходиться к локальным минимумам.

1.2.2 Mini Batch K-Means

Алгоритм *Mini Batch K-Means* [4] является модификацией *K-Means*, в которой используются мини-батчи для сокращения времени вычислений, в то время как функция для оптимизации остаётся прежней. Мини-батчи это

подмножества из входных данных, которые выбираются случайным образом в каждой итерации. Такая технология значительно снижает объём вычислений, необходимый для сходимости к локальному решению. В сравнении с другими алгоритмами, уменьшающими время сходимости, алгоритм *Mini Batch K-Means* показывает результаты, которые в общем лишь немного хуже, чем стандартный алгоритм.

Алгоритм выполняет итерации между двумя основными этапами, схоже с тем, как это происходит в стандартном *K-Means*. На первом шаге *b* точек выбираются из датасета случайным образом и образуют мини-батч. Затем они соотносятся с ближайшими центроидами. На втором шаге значения координат центроидов обновляются. На контрасте с обычным *K-Means*, это происходит на основании каждой из точек. Для каждой точки мини-батча назначенный центроид пересчитывается, путём вычисления кумулятивного среднего для точки и всех других точек, прежде отнесённых к центроиду. Такая особенность привносит эффект снижения скорости изменения центроидов в единицу времени. Эти шаги повторяются до тех пор, пока сходимость не будет достигнута или не пройдёт заданное число итераций.

1.2.3 DBSCAN

Алгоритм *DBSCAN* (англ. Density-based spatial clustering of applications with noise, DBSCAN) [5] рассматривает кластеры как области высокой плотности, разделённые областями низкой плотности. Благодаря такому более универсальному подходу, кластеры, найденные этим алгоритмом, могут быть произвольной формы, в отличие от *K-Means*, который предполагает выпуклые кластеры. Ключевой для алгоритма *DBSCAN* является концепция основных точек, или же точек из областей высокой плотности. Кластер формируется некоторым числом основных точек, расположенных близко друг к другу (близость определяется различными метриками дистанции) и множеством не основных точек, которые находятся

близко к основным. Алгоритм параметризуется двумя значениями, *min_samples* и *ε*, которые формально определяют то, о чём мы говорим как о *плотности.* Большие значения *min_samples*, или напротив, маленькие *ε* сигнализируют о большей плотности, необходимой чтобы точки сформировали кластер.

Говоря более формально, мы определяем точку как основную, если в пределах расстояния *ε* от неё находятся *min_samples* других точек из набора данных, иначе их называют *соседями* основной точки. Это говорит нам о том, что основная точка находится в плотной области векторного пространства. Кластер, таким образом определяется набором основных точек, который может быть сформирован путём рекурсивного взятия одной основной точки, определения среди ей соседей других основных точек, определения среди её соседей тех точек, которые не являются основными, и так далее. Интуитивно понятно, что не основные точки, принадлежащие кластеру, вероятнее всего находятся рядом с его границей.

Каждая основная точка принадлежит кластеру по определению. Каждая не основная точка, находящаяся на расстоянии как минимум *ε* от любой основной точки, алгоритмом определяется как выброс.

Тогда как параметр *min_samples* контролирует то, насколько алгоритм терпим к шуму (на больших и зашумленных наборах данных часто полезно увеличивать значение этого параметра), параметр *ε* критически важно выбрать и настроить под конкретные данные и функцию дистанции. Когда это значение выбрано слишком низким, большинство точек данных не будут кластеризованы вовсе, (станут помечены как «шум»). Слишком высокие значения *ε* приводят к тому что близкие кластеры объединяются в один, в итоге весь набор данных представляется единственным кластером. Существуют некоторые эвристики для выбора этого параметра, например на основании вычисления значения в точке перегиба на графике расстояний до ближайших соседей [6].

Временная сложность алгоритма $O(n^2)$. В случае данных с небольшой размерностью и высокой разреженностью можно снизить это значение до $O(nlogn)$, используя такие структуры данных как *KD*-деревья, позволяющие эффективно выбирать все точки данных, находящиеся в пределах расстояния ε от выбранной точки.

Алгоритм *DBSCAN* отличается следующими преимуществами:

– не требует определять заранее ожидаемое число кластеров;

– способен отыскивать кластеры произвольной формы, даже кластеры, окруженные (без пересечения) другими кластерами;

– устойчив к вылетам и способен определять шумы;

– определяется только двумя параметрами и практически не зависит от порядка, в котором представлены точки в наборе данных;

Недостатки алгоритма:

– *DBSCAN* не способен раскластеризовать данные с сильно неоднородной плотностью, так как комбинация *min_samples* и $\boldsymbol{\varepsilon}$ не может быть подобрана одинаково подходящей для всех кластеров;

– Если данные и масштаб в достаточной степени не изучены, то определить осмысленное значение $\boldsymbol{\varepsilon}$ может быть затруднительно.

1.2.4 OPTICS

Алгоритм *OPTICS* [7] можно рассматривать как обобщение DBSCAN, использующее вместо одного значения *eps* множество значений из интервала. Точка *p* считается основной точкой, если по меньшей мере *MinPts* точек находятся в её ε-окрестности. В отличие от DBSCAN, *OPTICS* также рассматривает точки из более плотных кластеров, так что каждой точке приписывается основное расстояние. Это минимальное значение радиуса окрестности, необходимое чтобы классифицировать рассматриваемую точку как основную. Если точка не является основной, то основное расстояние не определено:

$$core - dist_{\varepsilon,MinPts(p)} = \begin{cases} \text{НЕОПРЕДЕЛЕНО}, & |N_\varepsilon(p)| < MinPts \\ Min\left(dist\left(p_i \subset N_\varepsilon(p)\right)\right) \end{cases}, \quad (2)$$

Расстояние достижимости от точки p до точки q определяется как максимальное из значений Евклидова расстояния (или другой метрики) между p и q и основного расстояния точки p. Если q не основная точка, то расстояния достижимости неопределено.

$$reachability - dist_{\varepsilon,MinPts}(o,p) =$$
$$\begin{cases} \text{НЕОПРЕДЕЛЕНО}, & |N_\varepsilon(p)| < MinPts \\ max\left(core - dist_{\varepsilon,MinPts}(p), dist(p,o)\right), & |N_\varepsilon(p)| \geq MinPts, \end{cases} \quad (3)$$

Если p и o являются ближайшими соседями, и если $\varepsilon' < \varepsilon$, можно предположить, что p и o принадлежат одному кластеру.

Результатом выполнения алгоритма является упорядочивание точек в соответствии со значениями расстояния достижимости. На основании этой информации может быть построен график достижимости, в котором плотность точек отражается на оси *Y*, а точки упорядочены так, что две ближайшие являются на графике соседними. Отсечение графика по единственному значению плотности даёт результат, похожий на результаты алгоритма *DBSCAN*: все точки выше отсечки классифицируются как выбросы, каждый разрыв на графике, при чтении его слева направо определяет новый кластер.

Используя алгоритм *OPTICS*, мы получаем возможность определения кластеров различной плотности. Помимо этого увеличивается вариативность способов соотнесения точек данных с кластерами. Однако время вычисления для него требуется несколько больше, авторы статьи [7] указывают на замедление в 1.6 раза по сравнению с *DBSCAN*.

1.2.5 Аффинное распространение ошибки

Аффинное Распространение Ошибки (англ. – Affinity Propagation) [8] определяет кластеры путём передачи сообщений между парами точек до тех пор, пока не будет достигнута сходимость. Небольшое число экземпляров, определенных как наиболее репрезентативные представители остальных точек, описывают весь набор данных. Сообщения, передаваемые между парами, отражают то, насколько подходит одной точке выступать представителем для другой, эта информация обновляется в зависимости от значений, характеризующих другие пары точек. Процесс итеративно повторяется до тех пор, пока сходимость, в результате которой определены все представители и соответствующие им кластеры, не будет достигнута.

Сообщения, пересылаемые между точками, принадлежат к двум категориям. Первая это *ответственность r(i, k)*, показатель того, насколько точно k описывает i. Вторая это *доступность a(i, k),* корректирующая показатель *r(i, k)*, учитывая значения от всех других точек, для которых k проверяется на роль представителя. В этом отношении представители выбираются, если 1) они достаточно близки ко многим другим точкам, 2) выбраны многими точками чтобы быть их представителями.

Для контроля числа кластеров в алгоритме используются два параметра: *предпочтение,* определяющее то, сколько экземпляров используется при выборе представителя, и *коэффициент затухания*, регулирующий сообщения *ответственности* и *доступности*, для того чтобы избежать численных осцилляций при пересчёте показателей.

Более формально, ответственность точки k относительно точки i задаётся следующим выражением:

$$r(i,k) \leftarrow s(i,k) - \max[\mathrm{a(i},k') + s(i,k')\forall k' \neq k], \quad (4)$$

где *s(i,k)* это сходство между точками i и k.

Доступность точки k относительно точки i определяется как:

$$a(i,k) \leftarrow \min\left[0, \mathrm{r(k,k)} + \sum_{i',s,t,i' \notin \{i,k\}} r(i',k)\right]. \quad (5)$$

Изначально, все значения r и a устанавливаются равными нулю, после чего вычисляются в итеративном процессе. Как уже было сказано, коэффициент затухания λ используется, чтобы избежать осцилляций:

$$r_{t+1}(i,k) = \lambda \cdot r_t(i,k) + (1-\lambda) \cdot r_{t+1}(i,k), \quad (6)$$

$$a_{t+1}(i,k) = \lambda \cdot a_t(i,k) + (1-\lambda) \cdot a_{t+1}(i,k), \quad (7)$$

где t это номер итерации.

Временная сложность алгоритма составляет $O(N^2T)$, где N – размер набора данных, T – число итераций до сходимости. Это делает алгоритм применимым в основном к небольшим и среднего размера наборам данных.

Преимущества алгоритма Распространения Аффинности:

– Не требует указания числа кластеров;

– Обладает хорошей точностью и эффективностью на небольших и средних наборах данных;

Недостатки алгоритма:

– Сложность при определении параметра *предпочтения*, который влияет на оптимальность кластеризации;

– Высокая временная сложность при использовании на больших объемах данных.

1.2.6 Метод Сдвига Среднего Значения

Кластеризация методом Сдвига Среднего Значения (англ. *Mean Shift*) [9] направлена на поиск уплотнений среди равномерной плотности точек. Это алгоритм, основанный на концепции центроидов. В этом методе положение центроида некоторой окрестности определяется как среднее значение координат всех точек в этой окрестности. На этапе постобработки

полученные центроиды фильтруются так, чтобы устранить дублирование и сформировать результирующий набор центроидов.

Положение центроида итерационно подстраивается с использованием алгоритма *Поиск Восхождением к Вершине*, который определяет локальный максимум путём оценки плотности вероятности. Имея позицию центроида *x* на шаге *t*, получаем следующее выражение для вычисления его позиции на шаге *t*+1:

$$x^{t+1} = x^t + m(x^t), \tag{8}$$

где m – вектор среднего сдвига, вычисляемый для каждого центроида, указывающий направление в сторону области максимальной плотности точек. Чтобы вычислить m, определим *N(x)* как подмножество соседей, находящихся в пределах определенного расстояния от *x*. Тогда m вычисляется согласно следующему выражению:

$$m(x) = \frac{\sum_{x_{j \in N(x)}} K(x_j - x) x_j}{\sum_{x_{j \in N(x)}} K(x_j - x)} - x, \tag{9}$$

где $K(x_j - x)$ – функция ядра, обычно используется распределение Гаусса.

Число кластеров определяется автоматически, на основании параметра *размера окна*, который определяет границы области, в которой выполняется поиск.

Алгоритм не столь хорошо масштабируется, так как требует множественное выполнение поиска ближайших соседей, во время вычислений. Однако он гарантированно сходится, можно оборвать выполнение алгоритма, когда изменения положений центроидов становятся незначительными.

Алгоритм Сдвига Среднего Значения обладает следующими преимуществами:

– не зависит от различных форм кластеров;

– способен обрабатывать произвольные пространства признаков;

– настраивается единственным параметром – размером окна;

– размер окна имеет физическое значение, в отличи, например,от параметров алгоритма K-Means.

Недостатки алгоритма:

– выбор параметра размера окна нетривиален;

– ширину канала часто приходится делать самонастраиваемым параметром.

1.2.7 Иерархическая кластеризация

Принцип Иерархической кластеризации [10] объединяет множество алгоритмов, основанных на формировании вложенных кластеров путем последовательного объединения или же разделения кластеров, сформированных на предыдущем шаге.

Иерархия кластеров представляется в виде дендрограммы. Корень дерева это уникальный кластер, содержащий в себе все остальные точки. Листья дерева представляют кластеры, состоящие лишь из одной точки.

Агломеративная кластеризация вычисляется путём использования подхода снизу-вверх. Каждая точка данных считается отдельным кластером, затем кластеры последовательно соединяются. *Критерий связи* определяет метрику, которая используется при объединении кластеров:

1) Одиночная связь. Расстояние между двумя кластерами определяется как наименьшее из попарных расстояний между точками, которые принадлежат этим кластерам:

$$L(r,s) = \min\left(D\left(x_{ri}, x_{sj}\right)\right). \quad (10)$$

2) Полная связь. Расстояние между кластерами это наибольшее из попарных расстояний между точками, которые принадлежат этим кластерам:

$$L(r,s) = \min(D\left(x_{ri}, x_{sj}\right)). \quad (11)$$

3) Усредненная связь. Расстояние между кластерами определяется как усредненная сумма расстояний от каждой из точек одного кластера до каждой из точек другого:

$$L(r,s) = \frac{1}{n_r n_s} \sum_{j=1}^{n_s} \sum_{i=1}^{n_r} D\left(x_{ri}, x_{sj}\right). \quad (12)$$

4) Связь Уорда. Расстояние между кластерами определяется как прирост суммы квадратов отклонений для каждого кластера:

$$\Delta = \sum_{i} (x_i - \bar{x})^2 - \sum_{x_i \in A} (x_i - \bar{a})^2 - \sum_{x_i \in B} (x_i - b)^2. \quad (13)$$

Объединению подлежат такие кластеры, которые приводят к минимальному изменению дисперсии. Используется в случае близко расположенных кластеров.

Агломеративная кластеризация обладает особенностью «богатый становится богаче», которая приводит к неравным размерам кластеров. В этом отношении одиночная связь является худшей стратегией, использование связи Уорда приводит к наиболее равным размерам. Однако некоторые метрики расстояния не могут использоваться в методе связи Уорда. В случае неевклидовых метрик усредненная связь является неплохой альтернативой. Одиночная связь, несмотря на то, что плохо устойчива к зашумленным данным, вычисляется достаточно эффективно и может быть использована в применении к наборам данных большого размера. Также одиночная связь даёт хорошие результаты для несферических форм кластеров.

Преимущества агломеративной кластеризации:

– не требует указания числа кластеров;

– может отражать реальную классификацию.

Недостатки агломеративной кластеризации:

– объединение кластеров необратимо;

– очень низкая скорость выполнения для больших наборов данных $O(N^2 \log N)$.

1.2.8 Спектральная кластеризация

Спектральная кластеризация [11] основана на спектральном анализе матрицы сходства между объектами. Он рассматривает каждую точку как узел графа и преобразует, таким образом, проблему кластеризации в проблему разбиения графа. Исполнение алгоритма состоит из трёх основных шагов:

1) Построение графа подобия, который представлен матрицей смежности *A*. Матрица смежности может быть сформирована на основании, например, *эпсилон-окрестностного графа* или с использованием метода *K-Means*.

2) Проецирование данных в пространство сниженной размерности. Этот этап выполняется чтобы учесть возможность того, что представители одного и того же кластера могут находиться на большом расстоянии в оригинальном пространстве. Снижение размерности пространства приводит к тому, что точки становятся ближе и могут впоследствии быть определены к одному кластеру при использовании традиционных алгоритмов кластеризации. Это осуществляется путём вычисления *Матрицы Лапласиана Графа.* Предварительно необходимо определить степень каждого узла графа, матрица степеней вычисляется как:

$$D_{i,j} = \begin{cases} \deg(v_i), & i = j \\ 0 \end{cases}, \quad (14)$$

где $deg(v_i)$ – число граней, выходящих из вершины. Матрица Лапласиана Графа в таком случае определяется как $L = D - A$. Для снижения размерности вычисляются собственные значения и собственные вектора этой матрицы. По собственным значениям определяется число кластеров в будущей кластеризации, матрица сниженной размерности, отображающая набор данных, строится из собственных векторов, выступающих в этой матрице столбцами.

3) Кластеризация данных. На этом этапе данные сниженной размерности кластеризуются одним из привычных алгоритмов – обычно алгоритмом *K-Means*.

Преимущества алгоритма Спектральной кластеризации следующие:

– применим к данным с произвольными формой и распределением;

– сокращение размерности делает анализ и визуализацию данных значительно проще;

– хорошо справляется с кластерами нелинейной формы.

Недостатки алгоритма:

– так как алгоритм основан на вычислении нескольких матриц, а также собственных значений и собственных векторов, на больших объемах данных временные затраты достаточно высокие;

– чувствителен к шумам и вылетам;

– требуется указание числа кластеров;

– хранение матрицы смежности затрачивает много памяти в случае большого набора данных.

1.1.9 BIRCH

Сбалансированное итеративное сокращение и кластеризация с использованием иерархии (*BIRCH*; англ – *Balanced Iterative reducing and clustering using hierarchies*) [12] – алгоритм интеллектуального анализа данных без учителя, используемый для иерархической кластеризации больших наборов данных. В этом алгоритме используется Дерево Признаков Кластеризации, сконструированное на основе входных данных. Листья дерева (Признаки Кластеризации) формируются из данных, над которыми было произведено сжатие с потерями. Каждому листу соответствует набор подкластеров Признаков Кластеризации, которые могут в свою очередь иметь дочерние листья.

Подкластеры Признаков Кластеризации содержат полную информацию, необходимую для кластеризации, что устраняет необходимость держать входные данные в памяти целиком. Эта информация включает:

– число точек в подкластере;

– линейная сумма (*n*-размерный вектор, хранящий сумму всех точек);

– квадратичная сумма (сумма квадратов *L2*-нормализации всех образцов);

– центроиды (чтобы избежать пересчета *linear sum / n_samples*);

– квадрат нормы позиций центроидов.

Алгоритм *BIRCH* использует два параметра – *пороговое значение* и *коэффициент ветвления*. Коэффициент ветвления ограничивает количество подкластеров в узле, а пороговое значение ограничивает расстояние от новой точки данных до уже существующих кластеров.

Алгоритм может рассматриваться как метод сокращения объёма данных, так как входные данные редуцируются до набора подкластеров, которые можно напрямую получить из листьев Дерева Признаков Кластеризации.

Процедуру кластеризации можно описать следующим образом:

– Новая точка данных помещается в корень Дерева Признаков Кластеризации, являющийся и узлом дерева. Затем она соотносится с подкластером корня, который после слияния будет иметь наименьший радиус, что регулируется пороговым значением и коэффициентом ветвления. Если в подкластерах есть дочерние узлы, то это будет повторяться итерационно, до тех пор, пока не будет достигнут лист. После того, как соответствующий подкластер найден, его параметры рекурсивно пересчитываются.

– Если радиус подкластера, образованного результате добавления точки, больше, чем квадрат порогового значения и если число подкластеров

больше значения коэффициента ветвления, то новой точке временно выделяется место. Берутся два наиболее удалённых подкластера и делятся на группы на основании расстояния между этими подкластерами.

– Если такой разделённый узел имеет покластер-родителя, и у него есть место для еще одного подкластера, то родительский подкластер делится на два. Если места нет, то тогда узел снова делится на два. Этот процесс повторяется рекурсивно до тех пор, пока не будет достигнут корень дерева.

1.1.10 Гауссовая Смесь

Модель Гауссовой Смеси [13] предполагает, что существует определенное число распределений Гаусса, и каждое из них описывает кластер. Таким образом, модель Гауссовой смеси группирует вместе точки, которые относятся к одному и тому же распределению.

Плотность вероятности *T*-мерного распределения Гаусса описывается формулой:

$$f(x|\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left(-\frac{1}{2}\frac{(\mathrm{x}-\ \mu)^T}{\Sigma}(\mathrm{x}-\mu)\right), \qquad (15)$$

где x – вектор входных данных, μ – вектор средних значений, Σ - матрица ковариации. Значения μ и Σ определяются с использованием техники *Минимизации Ожидания*. Минимизация Ожидания – это статистический алгоритм отыскания подходящих параметров модели.

Предположим, что нужно для набора данных определить метки k кластеров. Это значит, что имеется k распределений Гаусса с параметрами μ_1, μ_2, ..., μ_k и Σ_1, Σ_2, ..., Σ_k. Помимо этого каждому распределению соответствует ещё один параметр, определяющий плотность точек в распределении: π_1, π_2, ..., π_k. Тогда алгоритм Минимизации Ожидания можно описать следующим образом:

– **E-шаг**: для каждой точки x_i, рассчитывается вероятность того что она принадлежит к кластеру/распределению c_1, c_2, ..., c_k на основании следующей формулы:

$$r_{i,c} = \frac{\pi_c N(x_i; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'} N(x_i; \mu_{c'}, \Sigma_{c'})}, \quad (16)$$

$$\text{где } \pi = \frac{\text{число точек,отнесенных к кластеру}}{\text{общее число точек}}.$$

Это значение будет больше, если точка отнесена к верному кластеру и меньше в противном случае.

– **M-шаг**: Пересчитываются значения, характеризующие распределения:

$$\mu = n \sum_i r_{i,c} x_i, \quad (17)$$

$$\Sigma_c = n \sum_i r_{i,c} (x_i - \mu_c)^T (x_i - \mu_c), \quad (18)$$

$$\text{где } n = \frac{1}{\text{число точек,отнесенных к кластеру}}.$$

На основании обновленных значений, рассчитанных на этом шаге, вероятности для каждой точки вычисляются заново, и последующие обновления значений происходят итерационно. Процесс повторяется, чтобы максимизировать функцию логарифмического правдоподобия.

Алгоритм Гауссовой Смеси обладает следующими преимуществами:

– Гибкость. Практически любое распределение вероятностей может быть представлено в виде взвешенной суммы нормальных распределений.

– Устойчивость. Модель достаточно устойчива к выбросам, так как может учитывать наличие нескольких пиков в распределении.

– Скорость. Модель Гауссовой Смеси аппроксимирует исходные данные достаточно быстро.

– Пропущенные данные. Модель способна обрабатывать данные с пропущенными значениями через выделение недостающих переменных.

– Интерпретируемость. Параметры, такие как веса, средние значения, ковариации имеют известные и определённые интерпретации, что положительно сказывается на понимании структуры данных.

Недостатки алгоритма:

– Чувствительность к инициализации. Модель может быть чувствительной к стартовым значениям, особенно в случае смеси с большим числом компонент, что может негативно сказываться на сходимости.

– Предположение о нормальности данных. Модель подразумевает, что данные являются выборками из нормального распределения, что далеко не всегда подтверждается на практике.

– Число компонент. Выбор числа компонентов смеси может быть затруднительным, так как слишком большое число может приводить к переобучению, а слишком маленькое – к недообучению.

– Многомерные данные. Модель Гауссовой смеси становится вычислительно затратной, когда приходится иметь дело с многомерными данными, так как число параметров возрастает квадратично с ростом размерности.

1.2 Кластеризация откликов калориметров

1.2.1 Описание набора данных

Каждое попадание пучка частиц в телескоп детекторов будем называть *событием*. На этапе набора статистики дискретизованные отклики каждого детектора регистрируются в соответствии с событием, вызвавшим такой отклик. Показания детекторов объединяются в файлы, специфического формата, которые соответствуют некоторому временному интервалу.

Для извлечения калориметрических откликов был использован *C++* конвейер обработки данных, разработанный в научной группе *ТПУ/NA64*. В

результате был сформирован набор *.csv* файлов, содержащих следующую информацию: уникальный идентификатор события (*eventId*), уникальный идентификатор детектора (*detId*) и дискретизованный в 32 значения отклик детектора (*waveform*). Совокупный объем данных составил около 18 Гб (~ $1{,}5e^6$ строк).

По причине отсутствия предварительной обработки данные, образующие исходный набор, отличаются степенью своей достоверности. Помимо регулярных сигналов, там присутствуют также сигналы со значимым влиянием аппаратных шумов, а также сигналы, образованные недостаточно разнесенными во времени событиями.

Подобные особенности датасета, а также необходимость извлечения параметров функции, аппроксимирующей отклик калориметра для каждого события, обуславливают необходимость разработки конвейера для обработки сигналов детекторов.

1.2.2 Конвейер обработки данных

Код конвейера был написан на языке программирования Python, обладающего преимуществами быстроты разработки. Был использован следующий стек Python-библиотек:

– *Pandas* (чтение и агрегация файлов с данными);

– *Numpy* (представление данных в виде массивов, некоторые математические операции);

– *Scipy* (поиск пиков, аппроксимация);

– *Matplotlib* (визуализация данных);

– *Plotly* (визуализация данных).

Конвейер был спроектирован и запрограммирован как набор классов (*хэндлеров*), выполняющие некоторые атомарные процедуры обработки данных. Конвейер, описанный в виде диаграммы классов, представлен на рис. 1:

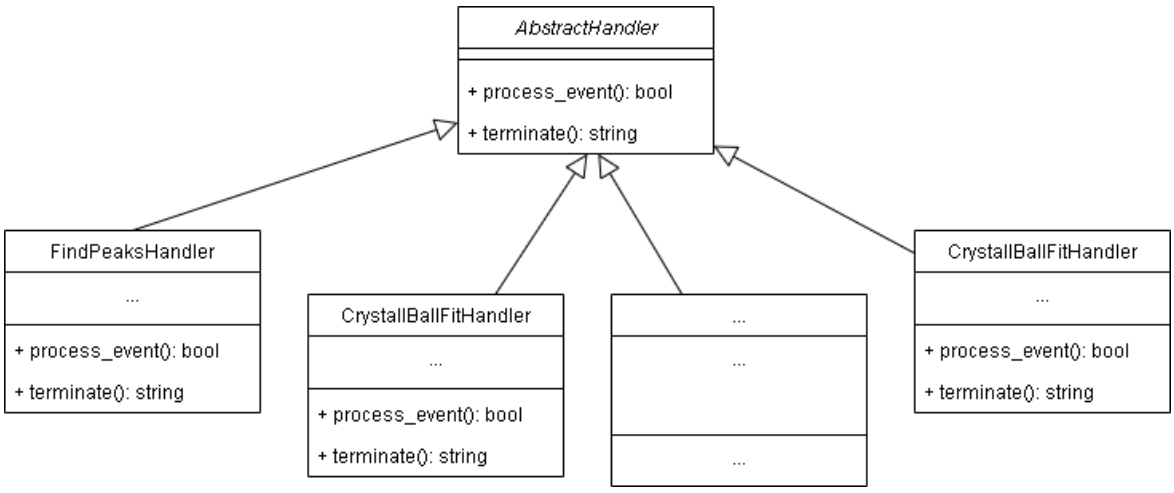

Рисунок 1 – Диаграмма классов конвейера обработки

Общий для всех хэндлеров функционал определяется абстрактным классом *AbstractHandler*. Метод *process_event()* обобщает действия при обработке каждого события, метод *terminate()* обобщает действия, выполняемые после того как все события будут пропущены через конвейер (например – визуализация, расчет статистических показателей). Эти методы реализуются в классах-наследниках, выполняющих функции фильтрации сигналов по количеству пиков, визуализации, аппроксимации сигнала и извлечения параметров аппроксимирующей функции и т.д.

1.2.3 Модель калориметрического отклика

Функция Crystal Ball [14] это функция плотности вероятности, используемая обычно для описания процессов, сопровождающихся потерями энергии. Она состоит из Гауссова ядра, а, начиная с определенного значения, из полиномиального хвоста. Функция является непрерывной также как и её первая производная. Функция Crystal Ball задаётся следующим выражением:

$$f(x,\alpha,n,\overline{x},\sigma,S) = S \cdot N \cdot \begin{cases} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right), \frac{x-\bar{x}}{\sigma} > -\alpha \\ A * \left(B - \frac{x-\bar{x}}{\sigma}\right)^{-n}, \frac{x-\bar{x}}{\sigma} \leq -\alpha \end{cases}, \quad (19)$$

где:

$$A = \frac{n}{|\alpha|}^{n} \cdot \exp\left(-\frac{|\alpha|^2}{2}\right),$$

$$B = \frac{n}{|\alpha|} - |\alpha|, \qquad N = \frac{1}{\sigma(C + D)},$$

$$C = \frac{n}{|\alpha|} \cdot \frac{1}{n-1} \cdot \exp\left(-\frac{|\alpha|^2}{2}\right),$$

$$D = \sqrt{\frac{\pi}{2}} \cdot \left(1 + \operatorname{erf}\left(\frac{|\alpha|}{\sqrt{2}}\right)\right);$$

$\bar{x}$ – точка, в которой функция плотности вероятности переходит от распределения Гаусса к полиномиальному хвосту; n – степень полиномиального хвоста, эмпирически значение n было выбрано равным 2; σ и α – параметры формы аппроксимирующей функции; S – коэффициент масштаба.

Выбор такой модельной функции обусловлен хорошим соответствием с формой отклика детекторов, наличием лишь 3 параметров, имеющих физическую интерпретацию и быстротой вычисления, относительно к примеру функции интегральной свёртки распределений Гаусса и Ландау [15], которая тоже используется для описания энерговыделения в калориметре.

Пример аппроксимации незашумленного события с использованием Метода Наименьших Квадратов, реализованного в библиотеке *scipy* представлен на рис. 2:

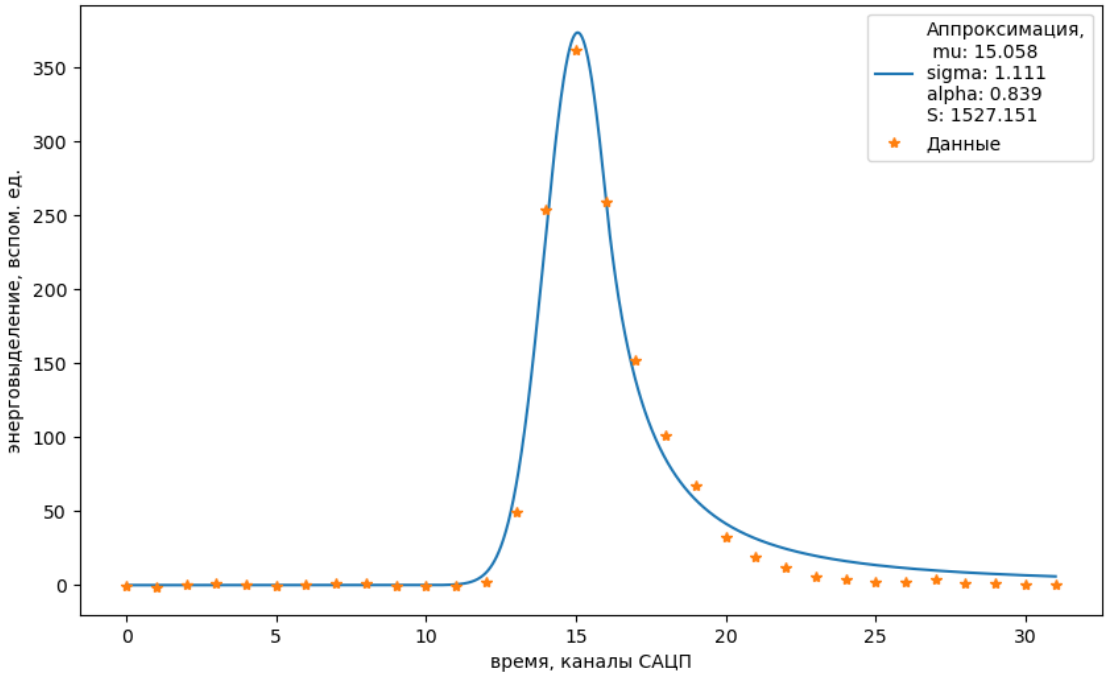

Рисунок 2 – Аппроксимация отклика калориметра *ECAL* функцией Crystal Ball

1.2.4 Кластеризация параметров

В результате аппроксимации исходного набора сигналов, были получены распределения параметров функции Crystal Ball, представленные на рис. 3 (параметр $\bar{x}$, положение максимального значения сигнала, был исключен из дальнейшего рассмотрения в силу того, что не отражает значимой физической информации и его значение во многом определяется эффектом временного дрейфа сигналов калориметра):

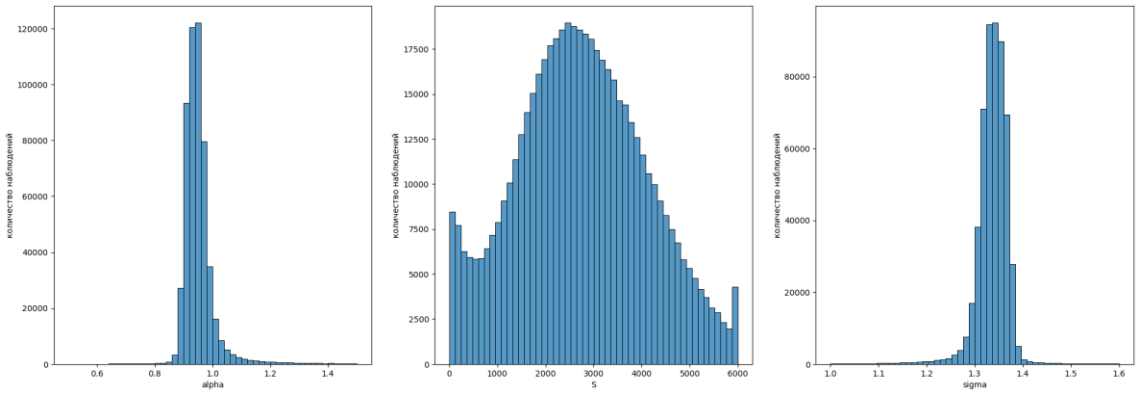

Рисунок 3 – Распределения параметров функции Crystal Ball

Большинство алгоритмов кластеризации используют для определения сходства между точками различные метрики дистанции. Поэтому для

осуществления качественной кластеризации необходимо подготовить данные – привести их к одному масштабу [16]. Распределения параметров, показанные на рис. 3 имеют явно не Гауссовый характер, поэтому использование стандартизации для масштабирования данных не подходит. Альтернативным вариантом является Минимакс нормализация, которая применима для большинства статистических распределений. Попарные распределения параметров после применения нормализации с использованием *MinMaxScaler* из библиотеки *sklearn* представлены на рис. 4:

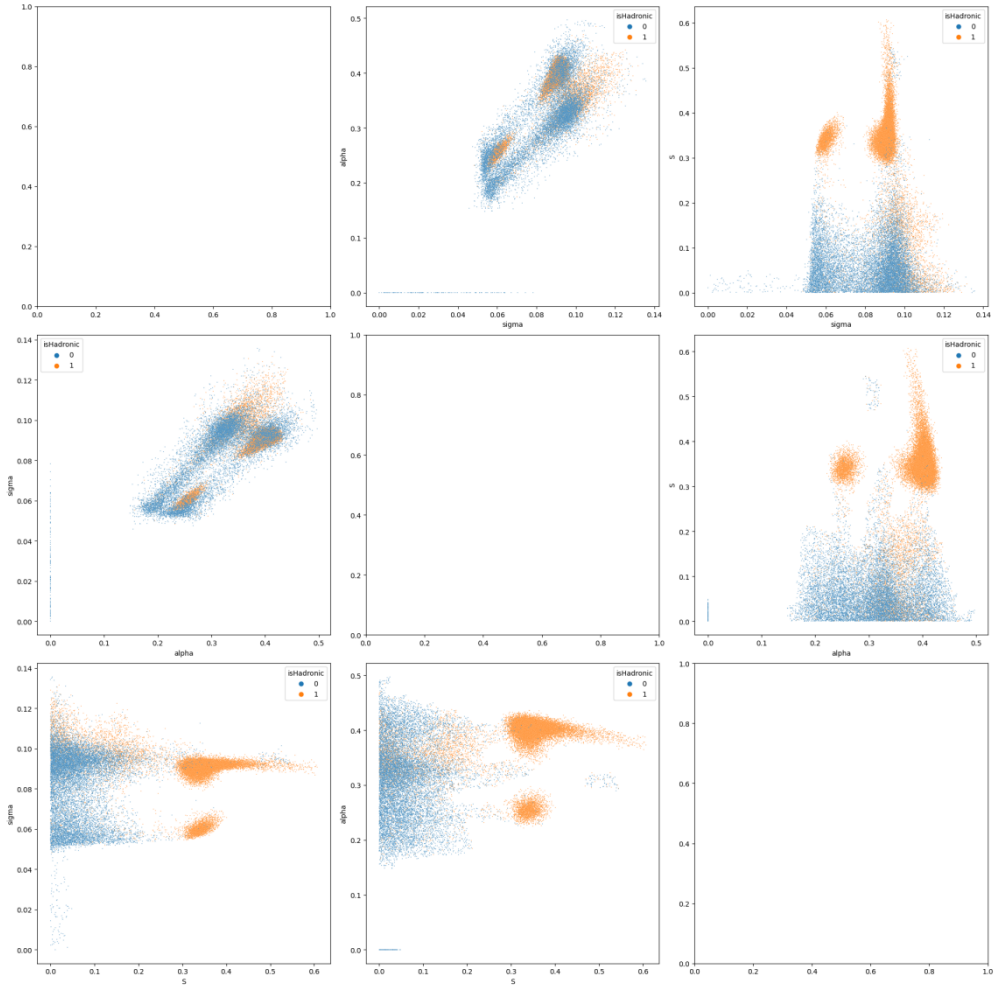

Рисунок 4 – Попарные распределения параметров функции Crystal Ball после нормализации с цветовым разделением в соответствии с природой инициирующего пучка

После проведения нескольких экспериментов для осуществления кластеризации параметров был выбран алгоритм *DBSCAN*, реализованный в библиотеке *sklearn*. Выбор обосновывается тем, что *DBSCAN* не требует

указания числа кластеров в качестве параметра, применим к распределениям различной формы, а также имеет приемлемое время вычисления и потребление памяти на рассматриваемом наборе данных, относительно других алгоритмов.

Как было указано в обзоре литературы, кластеризация методом *DBSCAN* определяется двумя параметрами – ε (максимальное расстояние между двумя точками, при котором они могут рассматриваться как соседи) и *min_samples* (минимальное количество точек, необходимое для образования кластера). Для выбора *min_samples* в литературе предлагается следующая эмпирика [17]:

$$min\_samples = \text{размерность набора данных} \cdot 2.$$

Определение значения ε осуществлялось в соответствии с [6] следующим образом:

Было определено среднее расстояние между *min_samples* ближайшими соседями для каждой точки с использованием классификатора *NearestNeighbors* из библиотеки *sklearn*. Значения средних дистанций были отсортированы в порядке возрастания. Значение ε было выбрано как значение в точке перегиба на графике сортированных средних дистанций (рис. 5):

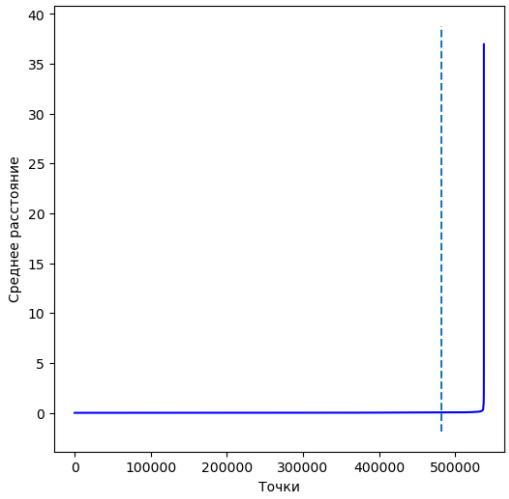

Рисунок 5 – Среднее расстояние от точки до *min_samples* ее ближайших соседей

1.2.5 Результаты кластеризации

Распределения параметров аппроксимирующей функции, отнесенных к соответствующим кластерам представлены на рис. 6 (точки, не отнесённые к действительным кластерам исключены из рассмотрения):

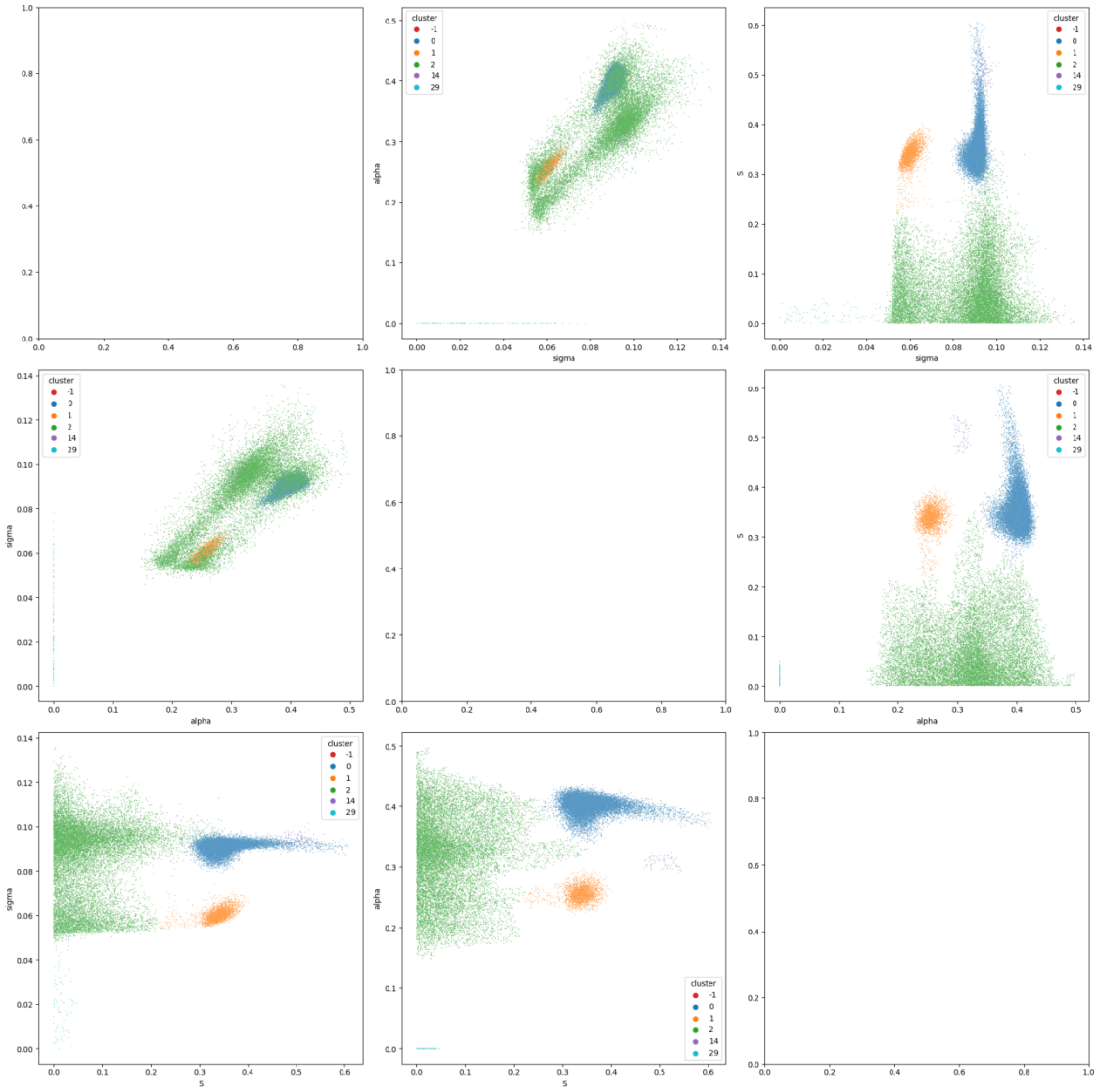

Рисунок 6 – Распределения кластеризованных параметров функции Crystal Ball

Оценивая полученные распределения, можем заметить, что кластеризация хорошо различает сигналы, соответствующие адронному и электронному пучку в области высокого энерговыделения (S-параметр функции Crystal Ball). Однако, адронные сигналы, соответствующие небольшому энерговыделению, плохо различимы с электронными, так как

распределены с невысокой плотностью внутри и вокруг электронных кластеров.

Данные, отражающие количественные соотношения сигналов в полученных кластерах представлены в таблице 1:

Таблица 1 – Доли адронной и электронной компоненты в полученных кластеров

Кластер	Доля сигналов от адронного пучка	Доля сигналов от электронного пучка
0	0.998	0.002
1	0,979	0,021
2	0,237	0,763
14	0,766	0,234
29	0,0	1,0
-1	0,761	0,239

Полученные соотношения свидетельствуют о том, что в кластерах 0, 1 и 29 с высокой точностью выделены сигналы от адронной компоненты, а также то, что кластер 2 и кластер соответствуют плохо разделимым событиям.

Для оценки качества кластеризации была построена конфузионная матрица классификации сигналов на адронные и электронные (рис. 7):

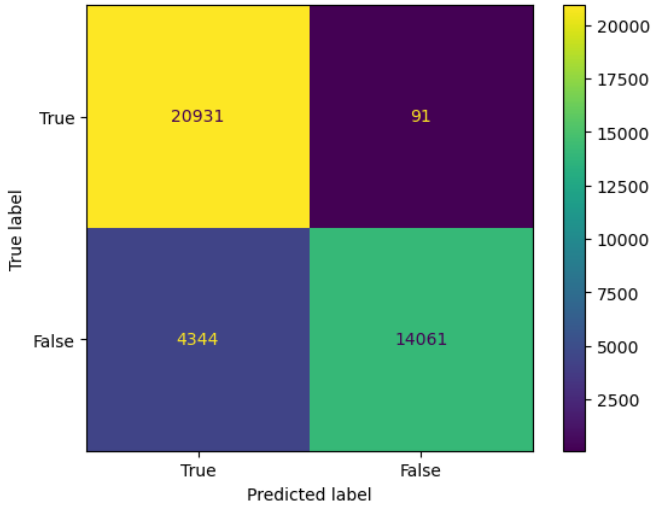

Рисунок 7 – Конфузионная матрица классификации откликов от адронных и электронных пучков

Значение F1-критерия, вычисленное по результатам классификации составляет 0,904, что свидетельствует о достаточно высоком качестве разделения откликов.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту :

Группа	ФИО
8ПМ1И	Крамойкину Ивану Алексеевичу

Школа	Инженерная школа информационных технологий и робототехники	Отделение школы (НОЦ)	ОИТ / Отделение информационных технологий
Уровень образования	Магистратура	**Направление**	09.04.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Бюджет проекта – не более 500000 рублей.* *Заработная плата инженера составляет 226000 рублей, руководителя 90000 рублей.*
2. *Нормы и нормативы расходования ресурсов*	*Накладные расходы – 15%,* *Районный коэффициент – 30%*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Отчисления на уплату во внебюджетные фонды – 30%*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Анализ потенциальных потребителей результатов исследования, оценка качества и перспективности проекта по технологии QuaD, SWOT-анализ*
2. *Разработка устава научно-технического проекта*	*Инициация проекта: определение заинтересованных сторон проекта, целей и результатов проекта*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*План проекта, определение трудоемкости выполнения работ, разработка графика проведения научного исследования, расчет бюджета разработки*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Описание потенциального эффекта*
Перечень графического материала:	
1. *Оценочная карта для QuaD-анализа разработки* 2. *Диаграмма Ганта* 3. *Матрица SWOT* 4. *График проведения НТИ* 5. *Бюджет затрат* 6. *Потенциальные риски*	

Дата выдачи задания	

Консультант:

Должность	ФИО	Ученая степень, Звание	Подпись	Дата
Доцент	Спицына Л.Ю.	к.экон.н		

Студент:

Группа / Group	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		

2 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

2.2 Предпроектный анализ

На этапе предпроектного анализа определяются цель проекта и его содержание, внутренние и внешние потребители, сильные и слабые стороны, возможности и угрозы, оценивается готовность проекта к выходу на рынок.

В работе производятся кластеризация откликов электромагнитных калориметров эксперимента NA64 CERN и анализ полученных кластеров. Была разработана методика, в которой сигналы аппроксимировались аналитической функцией, после чего параметры аппроксимации подлежали кластеризации. Кластеризованные сигналы предположительно соответствуют определённым сценариям развития ливней вторичных частиц в теле детектора, качество и информативность кластеризации можно оценить, используя некоторые методы визуализации и статистические критерии, программные инструменты для осуществления такого анализа были также подготовлены в ходе исследования и разработки.

Рассматриваемую методику предполагается использовать для извлечения физической информации и фильтрации данных на этапе анализа результатов эксперимента.

2.2.1 Потенциальные потребители разработки

Так как методика основывается на описании сигнала детектора аналитической функцией, это сильно ограничивает вариативность возможных потребителей исследования. Существует множество разновидностей калориметров, но результаты исследования будут применимы только к гетерогенным сцинтиллятор-свинцовым электромагнитным калориметрам, аналогам детектора ECAL эксперимента NA64, потому как для описания откликов других типов детекторов нужно подбирать и исследовать другие модельные функции.

В целевой сегмент потребителей исследования входят лаборатории, производящие эксперименты Физики Высоких Энергий на установках,

включающих в себя гетерогенные сцинтиллятор-свинцовые электромагнитные калориметры, или научные коллективы, ВУЗы и НИИ, связанные с анализом результатов таких экспериментов.

Причины заинтересованности исследованием у потенциальных потребителей отличаются: специалисту-физику важны результаты использования методики на данных, так как они содержат физическую информацию, специалист-аналитик может использовать исследование как основу для собственной процедуры обработки данных.

Таблица 1 – Сегментирование рынка

		Потребители	
		Специалисты-физики	Специалисты-исследователи данных
Методика кластеризации	Использование нейронных сетей	-	+
	Физическая модель	+	+

Были рассмотрены два варианта осуществления кластеризации сигналов. Карта сегментирования рынка, приведенная в таблице 1, показала, что выбранный вариант удовлетворяет интересы обоих групп потенциальных потребителей, поэтому в дальнейшем будет рассматриваться только вариант исполнения с физической моделью отклика.

В литературе представлены некоторые работы, в которых также проводилась кластеризации сигналов электромагнитных калориметров. Преимущественно для таких целей используются решения с использованием нейронных сетей. Однако, учитывая, что физическая модель откликов в этих исследованиях не используется, можно предположить, что результаты будут содержать значительно меньше физической информации, чем вариант кластеризации параметров физической модели. Поэтому, несмотря на некоторые сходства, назвать такие исследования конкурентными всё же нельзя, отличия в целях и методах кластеризации слишком существенны.

2.2.2 Анализ конкурентоспособности проекта

Как было отмечено выше, рассматриваемое исследование имеет довольно весомые отличия в цели и методике от других исследований, посвященных кластеризации откликов. Сравнение с существующими «конкурентами» представляется затруднительным и едва ли может быть выполнено с необходимой строгостью. Поэтому для оценки перспективности проекта был выбран Quad-анализ.

Технология Quality Advisor (Quad) – это гибкий инструмент измерения характеристик, описывающих качество новой разработки и её перспективность на рынке. Используя этот инструмент можно обосновать решение о целесообразности вложения денежных средств. Каждый показатель оценивается экспертным путем по стобалльной шкале, где 1 – наиболее слабая позиция, а 100 – наиболее сильная. Результат оценки, произведенной по этой технологии, представлен в таблице 2:

Таблица 2 – Quad-анализ разработки

Критерии оценки	Вес критерия	Средний балл	Максимальный балл	Относите-льное значение (3/4)	Средневзвешенное значение (5х2)
Технические критерии					
Производительность	0.07	90	100	0.9	0.063
Отказоустойчивость	0.2	80	100	0.8	0.160
Унифицированность	0.05	60	100	0.6	0.030
Безопасность	0.05	60	100	0.6	0.030
Потребность в ресурсах памяти	0.13	90	100	0.9	0.117
Функциональная мощность	0.1	85	100	0.85	0.085
Простота эксплуатации	0.03	70	100	0.7	0.021
Масштабируемость	0.06	85	100	0.85	0.051
Экономические критерии					
Конкурентоспособность продукта	0.07	50	100	0.5	0.035
Перспективность рынка	0.07	50	100	0.5	0.035
Цена	0.1	60	100	0.6	0.060
Финансовая эффективность научной разработки	0.07	75	100	0.75	0.053
Итого	1				0.740

Совокупное средневзвешенное значение критериев Quad-анализа составляет 0,74, попадает в интервал [0,60; 0,79]. Принято считать, что значения из этого интервала характеризуют перспективность разработки как «выше среднего». На основании результатов Quad-анализа можно предположить что исследование является способным для конкуренции с возможными аналогами.

2.2.3 SWOT-анализ

SWOT-анализ это техника стратегического планирования и стратегического управления, используемая для определения сильных сторон, слабостей, возможностей или угроз относительно планирования проекта. Результаты анализа представлены в таблице 3:

Таблица 3 – SWOT-анализ

Сильные стороны: С1. Возможность классификации событий на основании только откликов калориметров. С2. Невысокое вычислительное время. С3. Возможность определения зашумленных событий. С4. Масштабируемость конвейера обработки данных и доступность для модификаций.	Слабые стороны: Сл1. Зависимость качества кластеризации от аппроксимирующей функции. Сл2. Сложность оценки результатов кластеризации. Сл3. Зависимость от структуры входных данных.
Возможности: В1. Исключить классификационные детекторы из телескопа детекторов, снизив операционные расходы. В2. Извлечение физической информации из кластеризованных параметров аппроксимирующей функции.	Угрозы: У1. Влияние геополитической ситуации на возможность российских команд участвовать в экспериментах CERN. У2. Изменение параметров первичного пучка может повлиять на отклики детекторов и привести к ошибочной кластеризации.

Определим наличие зависимости между компонентами SWOT-анализа чтобы установить значимость потребности в стратегических изменениях.

Таблица 4 – Проектная матрица сильных сторон и возможностей

	С1	С2	С3	С4
В1	+	-	+	-
В2	+	-	+	-

Анализ предыдущей матрицы позволяет определить следующие связи между сильными сторонами и возможностями: С1В1, С1В2, С3В1, С3В2.

Таблица 5 – Проектная матрица слабых сторон и возможностей

	Сл1	Сл2	Сл3
В1	-	+	-
В2	+	-	-

Слабые стороны связаны с возможностями в случаях: Сл1В2, Сл2В1.

Таблица 6 – Проектная матрица сильных сторон и угроз

	С1	С2	С3	С4
У1	-	-	-	-
У2	+	-	+	-

Сильные стороны связаны с угрозами в следующих случаях: С1У2, С3У2.

Таблица 7 – Проектная матрица слабых сторон и угроз

	Сл1	Сл2	Сл3
У1	-	-	-
У2	+	-	-

Анализ проектной матрицы позволяет определить связь между слабыми сторонами и угрозами в случае Сл1У2.

Итоговая матрица SWOT-анализа, составленная с использованием предыдущих результатов, показана в таблице 8:

Таблица 8 – Итоговая матрица SWOT-анализа

	Сильные стороны: С1. Возможность классификации событий на основании только откликов калориметров. С2. Невысокое вычислительное время. С3. Возможность определения зашумленных событий. С4. Масштабируемость конвейера обработки данных и доступность для модификаций.…	Слабые стороны: Сл1. Зависимость качества кластеризации от аппроксимирующей функции. Сл2. Сложность оценки результатов кластеризации. Сл3. Зависимость от структуры входных данных.
Возможности: В1. Исключить классификационные детекторы из телескопа детекторов, снизив операционные расходы. В2. Извлечение физической информации из кластеризованных параметров аппроксимирующей функции.	Некоторые сильные стороны проекта определяют вероятность реализации возможностей. Если укреплять сильные стороны: повышать качество классификации и фильтрации событий, то обе возможности становятся легкореализуемыми.	Некоторые слабые стороны негативно влияют на вероятность осуществления возможностей, однако они неустранимы, как фундаментальные ограничения выбранной методики. Усиление слабых сторон – поднятие качества аппроксимации, автоматизация оценки кластеризации снизит влияние слабых сторон на реализацию возможностей.
Угрозы: У1. Влияние геополитической ситуации на возможность российских команд участвовать в экспериментах CERN. У2. Изменение параметров первичного пучка может повлиять на отклики детекторов и привести к ошибочной кластеризации.	Несмотря на то, что модернизация экспериментальной установки – достаточно регулярное явление, прецедентов со значительным влиянием на показания приборов еще не происходило, сформированная технологическая система достаточно устойчива. Угроза У2 может потенциально ослабить сильные стороны проекта, но маловероятно.	Угроза У1 неподконтрольна, поэтому её риск – один из тех что приходится принимать без возможности повлиять. Реализация угрозы У2 может проявиться в том что слабые стороны станут еще весомее, однако, маловероятно.

Подводя итоги SWOT-анализа можно заключить, что стратегические изменения при планировании проекта не требуются. Сильные стороны проекта положительно связаны с возможностями. Слабые стороны представляют некоторую угрозу реализациям возможностей, но их влияние можно ослабить в результате модернизаций и доработок методики, хоть и не избавиться полностью. Угрозы проекта неподконтрольны, это необходимые риски, которые приходится принимать. Будучи реализованными, они могут

ослабить сильные стороны, проявить слабые, но вероятность такого невысока.

2.2.4 Оценка готовности разработки к коммерциализации

В этой главе проект будет рассмотрен для оценки готовности к коммерциализации. Существуют случаи, в которых исследование может быть модифицировано до состояния продукта.

Таблица 9 – Бланк оценки степени готовности разработки к коммерциализации

№ п/п	Наименование	Степень проработанности исследования	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	4	4
2.	Определены перспективные направления коммерциализации научно-технического задела	3	2
3.	Определены отрасли и технологии (товары, услуги) для предложения на рынке	2	2
4.	Определена товарная форма научно-технического задела для представления на рынок	3	3
5.	Определены авторы и осуществлена охрана их прав	3	3
6.	Проведена оценка стоимости интеллектуальной собственности	2	2
7.	Проведены маркетинговые исследования рынков сбыта	2	2
8.	Разработан бизнес-план коммерциализации научной разработки	1	1
9.	Определены пути продвижения научной разработки на рынок	2	2
10.	Разработана стратегия (форма) реализации научной разработки	4	4
11.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	5	5
12.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	2	2
13.	Проработаны вопросы финансирования коммерциализации научной разработки	2	2
14.	Имеется команда для коммерциализации научной разработки	4	4
15.	Проработан механизм реализации разработки	5	5
	ИТОГО БАЛЛОВ:	44/75	43/75

Рассматриваемое исследование является уникальным и узкоспециализированным, оно рассчитано под один конкретный эксперимент и не предполагает выход на рынок и рост числа потребителей. Оценка,

произведенная в таблице 9, показывает значения около половины баллов от возможного максимума. Это приблизительно показывает, что готовность проекта к коммерциализации средняя, однако учитывая уникальность предполагаемого потребителя, коммерциализация кажется излишней.

Единственный приемлемый способ – коммерциализация путём передачи ноу-хау более компетентной в экономической области команде.

2.3 Инициация разработки

В рамках этого раздела будут рассмотрены цели проекта, а также ожидания, ограничения и требования со стороны потребителей.

Информация о потенциальных потребителях и их ожиданиях относительно рассматриваемого проекта представлена в следующей таблице:

Таблица 10 – Потребители проекта

Потребители проекта	**Ожидания потребителей проекта**
Физики и аналитики из научных групп, занимающиеся анализом результатов экспериментов Физики Высоких Энергий, в частности – из группы TPU/NA64.	Определение кластеров, соответствующих определенным сценариям развития ливней; Определение зашумленных событий; Быстродействие обработки данных.

Информация о целях и ограничениях проекта, а также об ожидаемых результатах отражена в таблице 11:

Таблица 11 – Цели и ожидаемые результаты проекта

Цели проекта	Реализовать алгоритм кластеризации откликов калориметров эксперимента NA64 CERN для определения групп событий, соответствующих определенным сценариям разлития ливней вторичных частиц.
Ожидаемые результаты проекта	Программная реализация автоматизированного конвейера предобработки данных и алгоритма кластеризации.
Критерии принятия проекта	Соответствие полученных кластеров действительной физике оценивается экспертом в области калориметрии.
Требования к результатам проекта	Завершение проекта в установленные сроки
	Надежность реализации алгоритма
	Возможность определения недостоверных событий
	Временная эффективность при характерных объемах входных данных

2.3.1 Организационная структура проекта

В таблице 12 представлена рабочая группа разработки, определена роль и основные функции каждого участника в разработке.

Таблица 12 – Рабочая группа разработки

№	ФИО, основное место работы, должность	Роль в разработке	Функции	Трудозатраты, час
1	Губин Е. И., ТПУ ОИТ ИШИТР, доцент	Руководитель	Утверждение основных разделов, выдача заданий к использованию, координирование деятельности исполнителя	20
2	Крамойкин Иван Алексеевич, ТПУ ОИТ ИШИТР, магистрант гр. 8ПМ1И	Исполнитель	Выполнение поставленных задач	320
Итого:				340

2.3.2 Положения и ограничения

Положения и ограничения представлены в следующей таблице:

Таблица 13 – Положения и ограничения

Фактор	Ограничения/положения
1. Бюджет проекта	500 000 руб.
1.1 Источники финансирования	Общие фонды / Заём в банке
2. Сроки проекта:	1 Февраля 2023 – 1 Июня 2023
2.1 Дата принятия плана управления проектом	5 Февраля 2023
2.2 Дата завершения проекта	1 Июня 2023

Определение требований и пожеланий к проекту со стороны потребителей позволяет осуществить планирование проекта и оценку его эффективности.

2.4 Планирование управления разработкой

Планирование управления разработкой помогает обеспечить более эффективную и успешную реализацию проекта. Оно включает в себя расписание работ и бюджета.

2.4.1 План разработки

Для того чтобы определить расписание работ в рамках проекта необходимо определить ключевые этапы проекта и их длительности. План работ представлен в таблице 14.

Диаграмма Ганта – это визуальное представление графика работ, построенное согласно плану проекта. На ней отражены задачи, последовательность их выполнения, исполнители задания и приблизительные временные затраты при выполнении задания. Диаграмма Ганта для рассматриваемого проекта представлена в таблице 15.

Таблица 14 – Временные рамки проектирования и исследования

Задание	Трудоёмкость выполнения задания						Длительность выполнения задания в рабочих днях, T_{pi}		Длительность выполнения задания в календарных днях, T_{ki}	
	t_{min}, человеко-дней		t_{max}, человеко-дней		$t_{ожi}$, человеко-дней					
	Руководитель	Исполнитель	Руководитель	Исполнитель	Руководитель	Исполнитель	Руководитель	Исполнитель	Руководитель	Исполнитель
Составление задания для исследования	3		7		5		5		7	
Обзор литературы		4		7		5		5		7
Определение области и границ исследования	4		7		5		5		7	
Извлечение данных		5		10		7		7		10
Подготовка инфраструктуры конвейера обработки		7		14		10		10		14
Выбор и модификация аппроксимирующей функции		10		20		15		15		21
Настройка алгоритма кластеризации		1		2		1		1		1
Анализ кластеризации		7		14		10		10		14
Расчет статистики Байеса по результатам кластеризации		3		7		5		5		7
Экспертная оценка результатов кластеризации	5		7		6		6		8	
Оформление итогового отчета		7		12		9		9		13

Таблица 15 – Диаграмма Ганта

№	Задание	Исполняющий	T_{ki}, дни	Длительность задания																
				Февраль				Март				Апрель				Май				
				1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	Составление задания для исследования	Руководитель	7																	
2	Обзор литературы	Исполнитель	7																	
3	Определение области и границ исследования	Руководитель	7																	
4	Извлечение данных	Исполнитель	10																	
5	Подготовка инфраструктуры конвейера обработки	Исполнитель	14																	
6	Выбор и модификация аппроксимирующей функции	Исполнитель	21																	
7	Настройка алгоритма кластеризации	Исполнитель	1																	
8	Анализ кластеризации	Исполнитель	14																	
9	Расчет статистики Байеса по результатам кластеризации	Исполнитель	7																	
10	Экспертная оценка результатов кластеризации	Руководитель	7																	
11	Оформление итогового отчета	Исполнитель	13																	
	– Исполнитель		– Руководитель																	

2.4.2 Бюджет проекта

Бюджет проекта включает в себя все затраты, связанные с выполнением проекта. Стоимость рассматриваемого проекта состоит из:

– затрат на основную и дополнительную заработные платы исполнителей;

– стоимость специального оборудования;

– затрат на социальные выплаты;

– прочих расходов.

2.4.2.1 Стоимость специализированного оборудования

Этот раздел включает в себя все затраты, связанные с покупкой специализированного оборудования, необходимого для выполнения проекта, они отражены в таблице 16:

Таблица 16 – Стоимость специализированного оборудования

Наименование	**Количество**	**Стоимость за единицу, руб**	**Общая стоимость оборудования, руб**
Персональный компьютер	1	40 000	40 000
Итого, руб	**40000**		

Стоимость специализированного оборудования рассчитывается в виде амортизационных отчислений.

Амортизация – это процесс переноса по частям стоимости основных средств и нематериальных активов по мере их физического или морального износа на себестоимость производимой продукции.

Рассчитаем амортизационные отчисления линейным способом. Стоимость оборудования составила 40000 руб. Срок полезного использования компьютера составляет 3 года. Тогда процент ежегодных отчислений на амортизацию ПК рассчитывается следующим образом

$$N_P = \frac{1}{3} \cdot 100\ \% = 33{,}33\ \%,$$

Амортизационные отчисления за ПК с учётом длительности проекта составили:

$$D_P = 30000 \cdot \frac{N_D}{100\ \%} \cdot \frac{T}{365} = 40000 \cdot \frac{33{,}33\ \%}{100\ \%} \cdot \frac{109}{365} = 3981 \text{ руб},$$

где T – число рабочих дней.

2.4.2.2 Заработная плата

Сумма отчислений, приходящихся на выплаты сотрудникам, рассчитывается на основании интенсивности труда при выполнении задания и установленной системы окладов и тарифов.

Отраслевая система оплаты труда в ТПУ предполагает следующий состав заработной платы:

1) оклад в ТПУ определяется соответственно занимаемой должности;

2) поощрительные выплаты – устанавливаются начальником департамента за эффективную работу и различного рода дополнительную ответственность;

3) другие выплаты, районный коэффициент

Возьмём значение премиального коэффициента оплаты труда равное 30 %, коэффициент доплат и надбавок примем равным 20 %. Основная заработная плата сотрудника определяется следующей формулой:

$$S_{\mathrm{b}} = S_r \cdot T_w, \quad (1)$$

где S_r – регулярный оклад работника;

T_{p}- длительность работы в рабочих днях.

Дополнительная заработная плата:

$$S_{\mathrm{add}} = 0{,}15\, S_{\mathrm{b}} \quad (2)$$

Средняя дневная заработная плата при пятидневной рабочей неделе:

$$S_{\mathrm{d}} = \frac{S_{\mathrm{M}} \cdot M}{F_d}, \quad (3)$$

где S_{M} – месячная заработная плата сотрудника, RUB;

F_d – число рабочих дней в месяце,

M – число месяцев, в течение которых сотрудник работал без отпусков.

Полная заработная плата может быть рассчитана как:

$$S_F = S_b + S_{add}, \quad (4)$$

В соответствии с приказом «Об установлении размера должностных окладов по отдельным профессиональным квалификационным группам»,

доцент кафедры, кандидат физико-математических наук получает в ТПУ оклад в размере 37700 руб. Размер оклада инженера-исследователя, не имеющего научной степени, составляет 23800 руб. Рассчитаем суммарную заработную плату для исполнителей рассматриваемого проекта:

Месячная ЗП:

– руководитель

$$S_b = S_r \cdot (1 + k_{\text{pr}} + k_d) \cdot k_r = 37700 \cdot (1 + 0{,}3 + 0{,}2) \cdot 1{,}3 = 73515 \text{ руб}$$

$$S_F = S_b + S_{add} = 73515 + 0{,}15 \cdot 73515 = 84542 \text{ руб}$$

– исполнитель

$$S_b = S_r \cdot \left(1 + k_{\text{pr}} + k_d\right) \cdot k_r = 23800 \cdot (1 + 0{,}3 + 0{,}25) \cdot 1{,}3 = 46410 \text{ руб}$$

$$S_F = S_b + S_{add} = 46410 + 0{,}15 \cdot 46410 = 53372 \text{ руб}$$

Средняя дневная ЗП:

$$S_{\text{D.sup.}} = \frac{S_{\text{b.sup.}}}{F_d} = \frac{73515}{20{,}58} = 3572 \text{ руб}$$

$$S_{\text{D.ex.}} = \frac{S_{\text{b.ex.}}}{F_d} = \frac{46410}{20{,}58} = 2\,254 \text{ руб}$$

где среднее количество рабочих дней месяца определяется как:

$$F_d = \frac{\text{T}_w}{12} = \frac{247}{12} = 20{,}58.$$

Учитывая что руководитель проработал 22 дня, а исполнитель 87 дней, рассчитаем затраты на основную заработную плату исполнителей за время выполнения проекта:

$$S_{sup} = S_{\text{D.sup.}} \cdot t_{\text{sup}} = 3572 \cdot 22 = 78575 \text{ руб},$$

$$S_{ex} = S_{\text{D.ex.}} \cdot t_{\text{ex}} = 1762{,}4 \cdot 82 = 196162 \text{ руб}$$

Дополнительные выплаты исполнителям проекта:

$$S_{add.sup.} = 0{,}15 \cdot 73515 = 11027 \text{ руб},$$

$$S_{add.ex.} = 0{,}15 \cdot 46410 = 6962 \text{ руб}$$

Дневная дополнительная заработная плата:

$$S_{D.add.sup.} = \frac{11027}{20{,}58} = 536 \text{ руб},$$

$$S_{D.add.ex.} = \frac{6962}{20{,}58} = 338 \text{ руб}$$

Дополнительная заработная плата за всё время проекта:

$$S_{\text{add.sup.}} = S_{D.add.sup.} \cdot t_{sup} = 536 \cdot 22 = 11786 \text{ руб},$$

$$S_{\text{add.ex.}} = S_{D.add.eng.} \cdot t_{\text{eng}} = 338 \cdot 87 = 29424 \text{ руб}$$

Полная заработная плата исполнителей:

$$S_{F.sup.} = S_b + S_{add} = 78575 + 11786 = 90361 \text{ руб},$$

$$S_{F.ex.} = S_b + S_{add} = 196162 + 6962 = 225586 \text{ руб}.$$

2.4.2.3 Социальные выплаты

Страховые взносы на обязательное пенсионное, медицинское и социальное страхование начисляются на заработную плату сотрудников и уплачиваются из средств работодателя. Размер страховых выплат может быть рассчитан по формуле:

$$S_{\text{exb}} = k_{\text{exb}}(S_{\text{b}} + S_{add}), \quad (5)$$

где k_{exb}– коэффициент выплат во внебюджетные фонды.

Отчисления на социальные нужды составляют:

– 22 % пенсионное обеспечение застрахованных работников;

– 5,1 % на оказание медицинской помощи, профилактических мер охраны здоровья;

– 2,9 % выплаты, связанные со временной нетрудоспособностью штатных сотрудников;

Итого, суммарные страховые выплаты составляют 30 %.

$$S_{\text{exb}} = 0{,}3 \cdot (78575 + 11786 + 196162 + 29424) = 94784 \text{ руб}$$

2.4.2.4 Накладные расходы

При выполнении проекта могут возникнуть косвенные издержки – накладные расходы, возникающие дополнительно к основным затратам, например, на консультационные услуги, оплату коммунальных услуг, расход на услуги связи (телефон, Интернет).

Расчет накладных расходов осуществляется по формуле:

$$S_{\text{over}} = k_{\text{over}}(S_{\text{b}} + S_{add}), \quad (6)$$

где k_{over} – коэффициент накладных расходов, примем его значение равным 15 %.

$$S_{\text{over}} = 0{,}15 \cdot (78575 + 11786 + 196162 + 29424) = 47392 \text{ руб}.$$

2.4.2.5 Формирование бюджета затрат на исследование

После выполнения всех расчетов можно определить плановую себестоимость проекта. В таблице 17 представлен бюджет затрат:

Таблица 17 – Бюджет затрат

Наименование	Стоимость, руб.	Стоимость, %
Стоимость специализированного оборудования	3981	0,86
Заработная плата руководителя	90361	19,55
Заработная плата исполнителя	225586	48,82
Выплаты в социальные фонды	94784	20,51
Накладные расходы	47392	10,26
Итого:	462104	100

2.4.3 Риски разработки

Оценка рисков является важным элементом планирования проекта, так как профилактические меры относительно обнаруженных угроз позволяют избежать непредвиденных потерь во времени и финансах. В таблице 18 описаны возможные риски рассматриваемого проекта, а также рекомендации по их профилактике и/или устранению.

Таблица 18 – Потенциальные риски проекта

№	Риск	Потенциальное воздействие	Вероятность наступления	Влияние риска	Уровень риска	Способы смягчения риска	Условия наступления
1	Изменение параметров пучка/установки	Изменения откликов детекторов могут привести к ошибочной кластеризации	1	4	средний	Изучение истории технологических изменений эксперимента для оценки степени влияния	Недостаточный уровень ознакомления с предметной областью
2	Недостаточная оценка разработанного алгоритма командой эксперимента	Отказ потребителей поддерживать дальнейшее развитие исследование	2	2	низкий	Качественная презентация с демонстрацией результатов	Низкий уровень подготовки к демонстрациям промежуточных достижений
3	Изменение условий/возможности сотрудничества с потребителями	Невозможность завершить исследование	3	3	средний	Изучение правовых особенностей научного международного взаимодействия, установление и поддержание социальных контактов	Изменение геополитической ситуации
4	Недостаточная временная эффективность	Неконкуренто-способность алгоритма	1	1	низкий	Предварительный анализ технического решения для установления способов оптимизации	Недостаточный уровень знаний и навыков исполнителей проекта

Были определены риски проекта и возможные меры профилактики и минимизации рисков. Рассматриваемое исследование достаточно устойчиво к возможным угрозам, отсутствуют критические риски, которые невозможно было бы предусмотреть и смягчить.

2.5 Разработка экономической модели

2.5.1 Экономическая эффективность

Экономическая эффективность – это способность использовать ресурсы компании или организации таким образом, чтобы получить наилучший результат при минимальных затратах.

В реализации проекта одним из ключевых моментов является выбор алгоритма кластеризации. От правильности выбора зависят как качество кластеризации, так и временная эффективность. Произведем сравнение выбранного алгоритма с другими возможными решениями.

Интегральный финансовый показатель разработки определяется как

$$I_{fin}^{p} = \frac{F_{pi}}{Fmax}, \quad (7)$$

где F_{pi} – стоимость i-го варианта исполнения, F_{max} – максимальная стоимость проекта.

Интегральный показатель ресурсоэффективности решения определяется как:

$$I_m^a = \sum_{i=1}^{n} a_i b_i^a, \quad I_m^p = \sum_{i=1}^{n} a_i b_i^p, \quad (8)$$

a_i – весовой коэффициент для i-го параметра,

b_i^a, b_i^p – значения i-го критерия для разработки и аналогов, значения которых выбираются на основании экспертной оценки,

n – число критериев сравнения.

Вычисление интегрального показателя отражено в таблице 19:

Таблица 19 – сравнительная оценка характеристик вариантов исполнения проекта

Критерий	Весовой коэффициент	Выбраный алгоритм проекта (DBSCAN)	Альтернативный алгоритм 1 (K-Means)	Альтернативный алгоритм 2 (Affinity Propagation)
Временная эффективность	0.2	5	2	2
Адаптивность алгоритма	0.2	4	2	4
Информативность кластеризации	0.2	5	1	3
Надежность кластеризации	0.4	4	1	4
Итого:	1.0	18	6	13

$$I_{proj} = 5 \cdot 0{,}2 + 4 \cdot 0{,}2 + 5 \cdot 0{,}2 + 4 \cdot 0{,}4 = 4{,}4$$

$$I_{analogue1} = 2 \cdot 0{,}2 + 2 \cdot 0{,}2 + 1 \cdot 0{,}2 + 1 \cdot 0{,}4 = 1{,}4$$

$$I_{analogue2} = 2 \cdot 0{,}2 + 4 \cdot 0{,}2 + 3 \cdot 0{,}2 + 4 \cdot 0{,}4 = 3{,}4$$

Интегральные показатели эффективности рассматриваемого проекта ($\boldsymbol{I}_{\boldsymbol{fin}}{}^{\boldsymbol{p}}$) и аналогов ($\boldsymbol{I}_{\boldsymbol{fin}}{}^{\boldsymbol{a}}$) определяются согласно следующей формуле:

$$I_{fin}^{p} = \frac{I_m^p}{I_f^p}, \; I_{fin}^{a} = \frac{I_m^a}{I_f^a} \quad (9)$$

Сравнение интегральных показателей эффективности проекта и его аналогов определяет сравнительную эффективность проекта:

$$E_{av} = \frac{I_{fin}^{p}}{I_{fin}^{a}} \quad (10)$$

Таблица 20 – сравнительная эффективность проекта

Показатель	Проект	Аналог 1	Аналог 2
Интегральный финансовый показатель	0.97	0,97	0,97
Интегральный показатель ресурсоэффективности	4,4	1,4	3,4
Интегральный показатель эффективности	4,27	1,36	3,30
Сравнительная оценка вариантов исполнения проекта	1	0,32	0,77

Проведённая оценка эффективности, показатели которой отражены в таблице 20, показывает, что выбранный вариант исполнения проекта превосходит рассмотренные аналоги по интегральным показателям финансов, эффективности и ресурсоэффективности.

2.6 Выводы по разделу

В результате оценки коммерческого потенциала НТИ анализа были определены потребители проекта, произведена оценка качества проекта. Слабые и сильные стороны, угрозы проекта и возможности были определены в процессе SWOT-анализа, который показал, что стратегических изменений совершать в проекте не требуется. Оценка готовности проекта к коммерциализации ожидаемо привела к отрицательному заключению. Конкурентоспособность проекта была оценена выше среднего значения по технологии Quad.

На этапе инициации разработки были обозначены требования и ожидания к проекту.

При планировании были определены основные этапы развития проекта и временная нагрузка для каждого из участников. Бюджет проекта был оценен в 462104 рублей. Основной статьёй расходов является заработная плата исполнителей. Также были определены возможные риски проекта и меры по их минимизации.

Оценка экономической эффективности вариантов использования рассматриваемого проекта показала, что выбранный вариант предпочтительнее своих аналогов.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО		
8ПМ1И	Крамойкин Иван Алексеевич		
Школа	Инженерная школа информационных технологий и робототехники	**Отделение (НОЦ)**	ОИТ / Отделение информационных технологий
Уровень образования	магистратура	**Направление/ специальность**	09.04.04 Программная инженерия

Тема ВКР:

Кластеризация отклика калориметров эксперимента NA64 (CERN, SPS)	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* алгоритм кластеризации откликов калориметров *Область применения:* эксперименты Физики Высоких Энергий *Рабочая зона:* офис *Размеры помещения:* 11*11 м. *Количество и наименование оборудования рабочей зоны: ПК*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности <u>при разработке проектного решения:</u> – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	Российская Федерация. Законы. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ ГОСТ 12.2.032-78 Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования. ГОСТ 21889-76 Система "Человек-машина". Кресло человека-оператора. Общие эргономические требования.
2. Производственная безопасность <u>при разработке проектного решения</u> – Анализ выявленных вредных и опасных производственных факторов – Расчет уровня опасного или вредного производственного фактора	**Опасные факторы:** – Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает работающий **Вредные факторы**: – факторы, обладающие свойствами психофизиологического воздействия на организм человека; – факторы, связанные с отсутствием или недостатком необходимого искусственного освещения; – факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении рабочего. **Расчет**: расчёт системы искусственного освещения
3. Экологическая безопасность <u>при разработке проектного решения</u>	**Воздействие на селитебную зону**: отсутствует **Воздействие на литосферу**: утилизация

	оргтехники **Воздействие на гидросферу**: отсутствует **Воздействие на атмосферу**: отсутствует
4. Безопасность в чрезвычайных ситуациях <u>при разработке проектного решения</u>	**Возможные ЧС**: пожар **Наиболее типичная ЧС**: пожар
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Антоневич Ольга Алексеевна	к.б.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		

3 Социальная ответственность

Объектом исследования является алгоритм кластеризации откликов калориметров эксперимента NA64 CERN.

Разработка алгоритма, написание кода и тестирование выполняются с использованием ПК в условиях учебной аудитории 11-го корпуса Томского Политехнического университета.

Рабочие процессы, связанные с объектом исследования заключаются во взаимодействии с ПК и коммуникации со специалистами.

3.1 Правовые и организационные вопросы обеспечения безопасности

3.1.1 Социальные правовые нормы трудового законодательства

Продолжительность рабочего в режиме гибкого рабочего времени дня определялась по соглашению между работником и работодателем в соответствии с главой 16 ст. 102 ТК РФ [18].

Общие требования при обработке персональных данных работника, гарантии их защиты, хранение, использование и передача персональных данных работника, права работников в целях обеспечения защиты персональных данных, ответственность за нарушение норм, регулирующих обработку и защиту персональных данных работника, определены в соответствии со ст. 86, ст. 87, ст. 88, ст.89 и ст. 90 главы 14 ТК РФ [18]. Отношения, связанные с обработкой персональных данных регулируются ФЗ от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О персональных данных» [19].

3.1.2 Организационные мероприятия при компоновке рабочей зоны исследователя

Общие эргономические требования к рабочим местам при выполнении работ в положении сидя определяются в соответствии с ГОСТ 12.2.032-78 ССБТ [20].

В соответствии с данным документом конструкцией рабочего места должно быть обеспечено выполнение трудовых операций в пределах зоны досягаемости моторного поля.

Конструкцией производственного оборудования и рабочего места должно быть обеспечено оптимальное положение работающего, которое достигается регулированием:

– высоты рабочей поверхности, сиденья и пространства для ног.

– высоты сиденья и подставки для ног (при нерегулируемой высоте рабочей поверхности).

Конструкция регулируемого кресла оператора должна соответствовать общим эргономическим требованиям, установленным в ГОСТ 21889—76 [21]: обеспечивать человеку-оператору соответствующую характеру и условиям труда физиологически рациональную рабочую позу и длительное поддержание основной рабочей позы в процессе трудовой деятельности.

В соответствие с ГОСТ 22269—76 [21] органы управления должны располагаться в зоне досягаемости моторного поля и расположение органов управления должно обеспечивать равномерность нагрузки обеих рук и ног человека-оператора.

3.2 Производственная безопасность

В этом разделе будут рассмотрены основные аспекты производственной безопасности при выполнении исследования, а также при эксплуатации проектируемого решения с точки зрения возникающих опасных и вредных факторов.

В процессе разработки программного решения можно выделить три этапа: проектирование, разработка и эксплуатация. Опасные и вредные факторы, сопутствующие выделенным этапам, определены в соответствии с ГОСТ 12.0.003-2015 [22] и представлены в таблице 1:

Таблица 2 — Возможные вредные и опасные факторы

Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
1. Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает рабочий	ГОСТ 12.1.038-82 Система стандартов безопасности труда. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов ГОСТ 12.1.019-2017 Система стандартов безопасности труда. Электробезопасность. Общие требования и номенклатура видов защиты
2. Производственные факторы, обладающие свойствами психофизиологического воздействия на организм человека (активное наблюдение за ходом производственного процесса, монотонность труда, перенапряжение анализаторов)	МР 2.2.9.2311 – 07 «Профилактика стрессового состояния работников при различных видах профессиональной деятельности» ГОСТ 12.2.032-78 Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования.
3. Производственные факторы, связанные с отсутствием или недостатком необходимого искусственного освещения	СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 2305-95* СанПиН 1.2.3685-21. Гигиенические требования к естественному, искусственному и совмещённому освещению жилых и общественных зданий
4. Опасные и вредные производственные факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении рабочего	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания

3.2.1 Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает рабочий

Электрический ток может оказывать опасное и вредное воздействие на человека в зависимости от его силы и продолжительности воздействия. Маленькая сила тока, проходящего через тело, может вызывать дискомфорт и судороги мышц. Сильный ток может привести к остановке сердца, ожогам, повреждению нервной системы и даже смерти.

В рамках текущей работы не производились контакты с открытыми источниками электрического тока. Ток, протекающий в компьютерной периферии (компьютерная мышь, клавиатура), не представляет существенной опасности для здоровья человека, так как не превышает предельных значений, определенных в ГОСТ 12.1.038-82 [23]. Требования безопасности при эксплуатации электрооборудования регламентируются следующими нормативными актами: Приказ Минтруда России от 15.12.2020 № 903н (ред. от 29.04.2022) «Об утверждении Правил по охране труда при эксплуатации электроустановок» [24] и Приказ Министерства энергетики РФ от 12 августа 2022 г. № 811 “Об утверждении Правил технической эксплуатации электроустановок потребителей электрической энергии” [25].

3.2.2 Производственные факторы, обладающие свойствами психофизиологического воздействия на организм человека

Продолжительная работа по разработке ПО, может вызывать значительное напряжение функций зрительного анализатора по причине регулярного взаимодействия с ПК. Помимо этого, повышенная концентрация и обработка больших объёмов информации может приводить к умственному и нервному истощению, а статичное положение тела, занимаемое при такого рода работе, приводит к утомлению органов опорно-двигательной системы.

В соответствии с МР 2.2.9.2311-07. 2.2.9. [26] разработка ПО относится к группе В (творческая работа в режиме диалога с ПЭВМ). В зависимости от уровня нагрузки за рабочую смену при работе с ПК устанавливается суммарное время регламентированных перерывов (таблица 2):

Таблица 2 — Суммарное время регламентированных перерывов в зависимости от продолжительности работы и категории трудовой деятельности с ПЭВМ

Категория работы ПЭВМ	Уровень нагрузки за смену при работе с ПЭВМ, час	Суммарное время перерывов, мин.	
		при 8-часовой рабочей смене	при 12-часовой рабочей смене
1	До 2	50	80
2	До 4	70	110
3	До 6	90	140

Во время регламентированных перерывов рекомендуется психологическая разгрузка в специально оборудованных комнатах, а также выполнение комплексов физических упражнений.

3.2.3 Производственные факторы, связанные с отсутствием или недостатком искусственного освещения.

Недостаточная освещенность рабочей зоны также считается одним из факторов, влияющих на работоспособность человека. Для промышленных предприятий оптимальная освещенность территории и помещений является важной и непростой технической задачей, решение которой обеспечивает нормальные гигиенические условия для работающего персонала. Правильно подобранные источники света и их проектирование создают условия для производственного труда, корректности выполнения технологических операций, соблюдения правил и техники безопасности.

Главной задачей светотехнических расчётов для искусственного освещения является определение требуемой мощности электрической осветительной установки для создания заданной освещённости.

При учете особенностей процесса работ на компьютере допускается применение системы общего равномерно освещения.

На поверхности рабочего стола в зоне размещения документов освещенность должна составлять 300 (при системе общего освещения) – 500 (при системе комбинированного освещения) люксов, а на поверхности экрана монитора – не превышать 300 лк.

Для общего освещения применяются газоразрядные лампы: дневной (ЛД), холодно-белой (ЛХБ), тепло-белой (ЛТБ) и белой цветности (ЛБ). Определим необходимое количество источников света для полного освещения аудиторного помещения с рабочим компьютером люминесцентными потолочными светильниками.

Световой поток для люминесцентных ламп, мощностью 56 Вт:

$$F = Ra \cdot P, \quad (1)$$

где Ra = 80Лм/Вт – минимальный индекс цветопередачи для люминесцентной лампы.

$$F = 80 \cdot 56 = 4480 \text{ Лм}.$$

Необходимое количество ламп для освещения лабораторной аудитории:

$$N = \frac{E \cdot S \cdot z \cdot k}{K \cdot F \cdot n}, \quad (2)$$

где E – освещенность, Лк (при системе общего освещения E = 300 Лк); K – переходный коэффициент, 4,5; n – коэффициент использования светового потока осветительной установки, 45 %; k – коэффициент запаса, 4,5; S – площадь освещаемого помещения, 121 м^2; z – поправочный коэффициент, учитывающий неравномерность освещения, 0.9.

$$N = \frac{300 \cdot 121 \cdot 0{,}9 \cdot 4{,}5}{4{,}5 \cdot 4480 \cdot 0{,}45} = 16{,}2 \text{ шт}.$$

Рассчитанное значение количества светильников округляем в большую сторону до целого числа. Получаем, что для надлежащего освещения аудитории необходимо 17 светильников.

Для защиты от недостаточной освещенности рабочей зоны естественное освещение по своему спектру является наиболее приемлемым, но не всегда его оказывается достаточно. Это связано во многом с режимом

работы. Обычно рекомендуется применять общее и комбинированное освещение. Нормы освещенности рабочего места соответствуют [27].

3.2.4 Производственные факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении рабочего.

Основными факторами, характеризующими микроклимат производственной среды, являются: температура, подвижность и влажность воздуха. Отклонение этих параметров от нормы приводит к ухудшению самочувствия работника, снижению производительности его труда и к возникновению различных заболеваний.

Работа в условиях высокой температуры сопровождается интенсивным потоотделением, что приводит к обезвоживанию организма, потере минеральных солей и водорастворимых витаминов, серьезным изменениям в деятельности сердечно-сосудистой системы, увеличению частоты дыхания, а также оказывает влияние на функционирование других органов и систем (ослабление внимания, ухудшение координации движений, замедление реакции тела и т.д.).

Высокая относительная влажность при высокой температуре воздуха способствует перегреву организма, при низкой же температуре увеличивается теплоотдача с поверхности кожи, что ведет к переохлаждению организма. Низкая влажность вызывает неприятные ощущения в виде сухости слизистых оболочек дыхательных путей работающего.

При нормировании метеорологических условий в производственных помещениях учитывают время года, физическую тяжесть выполняемых работ, а также количество избыточного тепла в помещении. Оптимальные и допустимые метеорологические условия температуры и влажности устанавливаются согласно СанПиН 1.2.3685-21 СанПиН 2.2.4.548-96 [28] и приведены в таблице 3.

Для удобства работы в помещении необходимо нормирование параметров микроклимата, то есть необходимо проведение мероприятий по контролю способов и средств защиты от высоких и низких температур, системы отопления, вентиляции и кондиционировании воздуха, искусственное освещение и т.п.

Таблица 3 – Оптимальные показатели микроклимата на рабочих местах

Период года	Категория работ по уровню энергозатрат, Вт	Температура воздуха, ℃	Температура поверхностей, ℃	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	Ia (до 139)	22-24	21-25	60-40	Не более 0,1
Тёплый	Ia (до 139)	23-25	22-26	60-40	Не более 0,1

Для поддержания данных санитарных норм достаточно иметь естественную неорганизованную вентиляцию помещения и местный кондиционер установки полного кондиционирования воздуха, обеспечивающий постоянство температуры, относительной влажности, скорости движения и чистоты воздуха. Необходима система центрального отопления, обеспечивающая заданный уровень температуры в зимний период по СНиП 41-01-2003 [29]. В зимний период в аудитории для поддержания необходимой температуры используется система водяного отопления. Эта система надежна в эксплуатации и обеспечивает возможность регулирования температуры в широких пределах. При устройстве системы вентиляции и кондиционирования воздуха в помещении лаборатории необходимо соблюдать определенные требования пожарной безопасности. В зимнее время в помещении необходимо предусмотреть систему отопления. Она должна обеспечивать достаточное, постоянное и равномерное нагревание воздуха. В помещениях с повышенными требованиями к чистоте воздуха должно использоваться водяное отопление.

3.3 Экологическая безопасность

Целью данного подраздела является выявление потенциальных опасных факторов влияния объекта исследования на окружающую среду, а также разработка мер, которая обеспечивает безопасность исследовательской деятельности для окружающей среды.

3.3.1 Анализ влияния процесса исследования на окружающую среду

При разработке ПО используются компьютеры, мониторы и прочая техника, ненадлежащая утилизация которой может оказывать негативное влияние на литосферу.

В соответствии с ГОСТ Р 53692-2009 «Ресурсосбережение. Обращение с отходами» [30], регламентирующем обращение с отходами, ПК и подобная оргтехника являются отходами 4го класса опасности (малоопасные отходы). Административный Кодекс РФ [31] запрещает выбрасывать технику наряду с обыкновенным мусором. Аккумуляторы, системы охлаждения и прочие детали оргтехники содержат такие вредные вещества как ртуть, свинец и мышьяк – 2й класс опасности согласно ст. 1 Федерального закона «Об отходах производства и потребления» [32]. Утилизацией таких отходов занимаются специализированные организации.

Для снижения ущерба окружающей среде предлагается использовать процедуру утилизации, при которой часть отходов (до 90 %) будет переработано для вторичного использования и только около 10 % будут непосредственно утилизированы.

Помимо контроля за утилизацией следует принимать меры по снижению количества отходов, а также приобретать технику, оказывающую наименьшее вредное влияние на окружающую среду как в эксплуатации, так и при утилизации.

3.4 Безопасность в чрезвычайных ситуациях

3.4.1 Анализ вероятных ЧС, которые могут возникнуть в лаборатории при проведении исследований

Пожар. В соответствии с СП 12.13130.2009. [33] аудиторное помещение, в котором проводилось исследование, отнесено к классу В по пожароопасности (содержатся твёрдые горючие вещества в холодном состоянии).

Горючие материалы: мебель, строительные материалы.

Источники возгорания: электрооборудование, проводка.

3.4.2 Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

В целях снижения риска возникновения пожара и минимизации возможного ущерба производятся профилактические мероприятия, которые подразделяются на организационно-технические, эксплуатационные и режимные. Организационно-технические мероприятия заключаются в проведении регулярных инструктажей сотрудников ответственным за пожарную безопасность, обучении сотрудников надлежащей эксплуатации оборудования и необходимым действиям в случае возникновения пожара, паспортизацию веществ, материалов и изделий в части обеспечения пожарной безопасности, изготовление и применение средств наглядной агитации по обеспечению пожарной безопасности [34]. К эксплуатационным мероприятиям относят профилактические осмотры оборудования. Мероприятия режимного характера включают установление правил организации работ и соблюдение противопожарных мер. Для предупреждения возникновения пожара необходимо соблюдение следующих правил пожарной безопасности:

– содержание помещений в соответствии с требованиями пожарной безопасности;

– надлежащая эксплуатация оборудования (правильное включение оборудования в сеть электропитания, контроль нагрева оборудования);

– обучение производственного персонала правилам пожарной безопасности;

– наличие, правильное размещение и использование средств пожаротушения.

В помещении с электрооборудованием, во избежание поражения электрическим током, целесообразно использовать углекислотные или порошковые огнетушители. Данные огнетушители предназначены для тушения загораний различных веществ и материалов, электроустановок под напряжением до 1000 В, горючих жидкостей. Химические и пенные огнетушители не допустимы. Огнетушители следует располагать на защищаемом объекте в соответствии с требованиями таким образом, чтобы они были защищены от воздействия прямых солнечных лучей, тепловых потоков, механических воздействий и других неблагоприятных факторов (вибрация, агрессивная среда, повышенная влажность и т. д.). Они должны быть хорошо видны и легкодоступны в случае пожара. Предпочтительно размещать огнетушители вблизи мест наиболее вероятного возникновения пожара, вдоль путей прохода, а также около выхода из помещения. Огнетушители не должны препятствовать эвакуации людей во время пожара. Согласно требованиям пожарной безопасности ГОСТ 12.1.004-91 [34], на этаже находится 2 огнетушителя ОП3 (огнетушители переносные порошковые), лестничные пролеты оборудованы гидрантами, имеется кнопка пожарной сигнализации.

3.5 Выводы по разделу

В этом разделе были рассмотрены различные вредные и опасные производственные факторы, оказывающие влияние на рабочего, находящегося на рабочем месте. Значения рассмотренных факторов не превышают показателей нормы, установленных в соответствующих нормативных документах.

Согласно Правилам Устройства Электроустановок рассматриваемое помещение относится к категории помещений без повышенной опасности [35].

Персонал, участвующий в разработке программного решения в соответствии с Правилами по охране труда при эксплуатации электроустановок относится к первой группе по электробезопасности – неэлектротехнический персонал [36]. Присвоение группы I по электробезопасности производится путем проведения инструктажа, который необходимости) проверкой приобретенных навыков безопасных способов работы и оказания первой помощи при поражении электрическим током.

Разработка ПО относится к Iб категории по уровню энергозатрат организма (140-174 Вт) согласно СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" [37].

Помещение, в котором проводилось исследование, отнесено к классу В по пожароопасности (содержатся твёрдые горючие вещества в холодном состоянии) в соответствии с СП 12.13130.2009 «Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности» [38].

Согласно постановлению Правительства Российской Федерации от 31 декабря 2020 года, N2398 «Критерии отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий» [39] объект (алгоритм кластеризации), оказывающий незначительное негативное воздействие на окружающую среду относится ко III категории.

Заключение

В результате выполнения дипломной работы был получен и сформирован исходный датасет для кластеризации откликов калориметров эксперимента NA64 (CERN/SPS).

Был спроектирован и разработан конвейер обработки данных на языке Python. В нём в виде конфигурируемого набора атомарных классов-обработчиков реализованы функции фильтрации сигналов, аппроксимации аналитической функцией, расчета статистических показателей и визуализации.

Была осуществлена кластеризация откликов с использованием алгоритма DBSCAN , реализована методика определения гиперпараметров для алгоритма.

Результаты кластеризации указывают на достаточно высокое качество, достигнутое при разделении откликов адронных и электронных пучков (значение F1-критерия равно 0,904). Алгоритм способен разделять отклики в области высоких значений энерговыделения, но точность классификации значительно снижается в области низких значений энерговыделения. Для достижения более высокой точности разделения откликов необходима более тщательная подготовка данных, возможно с разделением откликов отдельных ячеек калориметров SRD и дальнейшее исследование влияния гиперпараметров на качество кластеризации.

Список литературы

1. Niedermayer G. Investigations of calorimeter clustering in ATLAS using machine learning : дис. – 2017.

2. Valsecchi D. et al. Deep learning techniques for energy clustering in the CMS ECAL //Journal of Physics: Conference Series. – IOP Publishing, 2023. – Т. 2438. – №. 1. – С. 012077.] и [Whiteson S., Whiteson D. Machine learning for event selection in high energy physics //Engineering Applications of Artificial Intelligence. – 2009. – Т. 22. – №. 8. – С. 1203-1217.

3. Arthur D., Vassilvitskii S. k-means++: The advantages of careful seeding. – Stanford, 2006.

4. Sculley D. Web-scale k-means clustering //Proceedings of the 19th international conference on World wide web. – 2010. – С. 1177-1178.

5. Schubert E. et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN //ACM Transactions on Database Systems (TODS). – 2017. – Т. 42. – №. 3. – С. 1-21.

6. Rahmah N., Sitanggang I. S. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra //IOP conference series: earth and environmental science. – IoP Publishing, 2016. – Т. 31. – №. 1. – С. 012012.

7. Ankerst M. et al. OPTICS: Ordering points to identify the clustering structure //ACM Sigmod record. – 1999. – Т. 28. – №. 2. – С. 49-60.

8. Frey B. J., Dueck D. Clustering by passing messages between data points //science. – 2007. – Т. 315. – №. 5814. – С. 972-976.

9. Comaniciu D., Meer P. Mean shift: A robust approach toward feature space analysis //IEEE Transactions on pattern analysis and machine intelligence. – 2002. – Т. 24. – №. 5. – С. 603-619.

10. Nielsen F., Nielsen F. Hierarchical clustering //Introduction to HPC with MPI for Data Science. – 2016. – С. 195-211.

11. Ng A., Jordan M., Weiss Y. On spectral clustering: Analysis and an algorithm //Advances in neural information processing systems. – 2001. – Т. 14.

12. Zhang T., Ramakrishnan R., Livny M. BIRCH: an efficient data clustering method for very large databases //ACM sigmod record. – 1996. – Т. 25. – №. 2. – С. 103-114.

13. Yang M. S., Lai C. Y., Lin C. Y. A robust EM clustering algorithm for Gaussian mixture models //Pattern Recognition. – 2012. – Т. 45. – №. 11. – С. 3950-3961.

14. Gaiser J. E. Charmonium spectroscopy from radiative decays of the J/psi and psi'. – Stanford University, 1983.

15. Schröder S. Commissioning of a prototype hadronic calorimeter : дис. – Master Thesis, 2015.

16. Trebuňa P. et al. The importance of normalization and standardization in the process of clustering //2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI). – IEEE, 2014. – С. 381-385.

17. Sander J. et al. Density-based clustering in spatial databases: The algorithm gdbscan and its applications //Data mining and knowledge discovery. – 1998. – Т. 2. – С. 169-194.

18. Российская Федерация. Законы. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 25.02.2022) (с изм. и доп., вступ. в силу с 01.03.2022) — URL: http://www.consultant.ru/document/cons_doc_law_34683/ (дата обращения 24.03.2022)

19. Российская Федерация. Законы. Федеральный Закон от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О персональных данных». Режим доступа: https://normativ.kontur.ru/document?moduleId=1&documentId=447363 (дата обращения 31.05.2023)

20. ГОСТ 12.2.032-78 Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования. Режим доступа: https://allgosts.ru/13/180/gost_12.2.032-78 (дата обращения: 17.05.2023)

21. ГОСТ 21889-76 Система "Человек-машина". Кресло человека-оператора. Общие эргономические требования. Режим доступа: https://allgosts.ru/13/180/gost_12.2.032-78 (дата обращения: 17.05.2023)

22. ГОСТ 12.0.003-2015 Система стандартов безопасности труда. Опасные и вредные производственные факторы. Классификация Режим доступа: https://allgosts.ru/13/100/gost_12.0.003-2015 (дата обращения: 17.05.2023)

23. ГОСТ 12.1.038-82 Система стандартов безопасности труда. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов. Режим доступа: https://allgosts.ru/13/260/gost_12.1.038-82 (дата обращения: 18.05.2023)

24. Приказ Минтруда России от 15.12.2020 № 903н (ред. от 29.04.2022) «Об утверждении Правил по охране труда при эксплуатации электроустановок» (Зарегистрировано в Минюсте России 30.12.2022 № 61957). Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_372952/b3ff40ceea8ae665280131c2b50f9892cb958415/ (дата обращения: 18.05.2023)

25. Приказ Министерства энергетики РФ от 12 августа 2022 г. № 811 “Об утверждении Правил технической эксплуатации электроустановок потребителей электрической энергии” Режим доступа: https://www.garant.ru/products/ipo/prime/doc/405299745/ (дата обращения: 18.05.2023)].

26. МР 2.2.9.2311-07. 2.2.9. Состояние здоровья работающих в связи с состоянием производственной среды. Профилактика стрессового состояния работников при различных видах профессиональной деятельности. Методические рекомендации" Режим доступа: https://legalacts.ru/doc/mr-2292311-07-229-sostojanie-zdorovja-rabotaiushchikh-v/ (дата обращения: 28.05.23)

27. СНиП 23-05-95* Естественное и искусственное освещение (с Изменением № 1) Режим доступа: http://docs.cntd.ru/document/871001026 (дата обращения: 28.05.23)

28. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания. Режим доступа: http://docs.cntd.ru/document/901704046 (дата обращения: 29.05.23)

29. СНиП 41-01-2003 Отопление, вентиляция и кондиционирование Режим доступа: http://docs.cntd.ru/document/1200035579 (дата обращения: 29.05.23)

30. ГОСТ Р 53692-2009 «Ресурсосбережение. Обращение с отходами» Режим доступа: https://allgosts.ru/13/030/gost_r_53692-2009 (дата обращения: 29.05.23)

31. Кодекс Российской Федерации об административных правонарушениях" от 30.12.2001 N 195-ФЗ (ред. от 28.04.2023, с изм. от 17.05.2023) Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_34661/ (дата обращения: 29.05.23)

32. Российская Федерация. Законы. Федеральный Закон от 24.06.1998 N 89-ФЗ (ред. от 30.12.2008) «Об отходах производства и потребления». Режим доступа: http://pravo.gov.ru/proxy/ips/?docbody=&nd=102053807 (дата обращения 01.06.2023)

33. СП 12.13130.2009. Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности (в ред. изм. № 1, утв. приказом МЧС России от 09.12.2010 № 643) [Электронный ресурс]. Доступ из сборника НСИС ПБ.–2011.–№2 (45).

34. ГОСТ 12.1.004-91 Система стандартов безопасности труда. Пожарная безопасность. Общие требования Режим доступа: https://docs.cntd.ru/document/9051953 (дата обращения: 31.05.2023)

35. Правила устройства электроустановок (ПУЭ)7-ое издание(утв. приказом Минэнерго РФ от 8 июля 2002 г. N 204) Режим доступа: https://akak7.ru/docs/wp-content/uploads/2019/12/pue.pdf (дата обращения: 31.05.2023)

36. Правила по охране труда при эксплуатации электроустановок (с изменениями на 29 апреля 2022 года) Режим доступа: https://docs.cntd.ru/document/573264184 (дата обращения: 31.05.2023)

37. СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" (с изменениями на 30 декабря 2022 года) Режим доступа: https://docs.cntd.ru/document/573500115 (дата обращения: 31.05.2023)

38. СП 12.13130.2009 «Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности» Режим доступа: https://docs.cntd.ru/document/1200071156 (дата обращения: 31.05.2023)

39. Постановление Правительства Российской Федерации от 31 декабря 2020 года N 2398 «Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий». Режим доступа: https://docs.cntd.ru/document/573292854 (дата обращения 01.06.2023)

Приложение А

Раздел 4

NA64 (CERN, SPS) experiment calorimeter response clusterization.

Обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Крамойкин Иван Алексеевич		

Консультант-лингвист отделения (НОЦ) школы: ИШИТР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИЯ ШБИП,	Уткина А. Н	к.филос.н.		

Introduction

NA64 is a fixed-target experiment on proton supersynchrotron (SPS) in the European Center of Nuclear Researches (CERN). There are many special equipment related to experiment – about ten different detectors and various hardware.

The special parts of the experimental setup are calorimeters (detectors preordained for the energy measurement). The working principle is usually based on the proportionality of the number of photons, born in the material to the deposited energy. Photons reach the Photoelectronic Multipliers which produce the electrical signal.

The detector response depends on the electromagnetic shower specifics and on the initial beam properties. The selection of events related to special scenarios is the crucial interest of the experimental results analysis.

The present work proposes a method for clustering calorimeter responses, represented as a set of parameters of a function that approximates the detector signal. The relevance of the work lies in the possibility of improving the reconstruction of physical events in the NA64 experiment, in which the separation of signals from hadronic and electron beams is currently performed by a simple threshold selection of signals. The aim of the work is to implement a clustering algorithm to determine groups of calorimetric responses corresponding to the hadronic and electron nature of the initiating beam.

Following tasks were established for the aim achievement:

– To define a work structure in the framework of scientific investigation;

– To overview the relevant literature for the most appropriate clustering algorithm choice;

– To obtain the input data;

– To design and to implement the data pipeline;

– To implement clusterization on the preprocessed data;

– To analyze clusters obtained.

1. Clustering algorithms

1.1. Affinity propagation

Affinity Propagation [1] involves finding a set of exemplars that best summarize the data. Method simultaneously considers all data points as potential exemplars. By viewing each data point as a node in a network, method recursively transmits real-valued messages along edges of the network until a good set of exemplars and corresponding clusters emerges. As described later, messages are updated on the basis of simple formulas that search for minima of an appropriately chosen energy function. At any point in time, the magnitude of each message reflects the current affinity that one data point has for choosing another data point as its exemplar.

Affinity Propagation can be interesting as it chooses the number of clusters based on the data provided.

The main drawback of Affinity Propagation is its complexity. The algorithm has a time complexity of the $O(N^{2T})$ order, where N is the number of samples and T is the number of iterations until convergence. Further, the memory complexity is of the order $O(N^2)$ if a dense similarity matrix is used, but reducible if a sparse similarity matrix is used. This makes Affinity Propagation most appropriate for small to medium sized datasets.

1.2 Agglomerative clustering

Agglomerative clustering [2] involves merging examples starting from the point where each observation treated as separate cluster until the desired number of clusters is achieved. Linkage distance is a measurement of observations similarity. It is presented in four variations:

– For *Ward's* linkage, two clusters are merged based on minimum of their error sum of square values,

– For the *Single* linkage the closest minimum distance between two clusters is the reason for merge,

– For the *Complete* linkage, two clusters with the closest maximum distance are merged,

– *Average* linkage method uses the average pair-wise proximity among all pairs of objects in different clusters.

Agglomerative cluster has a “rich get richer” behavior that leads to uneven cluster sizes. In this regard, single linkage is the worst strategy, and Ward gives the most regular sizes. However, the affinity (or distance used in clustering) cannot be varied with Ward, thus for non-Euclidean metrics, average linkage is a good alternative. Single linkage, while not robust to noisy data, can be computed very efficiently and can therefore be useful to provide hierarchical clustering of larger datasets. Single linkage can also perform well on non-globular data.

1.3 BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [3] builds the Clustering Feature Tree (CFT) for the given data in attempt of the memory requirements of large datasets minimization by summarizing the information contained in dense regions as Clustering Feature (CF) entries. Each CF contains the following information: the number of data points in the cluster, the linear and square sums of data points.

This algorithm can be viewed as an instance or data reduction method, since it reduces the input data to a set of subclusters which are obtained directly from the leaves of the CFT. This reduced data can be further processed by feeding it into appropriate algorithm.

1.4 DBSCAN

DBSCAN [4] is short for Density-Based Spatial Clustering of Applications with Noise. It views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape. The central component to the DBSCAN is the concept of *core samples*, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). Any core sample is part of a cluster, by definition. Any sample that is not a core sample, and is at least *eps* in distance from any core sample, is considered an outlier by the algorithm.

You should use DBSCAN where:

– The dataset is moderately large. It can be applied even to really big data in case of optimized and parallel implementation.

– The similarity measurement function is simple and known in advance.

– You supposed to see data clumps of exotic forms, nested and anomaly clusters, low dimensional folds.

– Border density between clumps is less than lowest cluster density. It is better when clusters are separated at all.

– Elements complexity does not matter.

– Number of cluster entities can vary anyhow.

– Number of outliers does not matter in case they are distributed over large volume.

– Number of clusters does not matter

Inability of clusters merging through the holes and opposite, ability to merge pure distinct clusters through high-density bridges are the main algorithm's disadvantages.

1.5 K-Means

The K-Means [5] algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (see below). This algorithm requires the number of clusters to be specified. It scales well to large numbers of samples and has been used across a large range of application areas in many different fields.

The K-Means algorithm divides a set of *N* samples into *K* disjoint clusters *C*, each described by the mean of the samples in the cluster. The means are commonly called the cluster “centroids”, note that they are not, in general, points from *X*, although they live in the same space.

The K-means algorithm aims to choose centroids that minimize the *inertia*, or *within-cluster sum-of-squares* criterion.

Inertia can be recognized as a measure of how internally coherent clusters are. It suffers from various drawbacks:

– Inertia makes the assumption that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes.

– Inertia is not a normalized metric: we just know that lower values are better and zero is optimal. But in very high-dimensional spaces, Euclidean distances tend to become inflated (this is an instance of the so-called “curse of dimensionality”). Running a dimensionality reduction algorithm such as Principal component analysis (PCA) prior to k-means clustering can alleviate this problem and speed up the computations.

Given enough time, K-means will always converge, however this may be to a local minimum. This is highly dependent on the initialization of the centroids. As a result, the computation is often done several times, with different initializations of the centroids.

1.6 Mini-Batch K-Means

The Mini Batch K-Means [6] is a variant of the K-Means algorithm which uses mini-batches to reduce the computation time, while still attempting to optimise the same objective function. Mini-batches are subsets of the input data, randomly sampled for the each training iteration. These mini-batches drastically reduce the amount of computation required to converge to a local solution. In contrast to other algorithms that reduce the convergence time of k-means, mini-batch k-means produces results that are generally only slightly worse than the standard algorithm.

The algorithm iterates between two major steps, similar to vanilla k-means. In the first step, samples are drawn randomly from the dataset, to form a mini-batch. These are then assigned to the nearest centroid. In the second step, the centroids are updated. In contrast to k-means, this is done on a per-sample basis. For each sample in the mini-batch, the assigned centroid is updated by taking the streaming average of the sample and all previous samples assigned to that centroid.

This has the effect of decreasing the rate of change for a centroid over time. These steps are performed until convergence or a predetermined number of iterations is reached.

Mini Batch K-Means converges faster than K-Means, but the quality of the results is reduced.

1.7 Mean Shift

Mean Shift [7] is the non-parametric density based clustering algorithm. The main idea consists in data points shift along the highest density direction within certain radius. Algorithm implements shifting till all points will converge to the local density maximum points.

The algorithm based on the Kernel Density Evaluation (KDE) concept - the method of the base distribution evaluation with the probability density function. It works by placing the kernel in the each data point. The kernel is weighing function using in convolution. There are many different kernel types, but the most popular is the Gaussian kernel. The probability density function is generated by all kernels addition.

Mean Shift algorithm has the following advantages:

– estimates the number of clusters,

– resistant to outliers,

– has only one tuning parameter — the window size h, where h has the physical meaning.

The algorithm is not highly scalable, as it requires multiple nearest neighbor searches during the execution of the algorithm. The algorithm is guaranteed to converge, however the algorithm will stop iterating when the change in centroids is small.

1.8 OPTICS

The OPTICS [8] algorithm shares many similarities with the DBSCAN algorithm, but it is able to deal with different density clusters.

Like DBSCAN the OPTICS needs to parameters - ϵ (the maximum distance for the core points detection) and *MinPts* — the minimal number of points to form a cluster. OPTICS algorithm considers also points from the clusters with higher density. So the *core distance* and *reachibility distance* are assigned for the each point.

Then such definitions are used for the reachability plot creation. Cluster labels can be extracted by evaluating of such plot.

1.9 Spectral clustering

In Spectral Clustering [9], the data points are treated as nodes of a graph. Thus, clustering is treated as a graph partitioning problem. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. An important point to note is that no assumption is made about the shape/form of the clusters.

Spectral clustering involves 3 steps:

– Compute a similarity graph;

– Project the data onto a low-dimensional space;

– Create clusters.

Spectral clustering algorithm advantages:

– Does not make strong assumptions on the statistics of the clusters — Clustering techniques like K-Means Clustering assume that the points assigned to a cluster are spherical about the cluster center. This is a strong assumption to make, and may not always be relevant. In such cases, spectral clustering helps create more accurate clusters.

– Easy to implement and gives good clustering results. It can correctly cluster observations that actually belong to the same cluster but are farther off than observations in other clusters due to dimension reduction.

– Reasonably fast for sparse data sets of several thousand elements.

Spectral clustering algorithm disadvantages:

– Use of K-Means clustering in the final step implies that the clusters are not always the same. They may vary depending on the choice of initial centroids.

– Computationally expensive for large datasets. This is because eigenvalues and eigenvectors need to be computed and then we have to do clustering on these vectors. For large, dense datasets, this may increase time complexity quite a bit.

1.10 Gaussian Mixture

A Gaussian Mixture [10] model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians.

The Gaussian Mixture implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion to assess the number of clusters in the data.

2 Calorimeters response clusterization

2.1 Data description

For the calorimeters responses extraction it was used the *C++* data pipeline, developed in *TPU/NA64* research group. It was prepared the dataset, containing several *.csv* files. Collected data provide the following information: the unique event identifier (*eventID*); the unique detector identifier (*detID*) and detector response, discretized in 32 samples.

2.2 Data pipeline

For implementing the data preprocessing it was designed and developed the *Python* data pipeline. The technological stack used for such implementation is following:

– *Pandas* (data aggregation);

– *Numpy* (mathematical functions and data structure manipulations);

– *Scipy* (peaks detection and approximation);

– *Matplotlib* (data visualization);

– *Plotly* (advanced data visualization).

The pipeline was designed as the set of classes (*handlers*) providing unit procedures for the data processing. The common behavior for the each class is described in *Abstract Handler* abstract class. Other handlers, inherited from the Abstract Handler are implementing such functional, by parent methods override. Such methods can perform for example signals filtering by the number of peaks, signal approximation, data visualization.

2.3 The calorimeter response model

For the signal description it was used the *Crystal Ball* function [11]. This probability density function is commonly used in High Energy Physics for the loss of energy processes description. It consists of Gaussian kernel and polynomial tale.

Such choice for the approximation function is determined by the good correspondence with the detector response, well-interpreted physical parameters and computational speed.

2.4 Parameters clusterization

The *Crystal Ball* function parameters obtained in approximation process were clusterized by DBSCAN algorithm with hyperparameters *eps* = 0.006 and *min_samples* = 6.

Such choice of hyperparameters values can be described in the following way, according to corresponding literature [12], [13]:

min_samples equals to the dimensionality of dataset, multiplied by two;

eps value reflects the elbow position value in the grapth of sorted nearest neighbors mean distances.

2.5 Clustering results

By estimation of the clustering results it can be concluded that hadron and electronic signals are well-distinguished in the region of high values of deposited energy. In the region of low deposited energy values hadron signals are hardly distinguished from electronic signals.

It was calculated the F1 score by the signals classification obtained as result of the clusterization. The value of 0.904 shows that the method described performs well in the task of signals distinguishing.

Conclusion

As the result of work implementation the dataset for the calorimeters responses clusterization was obtained.

The python data processing pipeline was designed and developed in form of unitary handlers classes for the signals filtration, analytical function approximation, statistical scores calculation and data visualization.

The calorimeters responses clusterization was implemented by the DBSCAN algorithm, the methodic for the hyperparameter choice was performed.

The result of clusterization indicates the high quality obtained for the hadron and electron responses separation (the F1 score criterion value equals to 0.904). Algorithm can separate responses in the region of high values of deposited energy. The separation accuracy decreases for the low deposited energy regions. For reaching the greater accuracy data should be preprocessed more carefully.

References

1. Frey B. J., Dueck D. Clustering by passing messages between data points //science. – 2007. – Т. 315. – №. 5814. – С. 972-976.

2. Nielsen F., Nielsen F. Hierarchical clustering //Introduction to HPC with MPI for Data Science. – 2016. – С. 195-211.

3. Zhang T., Ramakrishnan R., Livny M. BIRCH: an efficient data clustering method for very large databases //ACM sigmod record. – 1996. – Т. 25. – №. 2. – С. 103-114.

4. Schubert E. et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN //ACM Transactions on Database Systems (TODS). – 2017. – Т. 42. – №. 3. – С. 1-21.

5. Arthur D., Vassilvitskii S. k-means++: The advantages of careful seeding. – Stanford, 2006.

6. Sculley D. Web-scale k-means clustering //Proceedings of the 19th international conference on World wide web. – 2010. – С. 1177-1178.

7. Comaniciu D., Meer P. Mean shift: A robust approach toward feature space analysis //IEEE Transactions on pattern analysis and machine intelligence. – 2002. – Т. 24. – №. 5. – С. 603-619.

8. Ankerst M. et al. OPTICS: Ordering points to identify the clustering structure //ACM Sigmod record. – 1999. – Т. 28. – №. 2. – С. 49-60.

9. Ng A., Jordan M., Weiss Y. On spectral clustering: Analysis and an algorithm //Advances in neural information processing systems. – 2001. – Т. 14.

10. Yang M. S., Lai C. Y., Lin C. Y. A robust EM clustering algorithm for Gaussian mixture models //Pattern Recognition. – 2012. – Т. 45. – №. 11. – С. 3950-3961.

11. Gaiser J. E. Charmonium spectroscopy from radiative decays of the J/psi and psi'. – Stanford University, 1983.

12. Rahmah N., Sitanggang I. S. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra //IOP conference series: earth and environmental science. – IoP Publishing, 2016. – Т. 31. – №. 1. – С. 012012.

13. Sander J. et al. Density-based clustering in spatial databases: The algorithm gdbscan and its applications //Data mining and knowledge discovery. – 1998. – T. 2. – C. 169-194.