ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

на правах рукописи

Борисяк Максим Александрович

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КОНТРОЛЯ КАЧЕСТВА ДАННЫХ В НАУЧНЫХ ЭКСПЕРИМЕНТАХ

РЕЗЮМЕ

диссертации на соискание ученой степени
кандидата компьютерных наук

Москва — 2020

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

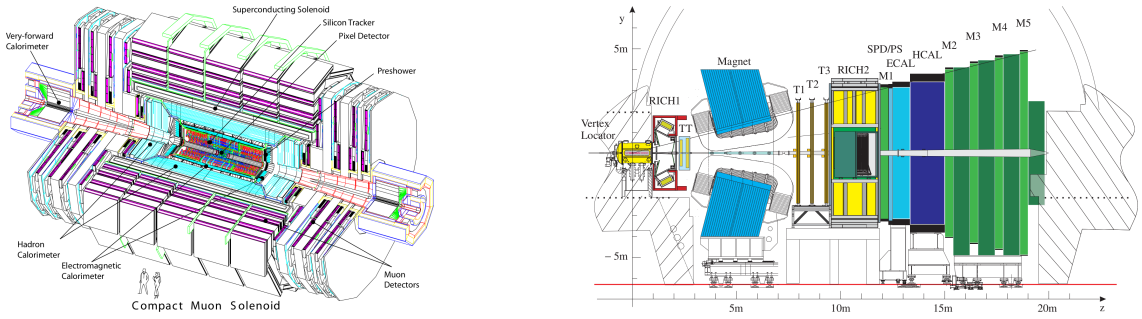
Научный руководитель: Устюжанин Андрей Евгеньевич, кандидат физико-математических наук, доцент базовой кафедры Яндекс департамента больших данных и информационного поиска факультета компьютерных наук НИУ ВШЭ, заведующий научно-учебной лабораторией методов анализа больших данных факультета компьютерных наук НИУ ВШЭ.

Содержание

1. Введение

Актуальность темы исследования. Получение данных и их обработка являются неотъемлемыми частями научных экспериментов. Во многих областях науки современные эксперименты полагаются на сложные детекторы и автоматизированные системы обработки данных. Например, в физике высоких энергий и астрофизике сбор данных и их обработка, как минимум, их начальные этапы, производятся исключительно автоматически. Для этих целей используются огромные вычислительные фермы: Большой Адронный Коллайдер производит миллионы событий в секунду, каждое из которых требует сложного анализа и должно быть обработано сразу после наблюдения [1, 2], современные обсерватории полагаются на большое число детекторов и производят значительный поток данных, например, планируемые вычислительные мощности для обсерватории Square Kilometre Array [3] превышают 100 PFLOPS [4].

Данные, полученные в современных научных экспериментах, обладают сложной структурой, и зачастую размерность данных превышает тысячи. Рисунок 1 демонстрирует сложность структуры современных экспериментальных установок: установка CERN CMS [1] состоит из множества подсистем, каждая из которых находится под контролем сложной электроники и сложного программного обеспечения, типичный размер события, считываемого детектором, находится в районе 0.5 Мб [2] при миллиарде столкновений в секунду. Упомянутый ранее Square Kilometre Array телескоп использует около 250 000 антенн, которые совместно считывают 2.5 Пб/сек сырых данных [5]. Методы машинного обучения демонстрируют высокую эффективность при обработке высокоразмерных данных и являются де-факто главными инструментами обработки данных в современных экспериментах [6–12].



(a) CERN CMS [1]. (b) CERN LHCb [13].

Рис. 1: Примеры современных экспериментальных установок.

Контроль качества данных (ККД, англ. data quality monitoring, DQM) является неотъемлемой частью систем получения и обработки данных. Главной задачей контроля качества данных является верификация валидности собираемых данных, иными словами, проверка того, что данные собраны при номинальных условиях эксперимента. Здесь аномалии определяются как отклонения от номинальных условий эксперимента и включают: человеческие ошибки, сбои в работе аппаратуры [14, 15], внешние события, например, сейсмическую активность [16] и даже наличие облаков [17, 18]. Отсутствие контроля за аномалиями приводит к искаженным данным, которые, в свою очередь, могут изменить результаты эксперимента [19] и привести к ложным открытиям. Например, наблюдения эксперимента OPERA [20], указывающие на нейтрино, распространяющиеся со сверхсветовой скоростью [21], были впоследствии объяснены техническими проблемами с аппаратурой [22].

Для современных экспериментальных установок контроль качества данных играет важную роль. Например, лазерно-интерферометрическая гравитационно-волновая обсерватория, LIGO [16], использует крайне чувствительную оптическую систему, поэтому сталкивается с различными источниками шума, включая внешние [23], в дополнение к сбоям в работе самой установки [24]. Искусственные сооружения могут существенно влиять на результаты гиперспектральной съемки, усложняя анализ состава почвы [25]. В медицине артефакты на изображениях, полученных с помощью магнитно-резонансной томографии, могут искажать результаты автоматической обработки этих изображений, что может привести к неправильным диагнозам [26]. В метеорологии неправильно сконфигурированные или плохо обслуживаемые метеорологические станции, ошибки при считывании инструментов, неточную дискретизацию и обработку данных часто относят к причинам ошибочных предсказаний [27]. Как и в случае с обработкой данных, системы контроля качества данных все больше и больше полагаются на алгоритмы машинного обучения, так как рассмотрение аномалий может только усложнить анализ данных [14, 24, 28–30].

На практике задачу контроля качества данных часто разбивают на две подзадачи: онлайн и оффлайн ККД (англ. online DQM, offline DQM). Онлайн ККД обычно решает проблемы, связанные с аппаратурой, и оперирует с сырыми данными или минимально обработанными данными [14, 15, 31, 32]. Структура данных зависит от конкретного детектора и различается от эксперимента к экспе-

рименту. Оффлайн ККД обычно проверяет данные на менее заметные иррегулярности, включая инспекцию результатов обработки данных. Оффлайн ККД обычно анализирует обработанные и агрегированные данные[1]. Это деление, однако, не является строгим, и некоторые эксперименты используют, например, дополнительные уровни ККД.

Более того, рассматривается задача, похожая на контроль качества данных — поиск различий между наблюдениями/экспериментальными данными и ожидаемыми результатами/теоретическими предсказаниями. Например, наблюдение бозона Хиггса [34, 35] можно рассматривать как иррегулярность в распределении инвариантной массы по отношению к так называемому фону (предсказания наилучшей теоретической модели, не содержащей бозон Хиггса). Слияние черных дыр [36] детектируется как неожиданно сильные осцилляции по отношению к фоновым шумам. Машинное обучение становится главным средством поиска несоответствий между теорией и экспериментальными данными при отсутствии альтернативных гипотез, ярким примером является поиск новой физики [37–43].

Поиск несоответствий между наблюдениями и теорией обычно рассматривается отдельно от контроля качества данных из-за различной природы отклонений и разницы в уровне обработки данных[2]. Так как данная работа рассматривает методы машинного обучения, различия между двумя задачами: контролем качества данных и поиском несоответствий между наблюдениями и теорией, не проводятся, и обе задачи рассматриваются как задачи обнаружения аномалий [38–43].

Терминология. В этой работе любое отклонение от номинальных условий эксперимента обозначается как аномальное. Номинальные условия эксперимента определяются самим экспериментом. Данные, полученные при аномальных условиях, называются аномальными или просто аномалиями. Несоответствие между теорией и экспериментом также называется аномалией.

Стоит обратить внимание на то, что эта терминология отличается от опре-

[1]Один из наиболее популярных методов заключается в сравнении оценок известных величин с их номинальными значениями [19, 33].

[2]Например, при поиске бозона Хиггса множество событий, каждое из которых занимает около 0.5 Мб дискового пространства, было агрегировано в несколько одномерных гистограмм [34]. В тоже время, контроль за тем же детектором использует низкоуровневые данные [14].

делений, использующихся в таких областях машинного обучения, как поиск выбросов. Последняя определяет аномалии или выбросы как наблюдения, которые значительно отличаются от остальной выборки [44, 45]. В случае контроля качества данных аномалии определяются не относительно остальной выборки, а относительно состояния эксперимента, который включает в себя состояние аппаратуры и состояние окружения. Несмотря на то, что аномалии в ККД с большой вероятностью значительно отличаются от нормальных данных, т.е. являются выбросами, они могут быть неотличимы от нормальных данных. Например, эксперимент CERN LHCb использует полупроводниковый трэкер, который состоит из множества полосок, которые регистрируют проходящие через них частицы [46]. Если часть этих полосок перестанет реагировать, что является аномалией, наблюдения все равно могут быть неотличимыми от наблюдений в нормальном состоянии, так как в некоторых редких, но вероятных случаях траектории частиц не пересекаются с полосками, которые вышли из строя.

Чтобы избежать путаницы в случаях, когда это не очевидно из контекста, задача нахождения аномалий (как определено в предыдущем параграфе) называется обнаружением аномалий, включая в себя задачи определения аномальных состояний детектора и задачу поиска разногласий между наблюдениями и теорией.

Цель и задачи исследования. Основная сложность, стоящая перед задачей контроля качества данных, кроется в свойствах аномальных данных. Некоторые аномалии могут быть неотличимы от нормальных примеров, особенно учитывая, что контроль качества данных производится только на части наблюдаемых величин (в терминологии, принятой в машинном обучении, признаков) или даже на агрегированных статистиках [27, 30, 47]. Крайне важно отличить такие случаи от остальных через соответствующие оценки на вероятности классов. Эта проблема также актуальна для поиска разногласий между теорией и экспериментальными наблюдениями, так как потенциальные несоответствия малы [37].

Более того, некоторые примеры аномалий или альтернативные гипотезы могут быть известны, а значит, должны быть учтены для получения адекватных оценок на вероятности классов [30]. В то же время, даже в случаях, когда примеры аномалий доступны, предположения о репрезентативности аномальной вы-

борки часто не могут быть выдвинуты [45]. Поэтому алгоритмы обнаружения аномалий должны быть устойчивы к новым типам аномалий, когда это возможно. С точки зрения машинного обучения такие требования означают, что задачи детектирования аномалий находятся в промежутке между обучением без учителя и обучением с учителем [48].

Как было замечено ранее, из-за природы аномалий системы контроля качества часто оперируют с сырыми или минимально обработанными данными. Большинство современных экспериментальных установок уникальны, и вместе с ними уникальна и структура получаемых данных. Это приводит к другой практически важной задаче: сбору данных для обучения алгоритмов обнаружения аномалий. Потенциально, два подхода могут быть применены:

- ручная разметка;
- автоматическая генерация нормальных примеров средствами компьютерных симуляций.

Также возможна комбинация этих подходов.

Первый подход, а именно ручная разметка, требует значительных трудозатрат [30, 47], поэтому алгоритмы, позволяющие сократить объем работ, крайне важны на практике. Подобные алгоритмы могут принять на себя значительную часть работы, что позволяет либо уменьшить затраты на разметку данных, либо увеличить объемы размеченной выборки при тех же затратах.

Второй подход опирается на факт, что многие эксперименты используют компьютерные симуляции [49–54]. Подобные симуляторы часто основываются на законах физики, выраженных в виде алгоритмов, например, в виде разностных схем для решения дифференциальных уравнений. Эти алгоритмы транслируют входные условия в наблюдаемые величины и часто зависят от параметров, определяющих законы физики, геометрию и прочие свойства. Компьютерные симуляции способны генерировать огромное количество примеров нормального поведения (в некоторых случаях также возможна симуляция некоторых типов аномалий). Эти примеры могут быть использованы для тренировки алгоритмов определения аномалий. Компьютерные симуляции особенно важны для поиска малых различий между экспериментальными данными и теоретическими предсказаниями, так как симуляции представляют собой теоретические модели [40, 43].

Компьютерные симуляции зачастую нуждаются в тонкой настройке: нахождению таких параметров симуляции, что выход симуляции наиболее близок к наблюдаемым данным [55–57]. Основное препятствие при тонкой настройке — высокая вычислительная сложность симуляторов [58], которая ведет к высокой сложности алгоритмов тонкой настройки, так как последние часто требуют больших выборок для своей работы.

Целью данной диссертации является разработка алгоритмов машинного обучения, решающих основные задачи, стоящие перед системами контроля качества, а именно:

- сбор данных:
  - алгоритмы для уменьшения трудозатрат на ручную разметку;
  - алгоритмы анализа аномалий;
  - эффективные алгоритмы тонкой настройки симуляторов;
- алгоритмы детектирования аномалий, способные учитывать известные примеры аномалий.

Для достижения этих целей должны быть выполнены следующие шаги:

- демонстрация того, что алгоритмы машинного обучения могут быть успешно применены для облегчения ручной разметки данных; тестирование этих алгоритмов на данных из больших экспериментальных установок;
- разработка методов анализа аномалий и тестирование этих методов на данных из больших экспериментальных установок;
- разработка методов для уменьшения вычислительной стоимости алгоритмов тонкой настройки;
- разработка методов детектирования аномалий, комбинирующих свойства одноклассовых и двухклассовых методов классификации, сравнение этих методов с ближайшими аналогами.

Рисунок 2 демонстрирует взаимосвязь между методами, рассматриваемыми в этой работе.

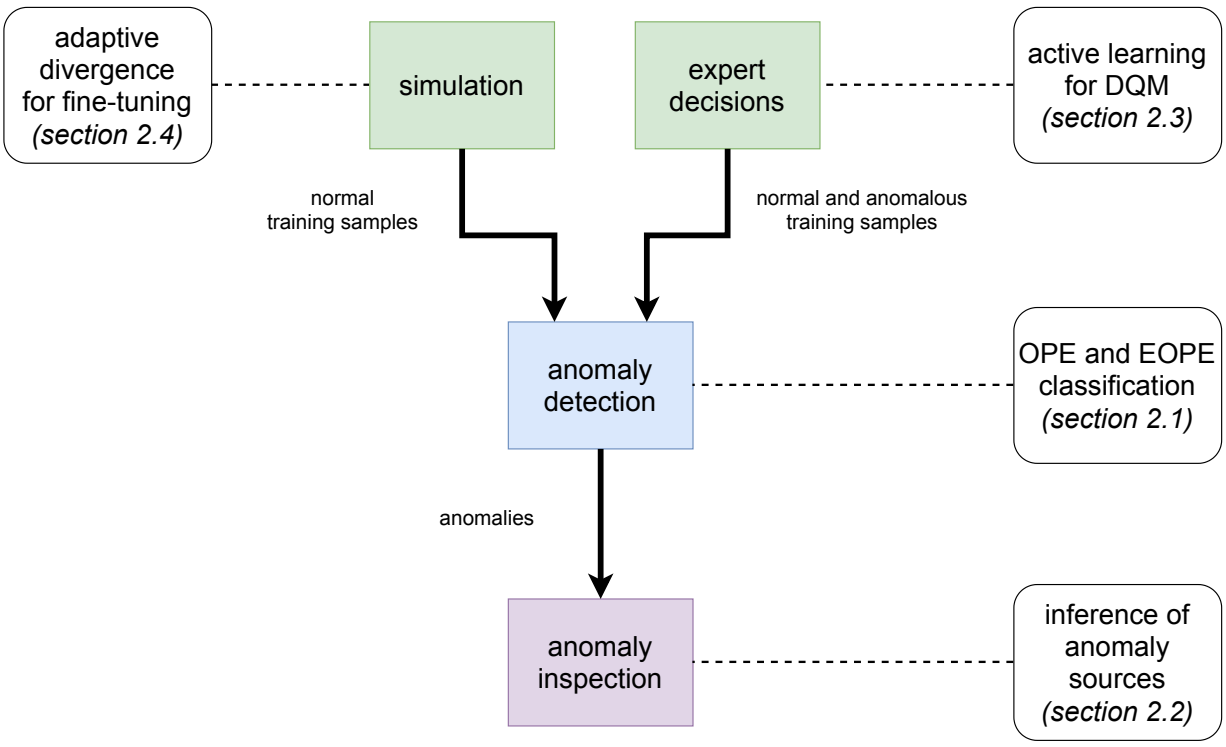


Рис. 2: Главные шаги систем контроля качества и соответствующие результаты диссертации.

Структура диссертации. Во второй главе приведено описание главных результатов диссертации. В секции 2.1 рассматриваются алгоритмы детектирования аномалий. Автор диссертации представляет новое семейство алгоритмов детектирования аномалий общего назначения, способных к работе в предположениях, часто рассматриваемых в системах контроля качества данных. Во-первых, рассматриваются существующие методы машинного обучения для детектирования аномалий и приводятся аргументы в пользу того, что существующие методы не покрывают наиболее распространенный случай ККД — большая репрезентативная выборка примеров нормального поведения и статистически не значимая или малая выборка аномальных примеров. Предложенное семейство алгоритмов полностью покрывает весь спектр задач детектирования аномалий: традиционные двухклассовые задачи (сбалансированная выборка), традиционные одноклассовые задачи и промежуточные задачи. Основные свойства предлагаемого метода строго доказаны, качество методов протестировано на наборе эталонных задач. Этот результат соответствует шагу «anomaly detection» на рисунке 2.

В секции 2.2 автор представляет новый алгоритм глубинного обучения, который при довольно общих предположениях выявляет источник аномалий, на-

пример, указывает на подсистемы, которые показывают аномальное поведение. Главным достоинством этого алгоритма является отсутствие требований на наличие меток для какой-либо подсистемы. Иными словами, алгоритм не требует дополнительной информации для тренировки. Этот алгоритм улучшает системы контроля данных, ускоряя выявление настоящих причин аномалий. Этот результат соответствует шагу «anomaly inspection» на рисунке 2.

В секции 2.3 рассматривается ручная разметка данных и предлагается использование активного обучения для снижения трудозатрат. Рассмотренный алгоритм постепенно тренируется на уже размеченных данных и способен принимать автоматические решения для примеров, похожих на те, которые уже были размечены экспертом. Данный алгоритм был протестирован на реальных данных, полученных экспериментом CERN CMS. Этот результат соответствует шагу «expert decisions» на рисунке 2.

Секция 2.4 посвящена основной проблеме, стоящей перед автоматической генерацией нормальных примеров средствами компьютерной симуляции — тонкой настройке симуляций (англ. fine-tuning). Основное внимание уделено высоким вычислительным затратам, связанным с алгоритмами тонкой настройки. Автор диссертации вводит новое семейство статистических дивергенций (англ. divergence) и новый класс соответствующих алгоритмов тонкой настройки. Данное семейство дивергенций сформулировано специально для уменьшения вычислительных затрат, связанных с их оценкой. Скорость работы алгоритмов тонкой настройки с предложенным семейством дивергенций протестировано на реалистичном примере с генератором событий Pythia. Этот результат соответствует шагу «simulation» на рисунке 2.

Степень разработанности темы исследования. Алгоритмы детектирования аномалий являются краеугольным камнем для систем контроля качества данных. Существующие подходы можно разделить на три категории: обучение с учителем, обучение без учителя и обучение с положительными и неразмеченными данными (англ. learning from positive and unlabeled data, PU learning).

В подходе с обучением с учителем детектирование аномалий рассматривается как задача бинарной классификации[3]. Такой подход демонстрирует хоро-

[3]В некоторых случаях класс аномалий разделен на подклассы (например, в работе [29]), что технически приводит к многоклассовой классификации. В данной работе не рассматриваются такие случаи, только аномальный и нормальный классы. Тем не менее, предложенные алгоритмы могут быть с легкостью адаптированы к много-

шие результаты в случаях относительно частых аномалий [14, 28, 30]. Однако, как показано в недавней работе [48], бинарная классификация неустойчива к малым или нерепрезентативным выборкам аномалий.

Методы обучения без учителя [59–63] широко применяются для обнаружения аномалий в случаях, когда аномалии редки или аномальная выборка нерепрезентативна, т.е. не покрывает всех возможных случаев аномального поведения. Некоторые алгоритмы используют ошибку реконструкции [30, 62]: их главная идея заключается в том, что алгоритм, натренированный на реконструкцию нормальных примеров, с малой вероятностью адекватно реконструирует аномалии, особенно если алгоритм натренирован как порождающая модель [64–66]. Метод описания данных через опорные вектора [67] (англ. Support Vector Data Description, SVDD) и схожий метод одноклассовых опорных векторов [61] (англ. one-class Support Vector Machine, one-class SVM) используют функцию потерь метода опорных векторов, при этом дополнительно уменьшая область, классифицированную как положительную. Как и все ядерные методы, основным недостатком метода описания данных через опорные вектора и метода одноклассовых опорных векторов является их высокая вычислительная сложность, что делает их непрактичными при работе с большими выборками[4]. Несколько алгоритмов определения аномалий строятся на похожих идеях: метод глубинного описания данных через опорные вектора [60] (англ. Deep SVDD) использует ограниченную нейронную сеть для нахождения нетривиального базиса для линейной версии метода описания данных через опорные вектора; одноклассовая нейронная сеть [59] (англ. one-class Neural Network) использует автокодировщик для нахождения базиса. Методы, основанные на деревьях принятия решений [68], используют эвристики, связанные с обучением деревьев принятия решений, однако, как и многие подобные методы, зачастую не справляются с задачами, в которых зависимости между признаками играют главную роль (что демонстрируется, например, в работах [59, 60, 69]).

Методы обучения без учителя показывают хорошее качество в случаях, когда классы не пересекаются или пересечение незначительно. С точки зрения систем контроля качества данных, главным недостатком данных подходов является игнорирование известных примеров аномалий, т.е. данные методы не

классовому случаю, например, введением дополнительного классификатора для аномалий.

[4]Например, две эталонных выборки, рассмотренные в работе [48], содержат более миллиона примеров.

могут предоставить достоверные оценки вероятностей классов в случаях с пересекающимися классами, зачастую предсказывая неоднозначные примеры как однозначно нормальные.

Обучение с положительными и неразмеченными данными [70] расматривает задачу близкую к задаче детектирования аномалий, а именно бинарную классификацию с размеченными положительными данными и неразмеченной смесью положительных и отрицательных примеров. Однако важное отличие между двумя областями состоит в том, что детектирование аномалий рассматривает случай нерепрезентативных или малых выборок, а не случай неполной информации о метках. Тем не менее, некоторые аналогии могут быть проведены. Некоторые методы обучения с положительными и неразмеченными данными рассматривают неразмеченные данные как отрицательный класс, что идейно похоже на метод, предложенный в данной диссертации [71, 72].

Другой важной задачей контроля качества данных является анализ аномалий. В данной диссертации рассматривается задача определения источников аномалий, т.е. определение подсистем, проявляющих аномальное поведение. Обычно такие задачи рассматриваются в области казуального/причинного вывода (англ. causal inference), обзор области можно найти в работе [73]. Как замечено в обзоре: «за каждым причинным выводом обязаны стоять предположения, не тестируемые с помощью наблюдений.» Насколько известно автору, предположения, рассматриваемые в данной диссертации, уникальны и не встречаются в литературе, в основном из-за того, что предположения включают отсутствие информации о метках для подсистем.

Третьей основной задачей, связанной с ККД, является сбор обучающих данных для алгоритмов обнаружения аномалий. Рассматриваются два подхода: ручная разметка и использование компьютерных симуляций. Минимизация человеческого труда при ручной разметке данных рассматривается в активном обучении — области машинного обучения, связанной с обучением на потоке данных или обратной связи от экспертов. Активное обучение рассматривает широкий спектр задач, различающихся, например, процедурами доступа к данным или базовой моделью данных [74]. Общий обзор активного обучения можно найти в работе [74]. В контексте контроля качества данных наиболее актуальным подходом является минимизация сбора данных (англ. minimization of data collection). Основная идея этого метода состоит в том, чтобы автоматически

принимать решения для однозначно нормальных или однозначно аномальных примеров, запрашивая метки экспертов во всех остальных случаях, с последующим обновлением модели [75]. Неоднозначность примера определяется различными эвристиками, например, измерением несогласия ансамбля классификаторов [76], использованием метрик «конфликта» (англ. conflict) и ”неизвестности” [77] (англ. ignorance) или использованием нечетких (англ. fuzzy) классификаторов [78].

Компьютерные симуляции часто требуют тонкой настройки параметров для конкретной экспериментальной установки. Методы тонкой настройки можно разделить на несколько категорий. Первая категория основывается на различных эвристиках для сопоставления истинного распределения и результата симуляции [55, 56, 79]. Основным недостатком этих методов является необходимость в специальных признаках, которые тщательно подобраны для удовлетворения предположений, лежащих в основе конкретной эвристики, что не всегда возможно на практике. Вторая категория тесно связана с генеративными моделями, в частности, с состязательными порождающими сетями [80] (англ. Generative Adversarial Networks) и выводом без правдоподобия [57, 81–84] (англ. likelihood-free inference). В эту категорию входят методы общего назначения, которые можно применять практически для любой компьютерной симуляции. Эти методы часто основаны на состязательном обучении [57] (англ. adversarial learning) или аналогичных подходах [81], что делает их вычислительно дорогими. Насколько известно автору, работа [85] является первой, в которой явно рассматривается вычислительная сложность методов тонкой настройки, в частности, для случаев с недифференцируемыми вычислительно затратными симуляциями.

Научная новизна. Основные результаты данной диссертации следующие.

- Предложено новое семейство алгоритмов для обнаружения аномалий. В отличие от традиционных одноклассовых методов классификации, предлагаемые методы сочетают в себе свойства двухклассовых и одноклассовых методов и способны решать проблемы в широком диапазоне предположений о природе аномалий.

- Предложен новый метод для определения источников аномалий. Метод протестирован на данных эксперимента CERN CMS. Алгоритм основан

на предположениях, которые часто выполнены для систем контроля качества данных, и не требует дополнительных меток для подсистем.

- Рассмотрен алгоритм активного обучения для значительного сокращения трудозатрат. Алгоритм протестирован на данных эксперимента CERN CMS;
- Представлено новое семейство статистических дивергенций, позволяющее значительно ускорить процедуры тонкой настройки по отношению к количеству обращений к симуляции.

Также следует отметить, что основные результаты этой работы могут быть применены к ситуациям вне систем контроля качества данных.

- Предложенные методы обнаружения аномалий являются универсальными методами, предназначенными для решения широкого спектра задач. Например, их можно легко адаптировать для задач вне систем контроля качества данных, для обучения на несбалансированных наборах данных [48] или для повышения устойчивости методов классификации [86].
- Предложенный метод для определения источников аномалий является методом общего назначения и может быть применен к промышленным задачам, которые согласуются с предположениями метода.
- Адаптивные дивергенции могут быть использованы при тренировке генеративных моделей общего назначения, например [87–89].

Практическая значимость. Результаты, полученные в ходе диссертационного исследования, напрямую применимы к системам контроля качества и позволяют:

- улучшить алгоритмы детектирования аномалий путем учета известных аномальных примеров;
- решать широкий спектр задач детектирования аномалий;
- улучшить анализ аномалий путем определения источников аномалий;
- значительно уменьшить затраты, связанные с тонкой настройкой вычислительно тяжелых компьютерных симуляций;
- значительно снизить трудозатраты при ручной разметке данных.

Методология и методы исследования. В работе используются методы теории вероятности и статистики, функциональный анализ, методы машинного обучения, экспертные знания из областей физики высоких энергий и астрофизики. Все алгоритмы были разработаны на языке программирования Python [90] с использованием библиотек numpy [91], scipy [92], scikit-learn [93], tensorflow [94], pytorch [95] и многих других. Все численные эксперименты воспроизводимы, код экспериментов доступен публично, ссылки на хранилища кода приведены в соответствующих работах.

Публикации и апробация работы. Все результаты данного диссертационного исследования были опубликованы в международных рецензируемых журналах.

Публикации повышенного уровня:

- (1 + epsilon)-class Classification: an Anomaly Detection Method for Highly Imbalanced or Incomplete Data Sets / M. Borisyak, A. Ryzhikov, A. Ustyuzhanin, D. Derkach, F. Ratnikov, O. Mineeva // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 72. — P. 1–22. (Scopus Q1)
  Вклад автора диссертации: синтез одноклассовой и двухклассовой классификаций, соответствующая функция потерь, вывод энергетической аппроксимации функции потерь, доказательства для асимптотического случая, эффективный алгоритм тренировки, эксперименты на эталонных задачах. Автор диссертации является главным автором данной работы.

- Adaptive divergence for rapid adversarial optimization / M. Borisyak, T. Gaintseva, A. Ustyuzhanin // PeerJ Computer Science. — 2020. — May. — Vol. 6. — P. e274. (Scopus Q1);
  Вклад автора диссертации: определение адаптивных дивергенций, формулировка нескольких конкретных семейств адаптивных дивергенций, вычислительно эффективный алгоритм для оценки адаптивных дивергенций, основанных на нейронных сетях и ансамблях деревьев, доказательства, эксперименты. Автор диссертации является главным автором данной работы.

Публикации стандартного уровня:

- Deep learning for inferring cause of data anomalies / V. Azzolini, M. Borisyak, G. Cerminara, D. Derkach, G. Franzoni, F. De Guio, O. Koval, M. Pierini, A. Pol, F. Ratnikov, F. Siroky, A. Ustyuzhanin, J-R. Vlimant. // Journal of Physics: Conference Series. — 2018. — sep. — Vol. 1085. — P. 042015. (Scopus Q3);
  Вклад автора диссертации: архитектура функция потерь для нейронной сети, доказательство, предварительные эксперименты на данных с экспериментальной установки CERN CMS.

- Towards automation of data quality system for CERN CMS experiment / M. Borisyak, F. Ratnikov, D. Derkach, A. Ustyuzhanin // Journal of Physics: Conference Series. — 2017. — oct. — Vol. 898. — P. 092041. (Scopus Q3).
  Вклад автора диссертации: эксперимент на данных с экспериментальной установки CERN CMS. Автор диссертации является главным автором данной работы.

2. Основные результаты

2.1. Детектирование аномалий

Алгоритмы детектирования аномалий лежат в основе любой системы контроля качества данных. Как упомянуто выше, под аномалией здесь понимается любое отклонение от номинальных условий эксперимента. Стоит обратить внимание, что это определение отличается от того, которое наиболее часто используется в областях обучения без учителя и детектирования выбросов. Ключевое отличие состоит в том, что аномалия здесь может быть представлена тем же вектором признаков, что и некоторое номинальное состояние. Иными словами, распределения аномальных и номинальных условий потенциально имеют пересекающиеся носители.

В работе [48] приводятся аргументы в пользу того, что такие обобщенные предположения более аккуратно описывают практические задачи детектирования аномалий, чем традиционные постановки задач [45]. Контроль качества данных часто происходит на обработанных данных, например, в работе [47] алгоритм обнаружения аномалий получает несколько статистик, агрегированных по большому количеству событий, что делает аномалии, присутствующие

только в нескольких событиях, практически неотличимыми от маловероятной, но возможной при номинальных условиях серии событий. Кроме того, в некоторых случаях условия эксперимента не полностью наблюдаемы, что потенциально приводит к тому, что некоторые аномалии дают те же значения наблюдаемых величин, что и при некоторых номинальных условиях. Например, трэкер CERN LHCb [46] состоит из большого количества кремниевых пикселей, которые регистрируют частицы, проходящие через них. Если группа таких пикселей перестает работать (что является аномалией), данные трэкера могут соответствовать некоторым редким, но возможным событиям, например, тем, в которых траектории частиц не проходят через эту группу пикселей.

Номинальные условия научных экспериментов часто определяются ограниченным множеством состояний. Более того, данных о номинальном поведении зачастую много, при этом аномалии относительно редки, например, в эксперименте CERN CMS только около 2% примеров аномальны [30]. Учитывая вышесказанное, выдвигаются следующие предположения, которые определяют обобщенную задачу детектирования аномалий:

- носитель аномального класса может быть не полностью отделимым от носителя номинального класса;
- некоторые виды аномалий могут не присутствовать в обучающей выборке.

В этих предположениях важно корректно идентифицировать неоднозначные примеры, особенно учитывая, что аномальность подобных примеров может быть определена экспертом при учете дополнительной информации.

В этой работе рассматриваются нейронные сети. Среди других популярных методов алгоритмы, основанные на деревьях принятия решений, плохо решают задачи, в которых присутствуют сильные связи между признаками (сравнение нейронных сетей и методов на основе деревьев принятия решений можно найти, например, в работах [59, 60, 69]); применение алгоритмов, основанных на методе опорных векторов [96], затруднительно из-за их высокой вычислительной сложности [97].

Технически задача детектирования аномалий в постановке, описанной выше, является задачей классификации. Пусть $\mathcal{X}$ — Банахово пространство, представляющее пространство всевозможных наблюдаемых в эксперименте вели-

чин. Тогда оптимальный баейсовский классификатор $f^* : \mathcal{X} \to [0, 1]$:

$$f^*(x) = \frac{P(x \mid \mathcal{C}^+)P(\mathcal{C}^+)}{P(x \mid \mathcal{C}^+)P(\mathcal{C}^+) + P(x \mid \mathcal{C}^-)P(\mathcal{C}^-)}; \tag{1}$$

где $\mathcal{C}^+$, $\mathcal{C}^-$ обозначают нормальный и аномальный классы и соответствующие апостериорные распределения.

Традиционно, бинарная классификация тренируется минимизацией кросс-энтропийной функции потерь (англ. cross-entropy loss function):

$$\mathcal{L}_2(f) = -P(\mathcal{C}^+) \mathop{\mathbb{E}}_{x \sim \mathcal{C}^+} \log f(x) - P(\mathcal{C}^-) \mathop{\mathbb{E}}_{x \sim \mathcal{C}^-} \log\left(1 - f(x)\right); \tag{2}$$

где $\mathbb{E}_{x \sim \mathcal{C}}$ обозначает условное среднее $\mathbb{E}_x\left[\cdot \mid \mathcal{C}\right]$.

Стоит заметить, что оптимальный классификатор (1) не определен вне объединения носителей $\operatorname{supp} \mathcal{C}^+ \cup \operatorname{supp} \mathcal{C}^-$, и, в общем, функция f^*, минимизирующая $\mathcal{L}_2$, может принимать любые значения для $x \notin \operatorname{supp} \mathcal{C}^+ \cup \operatorname{supp} \mathcal{C}^-$. В случае конечной выборки это ведет к отсутствию каких-либо гарантий на предсказания классификатора в областях, не покрытых тренировочной выборкой. Предсказания в этих областях могут зависеть от конкретной архитектуры сети, инициализации и даже от последовательности подвыборок (англ. mini-batches).

Такое поведение идет вразрез с предположениями, что некоторые типы аномалий не присутствуют в тренировочной выборке. Учитывая предположения о достаточно больших объемах выборки нормальных примеров, требуется, чтобы любой $x \notin \mathcal{C}^+$ был классифицирован как аномалия, т.е., $f^*(x) = 0$.

В работе [48] мы предлагаем добавление равномерно распределенного шума U к аномальному классу, где $\operatorname{supp} U = \Omega \subseteq \mathcal{X}$ — компакт, покрывающий оба класса. В этом случае функция потерь (2) становится:

$$\begin{aligned}
\mathcal{L}_{1+\varepsilon}(f) &= \frac{1}{2}\left(L^+(f) + \gamma\, L^-(f) + (1-\varepsilon)\, L^0(f)\right); \\
L^+(f) &= -\mathop{\mathbb{E}}_{x \sim \mathcal{C}^+} \log f(x); \\
L^-(f) &= -\mathop{\mathbb{E}}_{x \sim \mathcal{C}^-} \log(1 - f(x)); \\
L^0(f) &= -\mathop{\mathbb{E}}_{x \sim U} \log(1 - f(x));
\end{aligned} \tag{3}$$

при этом решение:

$$f^*_{1+\varepsilon}(x) = \frac{P(x \mid \mathcal{C}^+)}{P(x \mid \mathcal{C}^+) + (1-\varepsilon)\, C + \gamma\, P(x \mid \mathcal{C}^-)}; \tag{4}$$

где: $C = \text{const}$ плотность распределения U, $\varepsilon \in [0, 1]$ регулирует влияние L^0, и γ должно быть выставлено таким образом, чтобы:

$$\gamma + (1 - \varepsilon) = 2 \cdot \frac{P(\mathcal{C}^-)}{P(\mathcal{C}^+)};$$

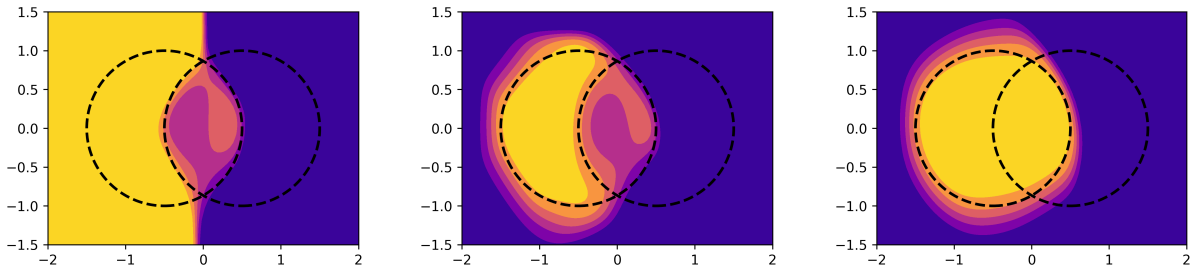
если априорные вероятности классов известны. Уравнение (3) задает так называемую OPE функцию потерь.

Как можно видеть из уравнения (4), для $\varepsilon = 1$ оптимальный классификатор является решением двухклассовой классификации (1), в том время как для $\varepsilon = 0$ и $\gamma = 0$ оптимальный классификатор является монотонным преобразованием $P(x \mid \mathcal{C}^+)$, иными словами, является одноклассовым решением. Промежуточные случаи позволяют привнести свойства одноклассового решения, а именно $f(x) = 0$ для $x \notin \operatorname{supp} \mathcal{C}^+$, в решение бинарной классификации. Другими словами, регуляризация L^0 смещает решение в сторону решения одноклассовой классификации, что делает оптимальный классификатор определенным вне носителя обоих классов, в то время как при ε, близкому к 1, регуляризация имеет слабое влияние в остальном пространстве. Рисунок 3 демонстрирует этот феномен на синтетических данных: в то время как двухклассовая классификация выдает корректные оценки на вероятности классов внутри носителя классов, предсказания положительны за пределами положительного класса. В то же время, одноклассовая классификация, корректно предсказывая отрицательный класс вне носителя положительного, игнорирует наличие отрицательных примеров, что приводит к некорректным оценкам внутри области пересечения носителей классов; классификатор, минимизирующий OPE функцию потерь, корректно предсказывает вероятности классов на всем пространстве.

Алгоритм 1 демонстрирует процедуру обучения, основанную на OPE функции потерь, которая далее обозначена как brute-force OPE.

OPE функция потерь приводит к решению с желаемыми свойствами, brute-force OPE демонстрирует главный недостаток этого подхода: оценка $\nabla L^0(f)$ по малой выборке крайне шумна, если $\mathcal{X}$ — высокоразмерное пространство, или если f аппроксимировано мощной моделью, что типично для случая нейронных сетей. Дисперсия оценки градиента напрямую влияет на сходимость стохастического градиентного спуска, что делает регуляризацию L^0 неэффективной на практике во многих случаях.

Для уменьшения дисперсии градиентов регуляризации мы предлагаем сле-



(a) двухклассовая классификация

(b) OPE классификация

(c) одноклассовая классификация.

Рис. 3: Демонстрация основного свойства OPE функции потерь. Примеры равномерно распределены внутри областей, обозначенных окружностями: левая окружность соответствует положительному классу, правая — отрицательному. Одноклассовое решение получено с помощью OPE функции потерь при $\gamma = 1$ и $\varepsilon = 0$. Обучающая выборка не показана для наглядности.

дующую регуляризацию:

$$\begin{aligned} L^E(g) &= \int_\Omega \exp(g(x))dx; \\ \text{где: } g(x) &= \sigma^{-1}(f(x)); \\ \sigma(\chi) &= \frac{1}{1+\exp(-\chi)}; \end{aligned}$$

которая именуется как энергетическая регуляризации везде далее, соответствующую функцию потерь, на которую мы ссылаемся как на energy OPE или EOPE.

В работе [48], мы показываем, что L^E приводит к решениям с желаемыми свойствами, т.е. функция потерь:

$$\mathcal{L}_1^E(g) = \frac{1}{2}\left[\mathop{\mathbb{E}}_{x\sim\mathcal{C}^+} \log\left(1+\exp(-g(x))\right) + (1-\varepsilon)L^E(g)\right] \tag{5}$$

приводит к одноклассовому решению.

Формально это свойство формулируется в следующей теореме.

Теорема 1 *Пусть $(\mathcal{X}, \|\cdot\|)$ — Банахово пространство, $P(x)$—непрерывная функция плотности вероятности, такая, что $\Omega = \operatorname{supp} P$ есть открытое множество в $\mathcal{X}$. Если непрерывная функция $g^* : \Omega \to \mathbb{R}$ минимизирует $\mathcal{L}_1^E$ (заданное уравнением 5) с $P(x \mid \mathcal{C}^+) = P(x)$, тогда существует строго возрастающая функция $s : \mathbb{R} \to \mathbb{R}$, такая, что $g^*(x) = s(P(x))$. Более того, $\lim_{y\to 0} s(y) = -\infty$, если $\inf_\Omega P = 0$.*

Алгоритм 1: Brute-force OPE

Input: normal data, anomalous data—samples from $\mathcal{C}^+$, $\mathcal{C}^-$, the latter might be absent; f_θ—a classifier with parameters θ.

Hyper-parameters: γ—the ratio of class priors; ε—the strength of regularization.

while *not converged* **do**

sample normal data $\{x_i^+ \sim \text{normal data}\}_{i=1}^m$;

sample known anomalies $\{x_i^- \sim \text{anomalous data}\}_{i=1}^m$;

sample pseudo-negative examples $\{x_i^0 \sim U[\Omega]\}_{i=1}^m$;

$\nabla L^+ \leftarrow -\sum_i \nabla_\theta \log f_\theta(x_i^+)$;

$\nabla L^- \leftarrow -\sum_i \nabla_\theta \log(1 - f_\theta(x_i^-))$;

$\nabla L^0 \leftarrow -\sum_i \nabla_\theta \log(1 - f_\theta(x_i^0))$;

$\theta \leftarrow \text{Adam}\left(\nabla L^+ + \gamma \nabla L^- + (1-\varepsilon)\nabla L^0\right)$

end

Доказательство теоремы 1 можно найти в соответствующей работе [48].

Значительным преимуществом L^E регуляризации является тот факт, что по сравнению с L^0, градиенты L^E могут быть оценены намного точнее:

$$\nabla L^E(g) = \frac{1}{Z}\int_\Omega \exp(g(x))\nabla g(x) = \mathop{\mathbb{E}}_{x\sim P_g} \nabla g(x). \tag{6}$$

Уравнение (6) основано на свойствах, широко применяемых для энергетических моделей. Это означает, что большинство алгоритмов, использующихся для тренировки энергетических моделей, могут быть применены здесь, в том числе, контрастная дивергенция (англ. contrastive divergence) с марковскими цепями и глубинные направленные порождающие сети [98] (англ. Deep Directed Generated Networks). Это приводит к целому семейству алгоритмов. Алгоритм 2 описывает общую процедуру тренировки моделей с EOPE функцией потерь, рисунки 4 и 5 демонстрируют результаты работы OPE и EOPE алгоритмов на синтетических данных.

Также стоит заметить, что EOPE функция потерь устойчива к неточным процедурам сэмплирования, поэтому была предложена неточная, но вычислительно эффективная процедура сэмплирования, которая демонстрирует качество сравнимое с точными марковскими цепями.

Предложенные алгоритмы были протестированы на наборе популярных эта-

Алгоритм 2: Energy OPE

Input: normal data, anomalous data—samples from $\mathcal{C}^+$, $\mathcal{C}^-$, the latter might be absent; g_θ—a classifier with parameters θ.

Hyper-parameters: γ—the ratio of class priors; ε—the strength of regularization; MCMC—a Monte-Carlo sampling procedure.

while *not converged* **do**

sample normal data $\{x_i^+ \sim \text{normal data}\}_{i=1}^m$;

sample known anomalies $\{x_i^- \sim \text{anomalous data}\}_{i=1}^m$;

sample pseudo-negative examples $\{x_i^0 \sim \text{MCMC}\,[x \mapsto \exp(g(x))]\}_{i=1}^m$;

$\nabla L^+ \leftarrow \sum_i \nabla_\theta \log(1 + \exp(-g_\theta(x_i^+))$;

$\nabla L^- \leftarrow \sum_i \nabla_\theta \log(1 + \exp(g_\theta(x_i^-)))$;

$\nabla L^E \leftarrow \sum_i \nabla_\theta g_\theta(x_i^0)$;

$\theta \leftarrow \text{Adam}\left(\nabla L^+ + \gamma \nabla L^- + (1-\varepsilon)\nabla L^E\right)$

end

лонных задач, включая изображения (MNIST, CIFAR-10, Omniglot), обнаружение аномалий (KDD-99) и данные из физики высоких энергий (HIGGS, SUSY). Все оригинальные задачи были изменены, чтобы соответствовать условиям задач обнаружения аномалий, рассматриваемых в этой работе: для многоклассовых задач один из классов был выбран как нормальный, остальные были помечены как аномальные, и только некоторые из аномальных классов присутствовали в обучающей выборке. Для задач бинарной классификации число аномальных примеров, использованных для обучения, варьировалось. Эффективность предложенных алгоритмов сравнивалась с несколькими современными одноклассовыми и двухклассовыми методами классификации, а также с обучением с частичным привлечением учителя. Результаты представлены в таблицах 6 - 8. Эксперименты показывают, что предлагаемые методы либо превосходят базовые методы, либо достигают сопоставимых результатов. Во-первых, как и ожидалось, качество методов OPE и EOPE улучшается с добавлением известных отрицательных примеров и быстро приближается к качеству двоичной классификации для полных (сбалансированных) выборок. Во-вторых, методы OPE и EOPE демонстрируют наилучшее относительное качество для выборок со значительно перекрывающимися классами (таблицы 6 и 7). В целом результаты

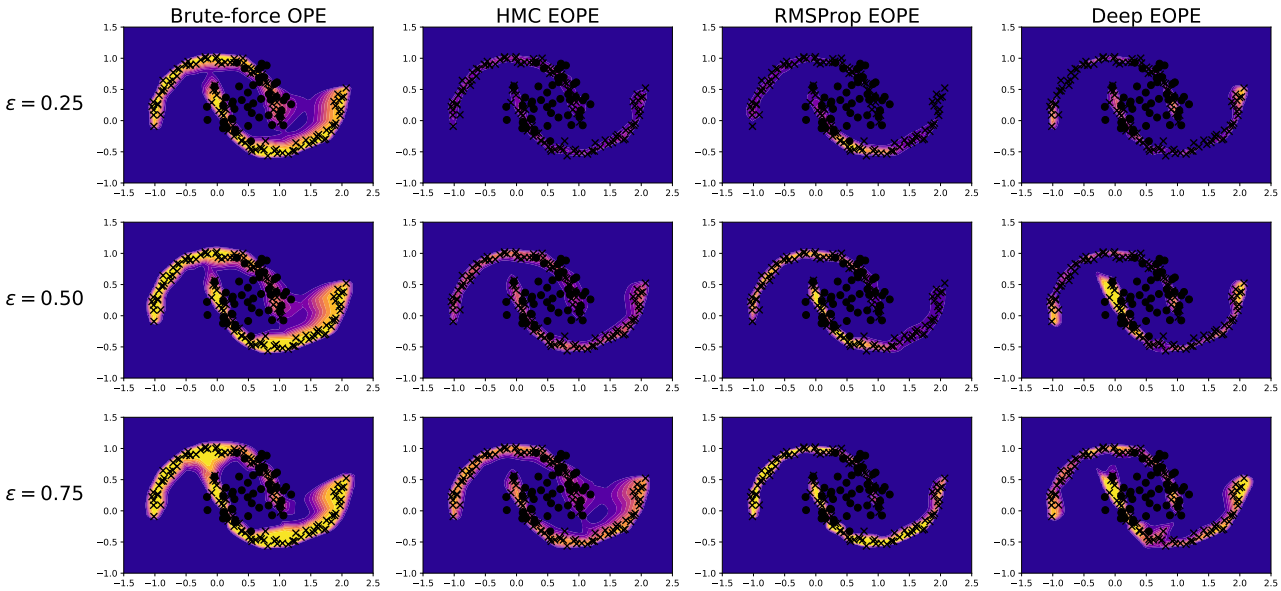
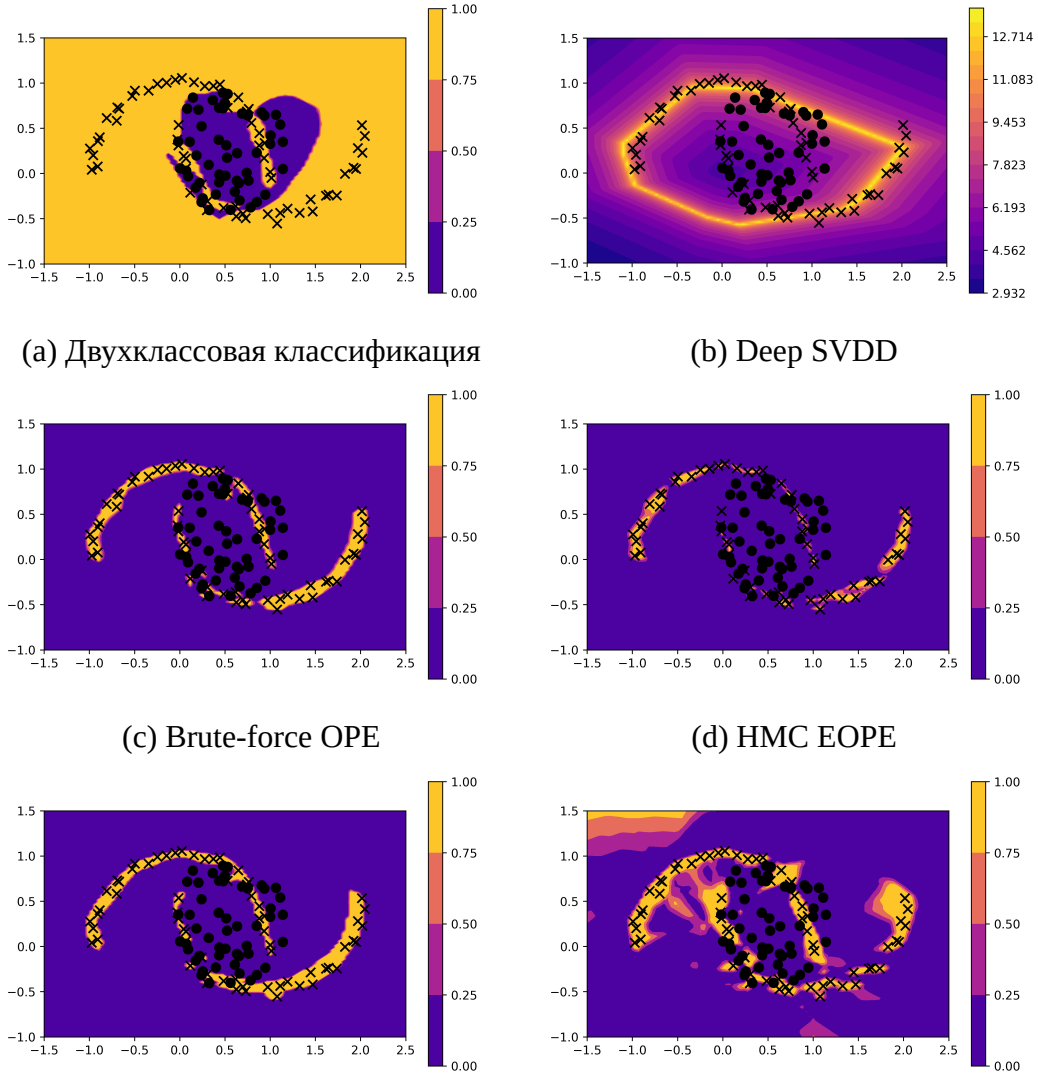


Рис. 4: Сравнение OPE и EOPE функций потерь для различных ε, для иллюстрации $\gamma = 1 - \varepsilon$. Для $\varepsilon < 1$ все функции потерь приводят к похожим решениям. Судя по результатам, EOPE регуляризация является более сильной регуляризацией, чем OPE регуляризация.

показывают, что OPE и EOPE хорошо подходят для решения задач обнаружения аномалий в контроля качества данных.



(a) Двухклассовая классификация

(b) Deep SVDD

(c) Brute-force OPE

(d) HMC EOPE

(e) RMSProp EOPE

(f) Deep EOPE

Рис. 5: Сравнение различных методов на синтетическом примере: положительные примеры (помеченные как ’x’) сгенерированы из набора данных «луны», отрицательные примеры (черные круги) равномерно распределены в круге с радиусом $\frac{1}{2}$. Для чистоты восприятия отображен отрицательный логарифм выхода Deep SVDD.

	one class	100	1000	10000	1000000
Robust AE	0.530 ± 0.002	0.530 ± 0.002	0.530 ± 0.002	0.530 ± 0.002	0.530 ± 0.002
Deep SVDD	0.497 ± 0.006	0.497 ± 0.006	0.497 ± 0.006	0.497 ± 0.006	0.497 ± 0.006
cross-entropy	-	0.496 ± 0.017	0.529 ± 0.007	0.566 ± 0.006	0.858 ± 0.002
semi-supervised	-	0.498 ± 0.003	0.522 ± 0.003	0.603 ± 0.002	0.745 ± 0.005
brute-force OPE	0.499 ± 0.009	0.500 ± 0.009	0.520 ± 0.003	0.572 ± 0.005	0.859 ± 0.001
HMC EOPE	0.491 ± 0.000	0.523 ± 0.005	$\mathbf{0.567} \pm 0.008$	$\mathbf{0.648} \pm 0.005$	0.848 ± 0.001
RMSProp EOPE	0.498 ± 0.002	0.494 ± 0.008	0.531 ± 0.008	0.593 ± 0.011	$\mathbf{0.861} \pm 0.000$
Deep EOPE	$\mathbf{0.531} \pm 0.000$	$\mathbf{0.537} \pm 0.011$	0.560 ± 0.008	0.628 ± 0.005	0.860 ± 0.001

Рис. 6: Результаты на задаче HIGGS. Первая строка показывает количество аномальных примеров в обучающей выборке.

	one class	100	1000	10000	1000000
Robust AE	0.394 ± 0.012	0.394 ± 0.012	0.394 ± 0.012	0.394 ± 0.012	0.394 ± 0.012
Deep SVDD	0.541 ± 0.022	0.541 ± 0.022	0.541 ± 0.022	0.541 ± 0.022	0.541 ± 0.022
cross-entropy	-	0.658 ± 0.033	0.736 ± 0.021	0.757 ± 0.036	0.871 ± 0.006
semi-supervised	-	0.715 ± 0.020	0.766 ± 0.009	$\mathbf{0.847} \pm 0.002$	0.876 ± 0.000
brute-force OPE	$\mathbf{0.648} \pm 0.035$	0.678 ± 0.025	0.729 ± 0.029	0.757 ± 0.036	0.871 ± 0.006
HMC EOPE	0.472 ± 0.000	$\mathbf{0.738} \pm 0.019$	$\mathbf{0.770} \pm 0.012$	0.816 ± 0.006	0.877 ± 0.000
RMSProp EOPE	0.443 ± 0.038	0.714 ± 0.019	0.760 ± 0.016	0.807 ± 0.004	0.877 ± 0.000
Deep EOPE	0.468 ± 0.118	0.670 ± 0.054	0.746 ± 0.024	0.813 ± 0.003	$\mathbf{0.878} \pm 0.000$

Рис. 7: Результаты на задаче SUSY. Первая строка показывает количество аномальных примеров в обучающей выборке.

	one class	1	2	4	8
Robust AE	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.972} \pm 0.006$
Deep SVDD	0.939 ± 0.014	0.939 ± 0.014	0.939 ± 0.014	0.939 ± 0.014	0.939 ± 0.014
cross-entropy	-	0.571 ± 0.213	0.300 ± 0.182	0.687 ± 0.268	0.619 ± 0.257
semi-supervised	-	0.315 ± 0.258	0.469 ± 0.286	0.758 ± 0.171	0.865 ± 0.087
brute-force OPE	0.398 ± 0.108	0.667 ± 0.175	0.394 ± 0.261	0.737 ± 0.187	0.541 ± 0.257
HMC EOPE	0.786 ± 0.200	0.885 ± 0.152	0.919 ± 0.055	0.863 ± 0.094	0.958 ± 0.023
RMSProp EOPE	0.765 ± 0.216	0.824 ± 0.237	0.770 ± 0.213	0.941 ± 0.048	0.960 ± 0.021
Deep EOPE	0.602 ± 0.279	0.767 ± 0.245	0.548 ± 0.279	0.763 ± 0.217	0.786 ± 0.267

Рис. 8: Результаты на задаче KDD-99. Первая строка показывает количество аномальных подклассов, входящих в обучающую выборку, максимум 1000 примеров из каждого аномального подкласса присутствует в обучающей выборке.

2.2. Определение источников аномалий

В большинстве реальных задач обнаружение аномалий сопровождается анализом причин аномалий. В сложных экспериментах установки обычно состоят из нескольких субдетекторов [1,15], каждый из которых считывает свой собственный набор значений, и определение подмножества субдетекторов, затронутых определенной аномалией, является важной задачей [99].

Кроме того, можно сделать следующее предположение: измеренные значения можно разбить на группы таким образом, чтобы аномалия, влияющая на подмножество этих групп, не влияла на значения из других групп. Такие группы далее называются каналами. Как правило, каждый субдетектор соответствует своему каналу, так как аномалия в субдетекторе не мешает работе другой аппаратуры, не затронутой этой аномалией.

В соответствующей работе [99] мы рассматриваем задачу определения подмножества каналов, затронутых аномалией. Кроме того, мы предполагаем, что метки для каналов, то есть индикаторы того, что аномалия затронула конкретный канал, недоступны, а именно присутствуют только глобальные метки, то есть индикаторы наличия аномалии по крайней мере в одном канале, без указания затронутых каналов.

Для каждого канала мы вводим свою нейронную сеть и объединяем их выходы следующей функцией активации:

$$\varphi(x) = \exp\left(\sum_{j=1}^{n} f^j(x^j) - n\right); \tag{7}$$

где: n — количество каналов, $x^j \in \mathcal{X}^j$ — вектор признаков, соответствующий j-ому каналу, $f^j : \mathcal{X}^j \mapsto [0,1]$ — сеть, соответствующая j-ому каналу.

Объединенная сеть тренируется путем минимизации кросс-энтропии (англ. cross-entropy) выхода функции φ по отношению к глобальным меткам.

Теорема 2 *Если доля аномальных примеров меньше $1/2$, количество каналов $n \geq 4$, для каждого канала $j \in \{1, \ldots, n\}$:*

$$\operatorname{supp} P(x^j \mid A^j) \cap \operatorname{supp} P(x^j \mid \bar{A}^j) = \varnothing;$$

где A^j и $\bar{A}^j$ события, обозначающие присутствие и отсутствие аномалии, затрагивающей j-ый канал, тогда решение $\{g^j : \mathcal{X} \mapsto [0,1]\}_{i=1}^n$, минимизирую-

щее кросс-энтропийную функцию потерь:

$$\mathcal{L}[\varphi] = -\frac{1}{N}\sum_{i}^{N}\left[y_i \log\varphi(x_i) + (1-y_i)\log(1-\varphi(x_i))\right]; \tag{8}$$

$$\varphi(x) = \exp\left(\sum_{j=1}^{n} f^j(x^j) - n\right); \tag{9}$$

раскладывает аномалии по каналам, т.е. для каждого канала j:

$$g^j(x^j) = \begin{cases} 1, & \text{если } \bar{A}^j; \\ 0, & \text{иначе}. \end{cases} \tag{10}$$

Доказательство теоремы может быть найдено в работе [99].

Неформально теорема говорит о том, что при определенных условиях сеть, натренированная вышеуказанным способом, восстанавливает метки для каждого канала. Интуиция, стоящая за доказательством, следующая: для каждого канала можно выделить 3 случая: аномалии нет, аномалия присутствует и затрагивает рассматриваемый канал, аномалия присутствует, но не затрагивает рассматриваемый канал. В первом и втором случаях функция потерь принимает минимальное значение, когда выход подсети, соответствующей рассматриваемому каналу, близок либо к 1 (аномалия отсутствует), либо к 0 (при наличии аномалии в канале). В третьем случае глобальная метка не соответствует метке канала, т.е. минимум функции потерь в этом случае достигается, если подсеть выдает ответ, расходящийся с глобальной меткой. Но из-за предположений, в частности, из-за условия на долю аномалий, и специального вида функции активации потери в этом случае всегда компенсируются уменьшением потерь для случая отсутствия аномалий, поэтому совместный минимум функции потерь для обоих случаев достигается, если подсеть выдает 1.

Предложенный метод был протестирован на данных, полученных экспериментом CERN CMS, которые были размечены вручную; основные метрики качества представлены на рисунке 9. Во-первых, выходы сетей ожидаемо сконцентрированы в районе 0 и 1, и только в редких случаях между ними. Во-вторых, все сети правильно определяют нормальные образцы с высокой точностью. Кроме того, для значительной части аномальных примеров все сети предсказывают аномалию. Однако следует отметить, что прогнозирование 1

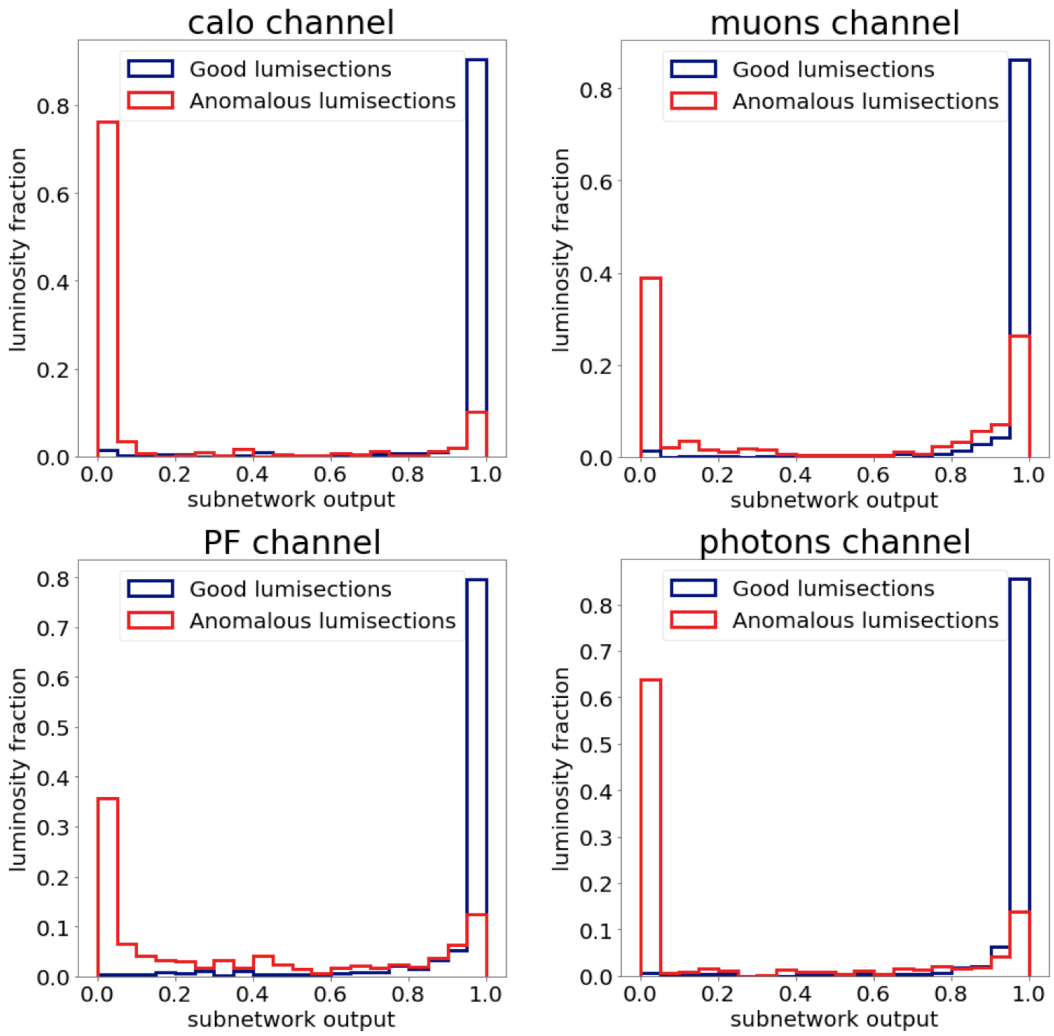


Рис. 9: Результаты предложенного метода на данных эксперимента CERN CMS: примеры обозначены как lumisections, luminosity — вес примера.

для аномального примера необязательно указывает на ошибку, так как аномалия могла произойти в другом канале. Чтобы дополнительно оценить эффективность метода, выходные данные каждой сети были оценены по меткам для отдельных подсистем эксперимента (рисунок 10): выходные данные каждой сети с высокой точностью предсказывают аномалии в подсистемах, связанных с ее каналом; в то же время, заметно уменьшение качества предсказаний на подсистемах, не связанных с соответствующим каналом. Например, аномалии в субдетекторе «мюоны» сильно коррелируют с выходами сети, которая соответствует «мюонному» каналу, и корреляция значительно ниже для других каналов. Обратите внимание, что ожидается положительная работа по отношению к субдетекторам, не связанным с соответствующим каналом, поскольку многие аномалии задевают несколько каналов одновременно. В целом, результаты сети

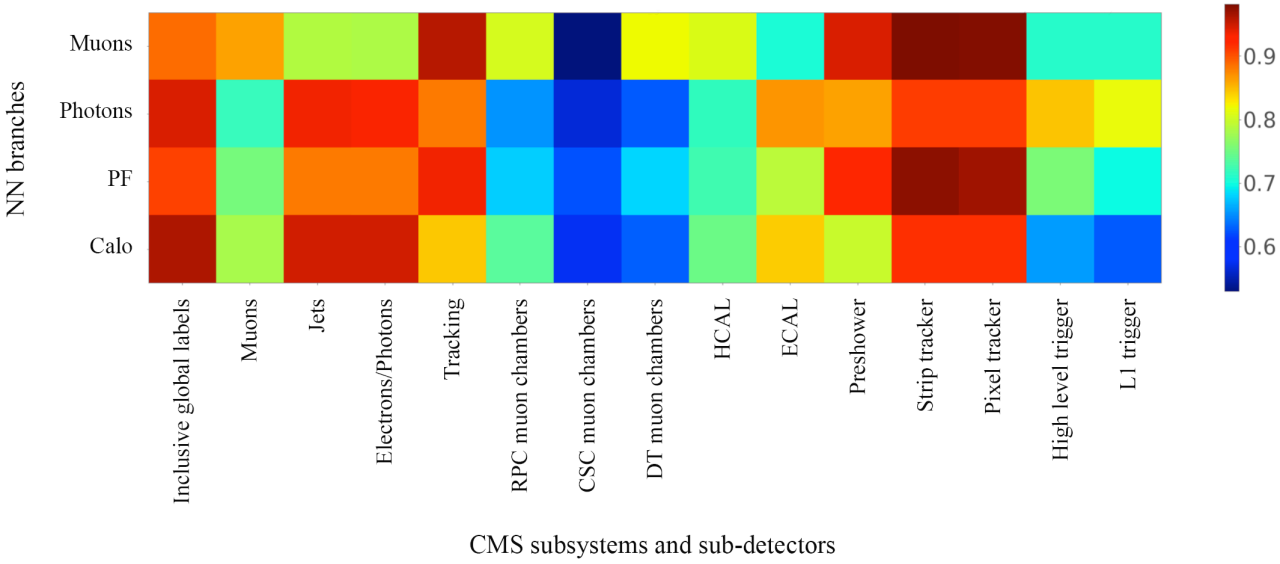


Рис. 10: Результаты предложенного метода на данных эксперимента CERN CMS: строки соответствуют каналам, столбцы — подсистемам установки, цвет обозначает значение ROC AUC метрики предсказаний соответствующей подсети против меток соответствующей подсистемы. Стоит заметить, что аномалии в подсистемах могут зависеть от наличия аномалий в других подсистемах и внешних событий.

согласуются с экспертной оценкой.

2.3. Ручная разметка данных

Методы, предложенные выше, полагаются на размеченные выборки для обучения и из-за высокой размерности обычно требуют больших обучающих выборок. Один из наиболее популярных способов получения размеченной выборки — ручная разметка экспертами. В научных экспериментах ручная разметка данных часто является нетривиальной и трудоемкой задачей. Например, в эксперименте CERN LHCb эксперту необходимо проверить несколько десятков гистограмм и графиков перед тем, как принимать решение относительно одного примера [33].

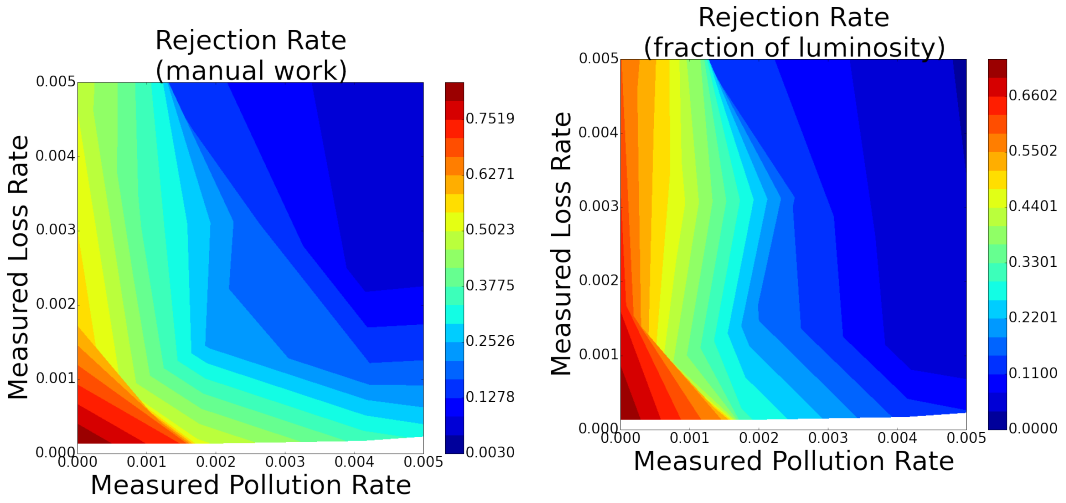
Алгоритм 3: Active learning system for manual labeling assistance.

Input: $L_0 \in \mathbb{R}, P_0 \in \mathbb{R}$ — constraints on loss and pollution rates
$\tau_L, \tau_P \leftarrow 0, 1$;
classifier $\leftarrow (x \mapsto 1/2)$;
$X, Y \leftarrow \varnothing, \varnothing$;
for $i = 1, \ldots, N$ **do**
  $x_i \leftarrow$ new sample;
  $p_i \leftarrow$ classifier(x_i);
  if $p_i > \tau_L$ **then**
    automatically label x_i as normal sample;
  else if $p_i < \tau_P$ **then**
    automatically label x_i as anomalious sample;
  else
    $y_i \leftarrow$ request expert label;
    $X, Y \leftarrow (X, x_i), (Y, y_i)$;
    compute predictions P on X with k-fold cross-validation;
    `// thresholds for acceptable loss and pollution rates`
    $\tau_L \leftarrow \max\{\tau \mid \hat{L}_\tau(P, Y) \leq L_0\}$;
    $\tau_P \leftarrow \min\{\tau \mid \hat{P}_\tau(P, Y) \leq P_0\}$;
    retrain classifier;
  end
end

Для решения этой проблемы мы рассматриваем активное обучение, основ-

ной целью которого является помощь экспертам путем принятия автоматических решений в случаях высокой уверенности в предсказаниях [75]. В случаях низкого уровня уверенности в автоматических предсказаниях, система делегирует решение эксперту, пример добавляется к обучающих выборке, и классификатор обучается заново. Достоверность в выборке определяется через оценку вероятности классов классификатором, пороговые значения для «безусловно нормальных» и «безусловно аномальных» определяются путем перекрестной проверки (англ. cross-validation) и сравнению результатов с внешними ограничениями на допустимый уровень загрязнения (англ. pollution rate) P_0 (доля аномальных примеров среди автоматически размеченных как «безусловно нормальные») и уровень потерь (англ. loss rate) L_0 (доля нормальных образцов, автоматически размеченных как «безусловно аномальных») [76]. Процедура описана в алгоритме 3.

Система была оценена на открытых данных эксперимента CERN CMS; результаты представлены на рисунке 11: стоит обратить внимание, что даже при самых жестких ограничениях (уровень загрязнения и уровень потерь менее 10^{-4}) система позволяет сэкономить не менее 20% ручного труда и даже небольшое ослабление ограничений (загрязнение и потери менее 10^{-3}) увеличивает процент сэкономленных усилий до более, чем 50%.



(a) Трудозатраты.

(b) Суммарная светимость примеров, переданных эксперту.

Рис. 11: Качество предложенного алгоритма на открытых данных эксперимента CERN CMS. Графики показывают пропорцию примеров, отправленных на ручную разметку (слева) и светимость этих примеров (справа) как функции от уровней потерь и загрязнения.

Обратите внимание, что система учится на данных, размеченных вручную, и постепенно заменяет экспертов; таким образом, производительность улучшается со временем, как показано на рисунке 12. На первых итерациях система запрашивает экспертные метки для большинства примеров; однако, по мере увеличения размера обучающей выборки, прогнозы системы становятся все более и более надежными, что отражается в постепенном уменьшении количества запросов к эксперту.

В заключение, данное исследование демонстрирует, что методы минимизации сбора данных могут значительно снизить затраты, связанные с ручной разметкой данных научных экспериментов. В свою очередь, это позволяет либо снизить затраты на обучение алгоритмов обнаружения аномалий, либо повысить качество этих алгоритмов, предоставляя больший объем данных для обучения.

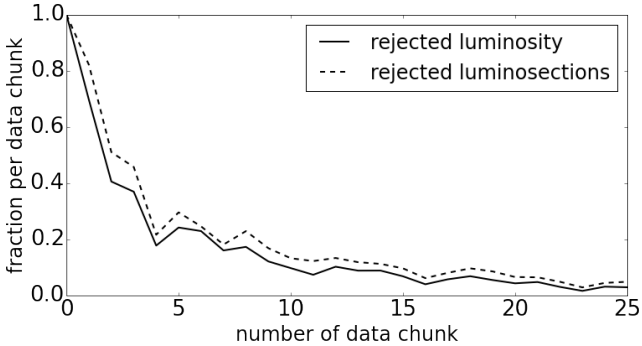


Рис. 12: Доля примеров (сплошная линия) и доля светимости (пунктирная линия), размеченных экспертом, для каждой итерации (data chunk).

2.4. Тонкая настройка компьютерных симуляций

Методы обнаружения аномалий основаны на предположении, что нормальная обучающая выборка велика. Поскольку в большинстве сложных экспериментов используются уникальные установки и, следовательно, уникальная структура данных, получение достаточно большой обучающей выборки является сложной задачей, в частности, из-за того, что контроль качества данных работает с необработанными или минимально обработанными данными (см., например, работу [14]).

Одним из основных источников обучающих данных для алгоритмов обнаружения аномалий является компьютерные симуляции (компьютерное моделиро-

вание). Многие научные эксперименты используют компьютерные симуляции. В некоторых областях, таких как физика высоких энергий, они играют важную роль в эксперименте. Например, генератор событий Pythia [100,101] широко используется в экспериментах CERN и моделирует результат протон-протонных столкновений. Другая компьютерная симуляция, GEANT [102], отвечает за симуляцию отклика детектора и часто используется в тандеме с генераторами событий. Как правило, генераторы событий имитируют результаты эксперимента при номинальных условиях, поскольку учет широкого диапазона возможных аномалий крайне затруднен. Тем не менее, смоделированные аномалии могут также учитываться алгоритмами детектирования аномалий, представленными выше.

Кроме того, компьютерные симуляции играют важную роль в поиске различий между теоретическими предсказаниями и наблюдениями, поскольку они фактически представляют теоретические модели [40].

Основная проблема, возникающая при попытке обучить алгоритмы машинного обучения на смоделированных данных, заключается в том, что большинство симуляций содержат параметры, значения которых точно не известны для конкретного эксперимента [55, 56]. Несоответствие между параметрами симуляции и настоящими значениями может привести к ухудшению качества предсказаний методов детектирования аномалий, обученных на таких данных. Такое несоответствие особенно проблематично, поскольку события, полученные плохо настроенной симуляцией, могут потенциально быть похожими на аномальное поведение или могут значительно затруднять анализ различий между теоретическими предсказаниями и наблюдениями [42].

Многие подходы пытаются обойти несоответствие между симуляцией и наблюдениями путем уменьшения влияния различий на модель. Такие подходы включают обучение переносу [103] (англ. transfer learning), обучение классификатора, статистически независимого от некоторых переменных [104], и использование контрольной переменной [105] (англ. control variable). Для задач детектирования аномалий некоторые методы неприменимы (например, контрольные переменные [105]); для других отсутствуют какие-либо гарантии, которые важны для обнаружения аномалий, например, обучение переносу [103] или обучение центральной статистики [104] (англ. learning to pivot).

Самый простой способ включить данные симуляции — найти такие значе-

ния параметров симуляции, чтобы распределение результатов моделирования точно соответствовало распределению, наблюдаемому в эксперименте. Этот процесс называется тонкой настройкой (англ. fine-tuning). Количество параметров симуляции обычно мало, например, в работе [56] авторы рассматривают около 20 параметров. Кроме того, тонкая настройка часто выполняется на обработанных данных, например, настройка генератора событий Pythia выполняется на примерно 400 признаках [55]. Таким образом, на практике для тонкой настройки требуется меньше реальных данных, чем для обучения алгоритма детектирования аномалий необработанных или минимально обработанных данных. В то же время, компьютерные симуляции являются источником априори нормальных примеров, которые не требуют ручной разметки.

Недавние исследования в области генеративных моделей, а именно порождающие состязательные сети [80] (англ. Generative Adversarial Networks), предоставляют процедуры тонкой настройки общего назначения. Одной из сложностей в применении состязательной оптимизации (англ. adversarial optimization) к поиску параметров симуляции является отсутствие градиентов симуляции по ее параметрам, так как симуляции включают генерацию множества случайных величин и не могут быть дифференцированы простым способом. Недавно опубликованный метод состязательной вариационной оптимизации [57] (англ. Adversarial Variational Optimization) решает эту проблему путем комбинирования алгоритма безградиентной оптимизации (вариационной оптимизации) и состязательного обучения, позволяя осуществлять тонкую настройку недифференцируемых симуляций. В данной работе любая процедура тонкой настройки, использующая состязательное обучение, называется состязательной оптимизацией (англ. adversarial optimization, AO).

AO минимизирует дивергенцию D между настоящим распределением P и распределением Q_θ, которое соответствует выходу симуляции с параметрами θ:

$$D(P, Q_\theta) \to_\theta \min . \tag{11}$$

AO использует так называемые состязательные дивергенции, т.е. дивергенции, которые могут быть представлены в виде оптимизационной задачи. Например, одна из самых популярных состязательных дивергенций, дивергенция Йенсена-Шенона (англ. Jensen-Shannon divergence):

$$\mathrm{JSD}(P, Q) = \log 2 - \min_{f \in \mathcal{F}} L(f, P, Q); \tag{12}$$

где $\mathcal{F}$ — множество всех функций вида $\mathcal{X} \to [0,1]$ и L — кросс-энтропийная функция потерь. Работа [85] фокусируется на дивергенции Йенсена-Шеннона; но все результаты могут быть применены к другим состязательным дивергенциям, например, расстоянию Вассерштейна.

Компьютерные симуляции сложных экспериментальных установок, как правило являются сложными в вычислительном отношении. Например, моделирование одного события столкновения протонов в детекторе CERN ATLAS занимает несколько минут на процессоре с одним ядром [58]. Из-за относительно высокой размерности, по крайней мере, высокой для методов безградиентной оптимизации, размеры выборок, требуемых для стандартных методов АО, велики, что приводит к большим вычислительным нагрузкам. Например, в работе [85] мы рассматриваем упрощенную версию реальной задачи тонкой настройки с одним параметром: состязательная байесовская оптимизация требует около $64 \cdot 10^3$ обращений к симуляции для уменьшения неопределенности настоящих параметров всего в 10 раз. Ожидается, что количество обращений к симуляции, необходимое для тонкой настройки, значительно выше в условиях с большим количеством оптимизируемых параметров.

В соответствующей работе [85] мы предлагаем новое семейство дивергенций, а именно адаптивные дивергенции, специально разработанные для уменьшения количества вызовов симуляции. Адаптивная дивергенция определяется на семействе псевдо-дивергенций.

Определение 1 *Функция $D : \Pi(\mathcal{X}) \times \Pi(\mathcal{X}) \to \mathbb{R}$ называется псевдо-дивергенцией, если:*

(P1) $\forall P, Q \in \Pi(\mathcal{X}) : D(P,Q) \geq 0$*;*

(P2) $\forall P, Q \in \Pi(\mathcal{X}) : (P = Q) \Rightarrow D(P,Q) = 0$*;*

где $\Pi(\mathcal{X})$ — множество всех распределений вероятности на $\mathcal{X}$.

Мы также накладываем ограничения на семейство псевдо-дивергенций.

Определение 2 *Семейство псевдо-дивергенций $\mathcal{D} = \{D_\alpha : \Pi(\mathcal{X}) \times \Pi(\mathcal{X}) \to \mathbb{R} \mid \alpha \in [0,1]\}$ упорядочено и полно относительно дивергенции Йенсена-Шеннона, если:*

(D0) *D_α — псевдо-дивергенция для всех $\alpha \in [0,1]$;*

(D1) $\forall P, Q \in \Pi(\mathcal{X}) : \forall 0 \leq \alpha_1 < \alpha_2 \leq 1 : D_{\alpha_1}(P, Q) \leq D_{\alpha_2}(P, Q)$;

(D2) $\forall P, Q \in \Pi(\mathcal{X}) : D_1(P, Q) = \mathrm{JSD}(P, Q)$.

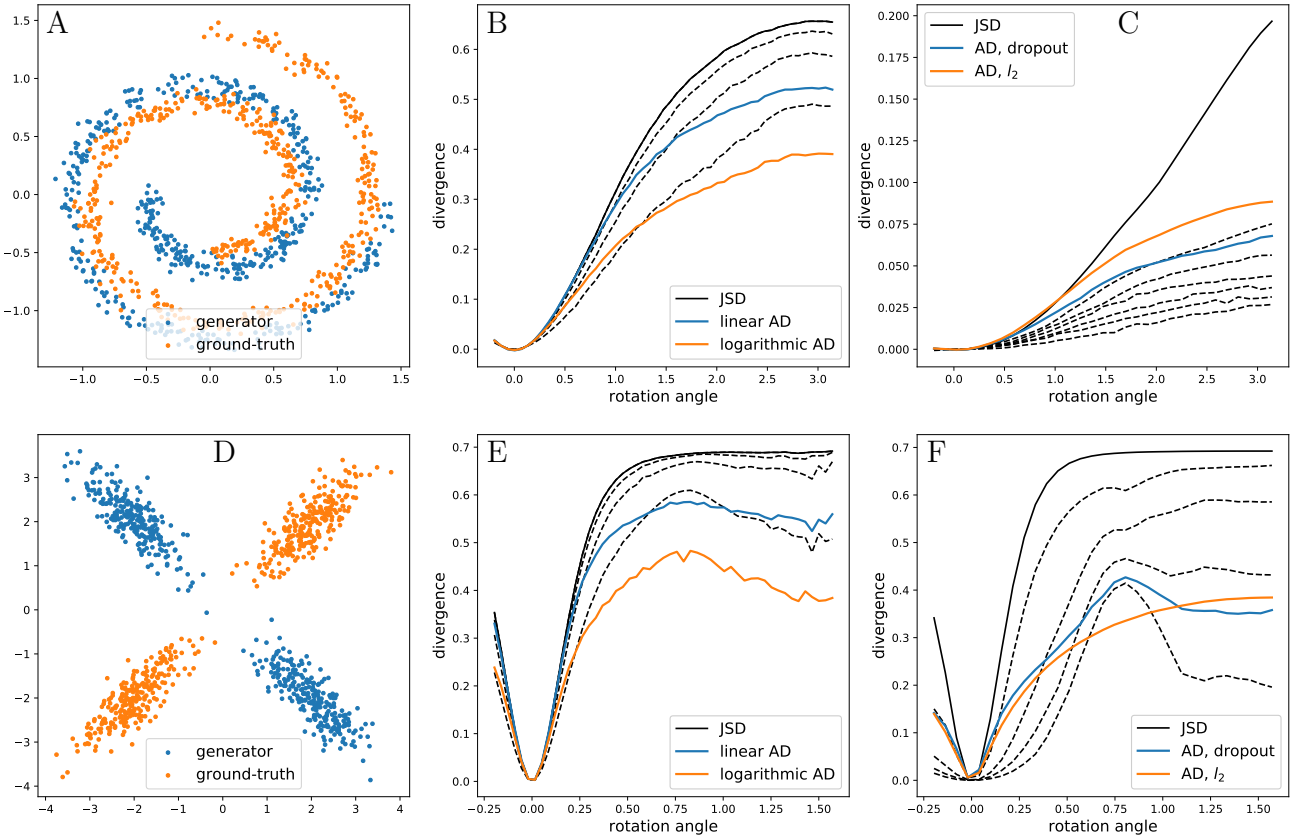


Рис. 13: Синтетические примеры. (A) and (D): настоящие распределения и примеры генераторов. Оба генератора являются повернутой версией настоящего распределения. (B) и (E): JSD — дивергенции Йенсена-Шеннона, оцененные с помощью ансамбля деревьев принятия решений; linear AD и logarithmic AD — адаптивные дивергенции, основанные на тех же моделях, что и JSD, с линейными (linear) и логарифмическими (logarithmic) функциями мощности. (C) and (F): JSD — дивергенции Йенсена-Шеннона и адаптивные дивергенции, оцененные с помощью полносвязных нейронных сетей.

Следующие определения вводят два типа семейств, упорядоченных и полных по отношению к дивергенции Йенсена-Шеннона.

Определение 3 *Семейство моделей* $\mathcal{M} = \{M_\alpha \subseteq \mathcal{F} \mid \alpha \in [0, 1]\}$ *полно и вложено, если:*

(N0) $(x \mapsto 1/2) \in M_0$;

(N1) $M_1 = \mathcal{F}$;

(N2) $\forall \alpha, \beta \in [0,1] : (\alpha < \beta) \Rightarrow (M_\alpha \subset M_\beta)$.

Теорема 3 *Если семейство моделей* $\mathcal{M} = \{M_\alpha \subseteq \mathcal{F} \mid \alpha \in [0,1]\}$ *полно и вложено, тогда семейство* $\mathcal{D} = \{D_\alpha : \Pi(\mathcal{X}) \times \Pi(\mathcal{X}) \to \mathbb{R} \mid \alpha \in [0,1]\}$*, где:*

$$D_\alpha(P,Q) = \log 2 - \inf_{f \in M_\alpha} L(f,P,Q), \tag{13}$$

является упорядоченным и полным относительно дивергенции Йенсена-Шеннона семейством псевдо-дивергенций.

Определение 4 *Для параметризованной модели* $M = \{f(\theta,\cdot) : \mathcal{X} \to [0,1] \mid \theta \in \Theta\}$*, функция* $R : \Theta \to \mathbb{R}$ *называется настоящей регуляризацией модели* M *если:*

(R1) $\forall \theta \in \Theta : R(\theta) \geq 0$*;*

(R2) $\exists \theta_0 \in \Theta : \left(f(\theta_0,\cdot) \equiv \frac{1}{2}\right) \wedge (R(\theta_0) = 0)$.

Теорема 4 *Для параметризованной модели* $M = \{f(\theta,\cdot) \mid \theta \in \Theta\}$*,* $M = \mathcal{F}$*, настоящей регуляризации* $R : \Theta \to \mathbb{R}$*, и строго возрастающей функции* $c : [0,1] \to [0,+\infty)$ *такой, что* $c(0) = 0$*, семейство* $\mathcal{D} = \{D_\alpha : \Pi(\mathcal{X}) \times \Pi(\mathcal{X}) \to \mathbb{R} \mid \alpha \in [0,1]\}$*:*

$$\begin{aligned}
D_\alpha(P,Q) &= \log 2 - \min_{\theta \in \Theta_\alpha(P,Q)} L(f(\theta,\cdot),P,Q); \\
\Theta_\alpha(P,Q) &= \operatorname*{Arg\,min}_{\theta \in \Theta} L_\alpha^R(\theta,P,Q); \\
L_\alpha^R(\theta,P,Q) &= L(f(\theta,\cdot),P,Q) + c(1-\alpha)R(\theta);
\end{aligned}$$

является упорядоченным и полным относительно дивергенции Йенсена-Шеннона семейством псевдо-дивергенций.

Определения 3 и 4 предоставляют практический способ конструирования упорядоченных и полных относительно дивергенции Йенсена-Шеннона семейств псевдо-дивергенций. Например, полное и вложенное семейство моделей может быть представлено в виде последовательности нейронных сетей.

В общем, наиболее значимые случаи упорядоченных и полных по отношению к дивергенции Йенсена-Шеннона семейств псевдо-дивергенций возникают при варьировании «мощности» базового классификатора f, где «мощность»

Алгоритм 4: Общая процедура вычисления адаптивной дивергенции поиском по сетке

Input: $\mathcal{D} = \{D_\alpha \mid \alpha \in [0,1]\}$ — ordered and complete w.r.t. Jensen-Shannon divergence family of pseudo-divergences;
ε — tolerance;
P, Q — input distributions.

$\alpha \leftarrow 0$;
while $D_\alpha(P,Q) < (1-\alpha)\log 2$ **do**
  $\alpha \leftarrow \alpha + \varepsilon$;
end
return $D_\alpha(P,Q)$

может означать количество узлов в нейронной сети или силу регуляризации, применяемой во время обучения. Таким образом, мы называем параметр α мощностью псевдо-дивергенции D_α относительно семейства $\mathcal{D}$ или просто мощностью псевдо-дивергенции, если семейство ясно из контекста. Важное свойство псевдо-дивергенций, определенных таким образом, заключается в том, что классификаторы с низкой мощностью, как правило, требуют небольшое количество примеров для обучения; следовательно, оценка псевдо-дивергенции, построенной на таких классификаторах, требует меньших выборок для оценки, чем псевдо-дивергенции с высокой мощностью. Следует отметить, что, хотя использование псевдо-дивергенции с низкой мощностью вместо настоящих дивергенций является привлекательным с точки зрения вычислительных ресурсов, их использование не гарантирует сходимости АО к настоящим параметрам из-за свойства (**P2**). В то же время, если псевдо-дивергенция $D(P,Q) > 0$ для некоторых P и Q, это автоматически означает, что $\mathrm{JSD}(P,Q) \geq D(P,Q) > 0$ и, следовательно, Q не является решением задачи тонкой настройки.

Адаптивная дивергенция использует тот факт, что некоторые параметры симуляции могут быть отвергнуты на основании вычислительно дешевых псевдо-дивергенций.

Определение 5 *Если семейство псевдо-дивергенций $\mathcal{D} = \{D_\alpha \mid \alpha \in [0,1]\}$ упорядочено и полно относительно дивергенции Йенсена-Шеннона, тогда адаптивная дивергенция $\mathrm{AD}_\mathcal{D}$, порожденная $\mathcal{D}$, определена как:*

$$\mathrm{AD}_\mathcal{D}(P,Q) = \inf\left\{D_\alpha(P,Q) \mid \mathrm{D}_\alpha(P,Q) \geq (1-\alpha)\log 2\right\}. \qquad (14)$$

Алгоритм 5: Boosted adaptive divergence

Input: X_P, X_Q — samples from distributions P and Q;
B — base estimator training algorithm;
N — maximal size of the ensemble;
$c : \mathbb{Z}_+ \to [0, 1]$ — capacity function;
ρ — learning rate;

```
F_0 ← 1/2;
i ← 0;
L_0 ← log 2;
for i = 1, ..., N do
    if L_i > c(i) log 2 then
        F_{i+1} ← F_i + ρ · B(F_i, X_P, X_Q);
        L_{i+1} ← L(F_{i+1}, X_P, X_Q);
        i ← i + 1;
    else
        return log 2 − L_i;
    end
end
return log 2 − L_N;
```

Следующая теорема, в сочетании с наблюдением, что $\mathrm{AD}(P, Q) \geq 0$, говорит о том, что адаптивная дивергенция является дивергенцией, а значит, гарантирует сходимость АО к таким параметрам симуляции, что распределение симуляции совпадает с реальными данными.

Теорема 5 *Если* $\mathrm{AD}_{\mathcal{D}}$ *— адаптивная дивергенция, порожденная упорядоченным и полным относительно дивергенции Йенсена-Шеннона семейством псевдо-дивергенций* $\mathcal{D}$*, тогда для любых* P *и* Q*:* $\mathrm{JSD}(P, Q) = 0$ *тогда и только тогда, когда* $\mathrm{AD}_{\mathcal{D}}(P, Q) = 0$.

Доказательство может быть найдено в соответствующей работе [85].

Алгоритм 4 демонстрирует общую процедуру вычисления адаптивной дивергенции методом поиска по сетке. Рисунок 13 демонстрирует поведение нескольких адаптивных дивергенций на искусственных данных.

Как можно видеть из определения, адаптивная дивергенция «переключается» между псевдо-дивергенциями в зависимости от распределений P и Q: ко-

Алгоритм 6: Adaptive divergence estimation by a regularized neural network

Input: X_P, X_Q — samples from distributions P and Q; $f_\theta : \mathcal{X} \to \mathbb{R}$ — neural network with parameters $\theta \in \Theta$; $R : \Theta \to \mathbb{R}$ — regularization function; c — capacity function; ρ — exponential average coefficient; β — coefficient for R_1 regularization; γ — learning rate of SGD.

$L_{\text{acc}} \leftarrow \log 2$

while *not converged* **do**

$x_P \leftarrow \text{sample}(X_P)$;

$x_Q \leftarrow \text{sample}(X_Q)$;

$\zeta \leftarrow c\left(1 - \frac{L_{\text{acc}}}{\log 2}\right)$;

$g_0 \leftarrow \nabla_\theta \left[L(f_\theta, x_P, x_Q) + \zeta \cdot R(f_\theta)\right]$;

$g_1 \leftarrow \nabla_\theta \|\nabla_\theta f_\theta(x_P)\|^2$;

$L_{\text{acc}} \leftarrow \rho \cdot L_{\text{acc}} + (1 - \rho) \cdot L(f_\theta, x_P, x_Q)$;

$\theta \leftarrow \theta - \gamma\,(g_0 + \beta g_1)$;

end

return $\log 2 - L(f_\theta, X_P, X_Q)$

гда P и Q далеки друг от друга, $\text{AD}_{\mathcal{D}}$ выбирает псевдо-дивергенции с низкой мощностью; когда Q приближается к P, адаптивная дивергенция использует псевдо-дивергенции с высокой мощностью, настоящая дивергенция $D_1 = \text{JSD}$ используется только для «доказательства» равенства распределений. Это свойство позволяет адаптивной дивергенции уменьшить количество обращений к симуляции, когда симуляция значительно отклоняется от реальных данных, не жертвуя свойствами сходимости АО.

Кроме того, предложены вычислительно эффективные процедуры для оценки адаптивной дивергенции в следующих случаях:

- семейство псевдо-дивергенций, удовлетворяющее определению 3, построенное на алгоритме бустинга (англ. boosting) — алгоритм 5;
- семейство псевдо-дивергенций, удовлетворяющее определению 4, построенное на нейронных сетях и настоящей регуляризации — алгоритм 6.

Предложенные алгоритмы были оценены на одном искусственном примере

и двух реалистичных задачах тонкой настройки, включая настройку генератора событий Pythia [100, 101], с использованием двух алгоритмов безградиентной оптимизации, а именно: байесовской оптимизации с гауссовскими процессами [106, 107] и состязательной вариационной оптимизации [57]. Основные результаты представлены на рисунках 14 и 15; дополнительные рисунки можно найти в оригинальной статье [85]. Как видно из рисунков, методы безградиентной оптимизации с адаптивной дивергенцией находят решения примерно на порядок ближе к истинностным в рамках того же бюджета на количество обращений к симуляции.

В заключение, адаптивная дивергенция уменьшает вычислительную нагрузку, связанную с тонкой настройкой, что, в свою очередь, позволяет значительно уменьшить любое несоответствие между наблюдениями и симуляцией. Стоит обратить внимание, что такое несоответствие напрямую приводит к смещению алгоритмов обнаружения аномалий, обученных на данных симуляции, и усложняет поиск разногласий между теорией и наблюдениями. Поэтому процедуры тонкой настройки, в которых используются адаптивные дивергенции, напрямую влияют на общую производительность систем контроля качества данных.

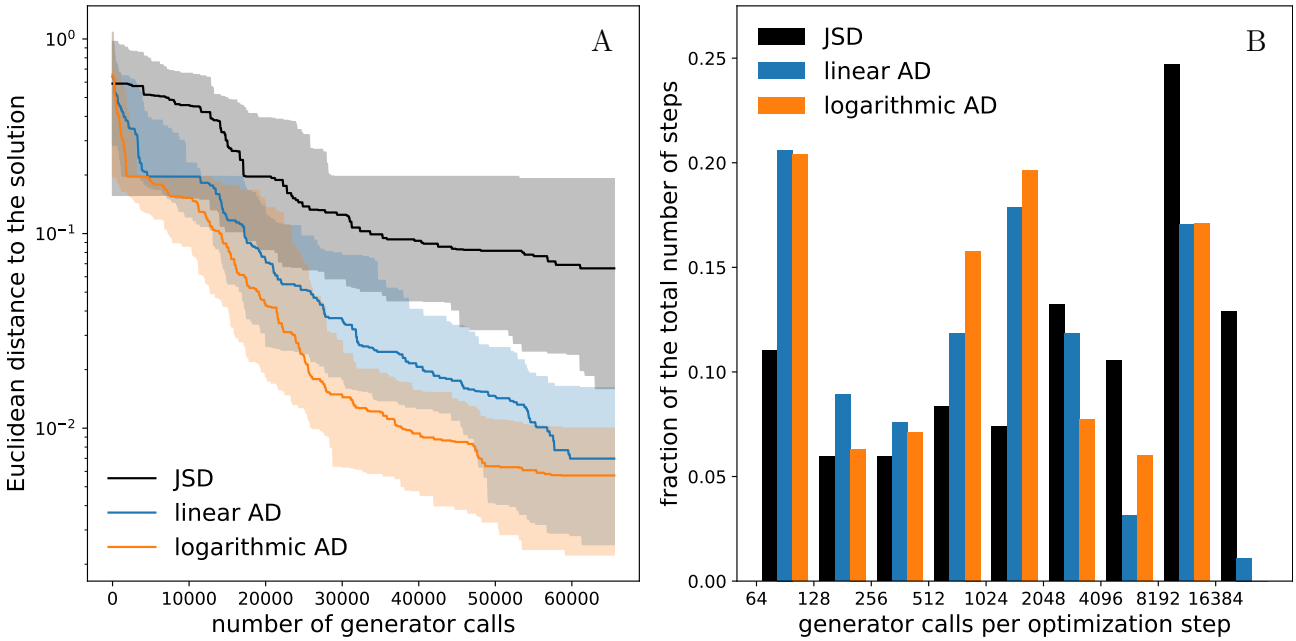


Рис. 14: Настройка параметров Pythia, CatBoost. (A) сходимость байесовской оптимизации на: дивергенции Йенсена-Шеннона (JSD), адаптивные дивергенции с линейной функцией мощности (linear AD) и логарифмической функцией мощности (logarithmic AD). Каждый эксперимент был повторен 100 раз; кривые интерполированы, медианы показаны сплошными линиями, полосы соответствуют 25-ой и 75-ой перцентилям. (B) Распределение вычислительных ресурсов на один шаг оптимизации.

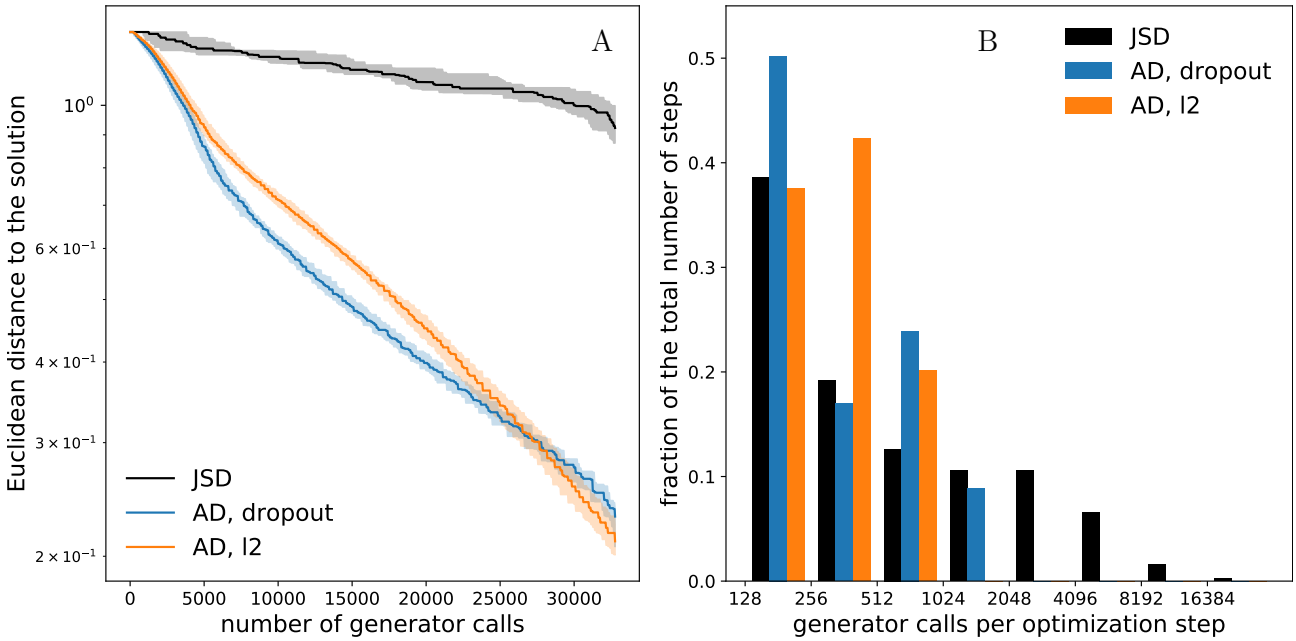


Рис. 15: Выравнивание Pythia, нейронные сети. (A) Сходимость состязательной вариационной оптимизации на: адаптивных дивергенциях, порожденных l_2 регуляризацией (AD, l_2) и случайным выбыванием (AD, dropout), и соответствующая базовая дивергенция с R_1 регуляризацией (JSD). Каждый эксперимент быд повторен 20 раз; кривые интерполированы, медианы показаны сплошными линиями, полосы соответствуют 25-ой и 75-ой перцентилям. (B) Распределение вычислительных ресурсов на один шаг оптимизации.

3. Заключение

Контроль качества данных и методы детектирования аномалий играют важную роль в научных экспериментах. Подходы машинного обучения к контролю качества данных и детектированию аномалий становятся все более и более востребованными с увеличением сложности экспериментов и повышением точности теоретических моделей.

В этой диссертации автор решает основные задачи, стоящие за контролем качества данных и поиском расхождений между теоретическими предсказаниями и наблюдениями. Во-первых, были рассмотрены алгоритмы обнаружения аномалий, автор расширил традиционную постановку задачи детектирования аномалий и:

1. предложил новое семейство методов детектирования аномалий, а именно алгоритмы OPE и EOPE классификации, способные учитывать известные примеры аномалий, таким образом охватывая весь спектр проблем между одноклассовыми и двухклассовыми задачами классификации; доказал основные свойства этих методов; продемонстрировал качество на многих эталонных задачах, в том числе из физики частиц.

Во-вторых, для того, чтобы увеличить возможности систем контроля качества данных, автор:

2. предложил новый метод для определения каналов, затронутых аномалией, который не требует дополнительных меток для обучения; доказал основные свойства метода; протестировал предложенный алгоритм на данных, собранных экспериментом CERN CMS.

В-третьих, для сбора обучающих выборок для алгоритмов детектирования аномалий и обеспечения возможности поиска различий между теорией и наблюдениями были рассмотрены два потенциальных источника размеченных данных, а именно ручная разметка и компьютерные симуляции. В результате автор:

3. протестировал алгоритм, основанный на активном обучении, для данных, собранных детектором эксперимента CERN CMS; продемонстрировал преимущества подхода в условиях ККД;

4. предложил новое семейство дивергенций, а именно адаптивные дивергенции, которые позволяют значительно ускорить тонкую настройку компьютерных симуляций; доказал основные свойства адаптивных дивергенций; оценил качество на реалистичных задачах тонкой настройки.

Кроме того, все методы, предложенные в данной диссертации, применимы за пределами научных экспериментов, в частности, методы OPE и EOPE классификации являются универсальными и могут быть применены для любых проблем детектирования аномалий; метод определения источников аномалий основан на общих предположениях и может применяться в промышленных условиях; адаптивные дивергенции не ограничиваются процедурами тонкой настройки и могут быть использованы в любых задачах состязательного обучения.

Список литературы

[1] The CMS experiment at the CERN LHC / The CMS Collaboration, S Chatrchyan, G Hmayakyan et al. // Journal of Instrumentation. — 2008. — aug. — Vol. 3, no. 08. — P. S08004–S08004.

[2] The CMS trigger system / V. Khachatryan, A.M. Sirunyan, A. Tumasyan et al. // Journal of Instrumentation. — 2017. — jan. — Vol. 12, no. 01. — P. P01020–P01020.

[3] The square kilometre array / Peter E Dewdney, Peter J Hall, Richard T Schilizzi, T Joseph LW Lazio // Proceedings of the IEEE. — 2009. — Vol. 97, no. 8. — P. 1482–1496.

[4] Broekema P Chris, van Nieuwpoort Rob V, Bal Henri E. The Square Kilometre Array science data processor. Preliminary compute platform design // Journal of Instrumentation. — 2015. — Vol. 10, no. 07. — P. C07004.

[5] An End-to-End Computing Model for the Square Kilometre Array / R. Jongerius, S. Wijnholds, R. Nijboer, H. Corporaal // Computer. — 2014. — Sep. — Vol. 47, no. 9. — P. 48–54.

[6] Neural networks and cellular automata in experimental high energy physics : Rep. / Paris-11 Univ. ; Executor: B Denby : 1987.

[7] Baldi Pierre, Sadowski Peter, Whiteson Daniel. Searching for exotic particles in high-energy physics with deep learning // Nature communications. — 2014. — Vol. 5, no. 1. — P. 1–9.

[8] Machine learning at the energy and intensity frontiers of particle physics / Alexander Radovic, Mike Williams, David Rousseau et al. // Nature. — 2018. — Vol. 560, no. 7716. — P. 41–48.

[9] Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science / Michael Zevin, Scott Coughlin, Sara Bahaadini et al. // Classical and Quantum Gravity. — 2017. — Vol. 34, no. 6. — P. 064003.

[10] Modern machine learning methods in HEP / Raphael Friese, Guenter Quast, Roger Wolf, Stefan Wunsch // Verhandlungen der Deutschen Physikalischen Gesellschaft. — 2017.

[11] Carrazza Stefano. Machine learning challenges in theoretical HEP // Journal of Physics: Conference Series / IOP Publishing. — Vol. 1085. — 2018. — P. 022003.

[12] Machine-learning in astronomy / Michael Hobson, Philip Graff, Farhan Feroz, Anthony Lasenby // Proceedings of the International Astronomical Union. — 2014. — Vol. 10, no. S306. — P. 279–287.

[13] LHCb reoptimized detector design and performance: Technical Design Report : Rep. / LHCb-TDR-009 ; Executor: S Cadeddu, P Dalpiaz, Z Guzik et al. : 2003.

[14] Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider / Adrian Alan Pol, Gianluca Cerminara, Cécile Germain et al. // Computing and Software for Big Science. — 2019. — Vol. 3, no. 1. — P. 3.

[15] Online data monitoring in the LHCb experiment / O Callot, S Cherukuwada, M Frank et al. // Journal of Physics: Conference Series / IOP Publishing. — Vol. 119. — 2008. — P. 022015.

[16] LIGO: the laser interferometer gravitational-wave observatory / BP Abbott, R Abbott, R Adhikari et al. // Reports on Progress in Physics. — 2009. — Vol. 72, no. 7. — P. 076901.

[17] Impact of aerosols and adverse atmospheric conditions on the data quality for spectral analysis of the HESS telescopes / Joachim Hahn, R De los Reyes, Konrad Bernlöhr et al. // Astroparticle Physics. — 2014. — Vol. 54. — P. 25–32.

[18] Mommert Michael. Cloud Identification from All-sky Camera Data with Machine Learning // The Astronomical Journal. — 2020. — mar. — Vol. 159, no. 4. — P. 178.

[19] CMS data quality monitoring: systems and experiences / Lassi Tuura, A Meyer, I Segoni, G Della Ricca // Journal of Physics: Conference Series. — 2010. — Vol. 219, no. 7. — P. 072020.

[20] The OPERA experiment in the CERN to Gran Sasso neutrino beam / R Acquafredda, T Adam, N Agafonova et al. // Journal of Instrumentation. — 2009. — Vol. 4, no. 04. — P. P04018.

[21] Brumfiel Geoff. Particles break light-speed limit. — 2011.

[22] Measurement of the neutrino velocity with the OPERA detector in the CNGS beam / T Adam, N Agafonova, A Aleksandrov et al. // Journal of High Energy Physics. — 2012. — Vol. 2012, no. 10. — P. 93.

[23] Stone Robert, Mukherjee Soma. Environmentally induced nonstationarity in LIGO science run data // Classical and Quantum Gravity. — 2009. — oct. — Vol. 26, no. 20. — P. 204021.

[24] George Daniel, Shen Hongyu, Huerta EA. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning // Physical Review D. — 2018. — Vol. 97, no. 10. — P. 101501.

[25] Li W., Wu G., Du Q. Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery // IEEE Geoscience and Remote Sensing Letters. — 2017. — Vol. 14, no. 5. — P. 597–601.

[26] Mimicking the human expert: Pattern recognition for an automated assessment of data quality in MR spectroscopic images / Bjoern H. Menze, B. Michael Kelm, Marc-André Weber et al. // Magnetic Resonance in Medicine. — 2008. — Vol. 59, no. 6. — P. 1457–1466.

[27] Identifying, attributing, and overcoming common data quality issues of manned station observations / Stefan Hunziker, Stefanie Gubler, Juan Calle et al. // International Journal of Climatology. — 2017. — Vol. 37, no. 11. — P. 4131–4145.

[28] Stankevicius Mantas, Marcinkevicius Virginijus, Rapsevicius Valdas. Comparison of Supervised Machine Learning Techniques for CERN CMS Of-

fline Data Certification. // Doctoral Consortium/Forum@ DB&IS. — 2018. — P. 170–176.

[29] Using Artificial Neural Networks for Glitch Identification in Advanced LIGO / Donald Moffa, Kyle Rose, Les Wade et al. // APS Meeting Abstracts. — 2018.

[30] Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment / Adrian Alan Pol, Virginia Azzolini, Gianluca Cerminara et al. // EPJ Web of Conferences / EDP Sciences. — Vol. 214. — 2019. — P. 06008.

[31] ATLAS online data quality monitoring / C. Cuenca Almenar, A. Corso-Radu, H. Hadavand et al. // 2010 17th IEEE-NPSS Real Time Conference. — 2010. — P. 1–5.

[32] The ALICE online data quality monitoring / Barthelemy von Haller, Adriana Telesca, Sylvain CHAPELAND et al. // 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research / SISSA Medialab. — Vol. 93. — 2011. — P. 024.

[33] IOP: LHCb data quality monitoring / M Adinolfi, A Ustyuzhanin, D Derkach et al. // J. Phys.: Conf. Ser. — Vol. 898. — 2017. — P. 092027.

[34] Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC / G. Aad, T. Abajyan, B. Abbott et al. // Physics Letters B. — 2012. — Vol. 716, no. 1. — P. 1 – 29.

[35] Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC / G. Aad, T. Abajyan, B. Abbott et al. // Physics Letters B. — 2012. — Vol. 716, no. 1. — P. 1 – 29.

[36] GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence / B. P. Abbott, R. Abbott, T. D. Abbott et al. // Phys. Rev. Lett. — 2017. — Oct. — Vol. 119. — P. 141101.

[37] Search for new physics with atoms and molecules / MS Safronova, D Budker, D DeMille et al. // Reviews of Modern Physics. — 2018. — Vol. 90. — P. 025008.

[38] Farina Marco, Nakai Yuichiro, Shih David. Searching for new physics with deep autoencoders // Phys. Rev. D. — 2020. — Apr. — Vol. 101. — P. 075021.

[39] Semi-supervised anomaly detection – towards model-independent searches of new physics / Mikael Kuusela, Tommi Vatanen, Eric Malmi et al. // Journal of Physics: Conference Series. — 2012. — jun. — Vol. 368. — P. 012032.

[40] Andreassen Anders, Nachman Benjamin, Shih David. Simulation assisted likelihood-free anomaly detection // Phys. Rev. D. — 2020. — May. — Vol. 101. — P. 095004.

[41] De Simone Andrea, Jacques Thomas. Guiding new physics searches with unsupervised learning // The European Physical Journal C. — 2019. — Vol. 79. — P. 1–15.

[42] Blance Andrew, Spannowsky Michael, Waite Philip. Adversarially-trained autoencoders for robust unsupervised new physics searches // Journal of High Energy Physics. — 2019. — Vol. 2019. — P. 47.

[43] Novelty detection meets collider physics / Jan Hajer, Ying-Ying Li, Tao Liu, He Wang // Phys. Rev. D. — 2020. — Apr. — Vol. 101. — P. 076015.

[44] Hodge Victoria, Austin Jim. A survey of outlier detection methodologies // Artificial intelligence review. — 2004. — Vol. 22. — P. 85–126.

[45] Chandola Varun, Banerjee Arindam, Kumar Vipin. Anomaly Detection: A Survey // ACM Comput. Surv. — 2009. — Jul. — Vol. 41, no. 3. — Access mode: `https://doi.org/10.1145/1541880.1541882`.

[46] LHCb VELO Upgrade Technical Design Report : Rep. : CERN-LHCC-2013-021. LHCB-TDR-013 ; Executor: LHCb Collaboration : 2013. — Nov.

[47] Towards automation of data quality system for CERN CMS experiment / M Borisyak, F Ratnikov, D Derkach, A Ustyuzhanin // Journal of Physics: Conference Series. — 2017. — oct. — Vol. 898. — P. 092041.

[48] (1 + epsilon)-class Classification: an Anomaly Detection Method for Highly Imbalanced or Incomplete Data Sets / Maxim Borisyak, Artem Ryzhikov, Andrey Ustyuzhanin et al. // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 72. — P. 1–22.

[49] Software for the LHCb experiment / Gloria Corti, Marco Cattaneo, Philippe Charpentier et al. // IEEE transactions on nuclear science. — 2006. — Vol. 53, no. 3. — P. 1323–1328.

[50] Petascale High Order Dynamic Rupture Earthquake Simulations on Heterogeneous Supercomputers / A. Heinecke, A. Breuer, S. Rettenberger et al. // SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. — 2014. — P. 3–14.

[51] TURBULENT MAGNETIC FIELD AMPLIFICATION FROM SPIRAL SASI MODES: IMPLICATIONS FOR CORE-COLLAPSE SUPERNOVAE AND PROTO-NEUTRON STAR MAGNETIZATION / Eirik Endeve, Christian Y. Cardall, Reuben D. Budiardja et al. // The Astrophysical Journal. — 2012. — may. — Vol. 751, no. 1. — P. 26.

[52] Using galaxy formation simulations to optimize LIGO follow-up observations / Elisa Antolini, Ilaria Caiazzo, Romeel Davé, Jeremy S Heyl // Monthly Notices of the Royal Astronomical Society. — 2017. — Vol. 466, no. 2. — P. 2212–2216.

[53] Perdikaris Paris, Grinberg Leopold, Karniadakis George Em. Multiscale modeling and simulation of brain blood flow // Physics of Fluids. — 2016. — Vol. 28, no. 2. — P. 021304.

[54] Agent-based simulation of a financial market / Marco Raberto, Silvano Cincotti, Sergio M. Focardi, Michele Marchesi // Physica A: Statistical Mechanics and its Applications. — 2001. — Vol. 299, no. 1. — P. 319 – 327. — Application of Physics in Economic Modelling.

[55] Skands Peter, Carrazza Stefano, Rojo Juan. Tuning PYTHIA 8.1: the Monash 2013 tune // The European Physical Journal C. — 2014. — Vol. 74, no. 8. — P. 3024.

[56] Ilten P., Williams M., Yang Y. Event generator tuning using Bayesian optimization // Journal of Instrumentation. — 2017. — apr. — Vol. 12, no. 04. — P. P04028.

[57] Louppe Gilles, Hermans Joeri, Cranmer Kyle. Adversarial Variational Optimization of Non-Differentiable Simulators // Proceedings of Machine Learning Research / Ed. by Kamalika Chaudhuri, Masashi Sugiyama. — Vol. 89 of Proceedings of Machine Learning Research. — PMLR, 2019. — 16–18 Apr. — P. 1438–1447.

[58] The ATLAS Collaboration. The ATLAS simulation infrastructure // European Physical Journal C: Particles and Fields. — 2010. — Vol. 70, no. 3. — P. 823–874.

[59] Chalapathy Raghavendra, Menon Aditya Krishna, Chawla Sanjay. Anomaly detection using one-class neural networks // arXiv preprint arXiv:1802.06360. — 2018.

[60] Deep one-class classification / Lukas Ruff, Nico Görnitz, Lucas Deecke et al. // International Conference on Machine Learning. — Stockholm, Sweden, 2018. — P. 4390–4399.

[61] Support vector method for novelty detection / Bernhard Schölkopf, Robert C Williamson, Alex J Smola et al. // Advances in Neural Information Processing Systems. — Denver, United States, 2000. — P. 582–588.

[62] Zhou Chong, Paffenroth Randy C. Anomaly detection with robust deep autoencoders // ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — Halifax, Canada, 2017. — P. 665–674.

[63] Chalapathy Raghavendra, Menon Aditya Krishna, Chawla Sanjay. Robust, deep and inductive anomaly detection // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. — Skopje, Macedonia, 2017. — P. 36–51.

[64] An Jinwon, Cho Sungzoon. Variational autoencoder based anomaly detection using reconstruction probability // Special Lecture on IE. — 2015. — Vol. 2, no. 1. — P. 1–18.

[65] Learning sparse representation with variational auto-encoder for anomaly detection / Jiayu Sun, Xinzhou Wang, Naixue Xiong, Jie Shao // IEEE Access. — 2018. — Vol. 6. — P. 33353–33361.

[66] Choi Hyunsun, Jang Eric, Alemi Alexander A. Waic, but why? generative ensembles for robust anomaly detection // arXiv preprint arXiv:1810.01392. — 2018.

[67] Tax David MJ, Duin Robert PW. Support vector data description // Machine learning. — 2004. — Vol. 54, no. 1. — P. 45–66.

[68] Liu Fei Tony, Ting Kai Ming, Zhou Zhi-Hua. Isolation forest // IEEE International Conference on Data Mining. — Washington, DC, United State, 2008. — P. 413–422.

[69] Baldi Pierre, Sadowski Peter, Whiteson Daniel. Searching for exotic particles in high-energy physics with deep learning // Nature communications. — 2014. — Vol. 5. — P. 4308.

[70] Bekker Jessa, Davis Jesse. Learning From Positive and Unlabeled Data: A Survey // CoRR. — 2018. — Vol. abs/1811.04820. — Access mode: `http://arxiv.org/abs/1811.04820`.

[71] Elkan Charles, Noto Keith. Learning Classifiers from Only Positive and Unlabeled Data // ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD ’08. — Las Vegas, Nevada, USA, 2008. — P. 213–220.

[72] Northcutt Curtis G., Wu Tailin, Chuang Isaac L. Learning with Confident Examples: rank Pruning for Robust Classification with Noisy Labels // Conference on Uncertainty in Artificial Intelligence, UAI. — Sydney, Australia, 2017.

[73] Pearl Judea. Causal inference in statistics: An overview // Statist. Surv. — 2009. — Vol. 3. — P. 96–146.

[74] Lughofer Edwin. On-line active learning: a new paradigm to improve practical useability of data stream modeling methods // Information Sciences. — 2017. — Vol. 415. — P. 356–376.

[75] RayChaudhuri T., Hamey L. G. C. Minimisation of data collection by active learning // Proceedings of ICNN’95 - International Conference on Neural Networks. — Vol. 3. — 1995. — P. 1338–1341.

[76] Burbidge Robert, Rowland Jem J., King Ross D. Active Learning for Regression Based on Query by Committee // Intelligent Data Engineering and Automated Learning - IDEAL 2007 / Ed. by Hujun Yin, Peter Tino, Emilio Corchado et al. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2007. — P. 209–218.

[77] Lughofer Edwin. Single-pass active learning with conflict and ignorance // Evolving Systems. — 2012. — Vol. 3. — P. 251–271.

[78] Lughofer Edwin. Evolving fuzzy systems-methodologies, advanced concepts and applications. — Springer, 2011. — Vol. 53.

[79] Toni Tina, Stumpf Michael PH. Simulation-based model selection for dynamical systems in systems and population biology // Bioinformatics. — 2009. — Vol. 26, no. 1. — P. 104–110.

[80] Generative adversarial nets / Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza et al. // Advances in neural information processing systems. — 2014. — P. 2672–2680.

[81] Meeds Edward, Leenders Robert, Welling Max. Hamiltonian ABC // Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence. — UAI'15. — Arlington, Virginia, USA : AUAI Press, 2015. — P. 582–591.

[82] Cranmer Kyle, Pavez Juan, Louppe Gilles. Approximating likelihood ratios with calibrated discriminative classifiers // arXiv preprint arXiv:1506.02169. — 2015.

[83] Experiments using machine learning to approximate likelihood ratios for mixture models / K Cranmer, J Pavez, Gilles Louppe, WK Brooks // Journal of Physics Conference Series. — 2016.

[84] Tran Minh-Ngoc, Nott David J, Kohn Robert. Variational Bayes with intractable likelihood // Journal of Computational and Graphical Statistics. — 2017. — Vol. 26, no. 4. — P. 873–882.

[85] Borisyak Maxim, Gaintseva Tatiana, Ustyuzhanin Andrey. Adaptive divergence for rapid adversarial optimization // PeerJ Computer Science. — 2020. — May. — Vol. 6. — P. e274.

[86] Jin Long, Lazarow Justin, Tu Zhuowen. Introspective classification with convolutional nets // Advances in Neural Information Processing Systems. — Long Beach, California, United States, 2017. — P. 823–833.

[87] Progressive growing of gans for improved quality, stability, and variation / Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen // arXiv preprint arXiv:1710.10196. — 2017.

[88] StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation / Yunjey Choi, Minje Choi, Munyoung Kim et al. // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition / IEEE. — 2018. — P. 8789–8797.

[89] Unrolled Generative Adversarial Networks / Luke Metz, Ben Poole, David Pfau, Jascha Sohl-Dickstein // ICLR. — 2016.

[90] Millman K. J., Aivazis M. Python for Scientists and Engineers // Computing in Science Engineering. — 2011. — Vol. 13, no. 2. — P. 9–12.

[91] van der Walt S., Colbert S. C., Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation // Computing in Science Engineering. — 2011. — Vol. 13, no. 2. — P. 22–30.

[92] SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python / Pauli Virtanen, Ralf Gommers, Travis E. Oliphant et al. // Nature Methods. — 2020. — Vol. 17. — P. 261–272.

[93] Scikit-learn: Machine Learning in Python / Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort et al. // Journal of Machine Learning Research. — 2011. — Vol. 12, no. 85. — P. 2825–2830.

[94] Abadi Martín, Agarwal Ashish, Barham Paul et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. — 2015. — Software available from tensorflow.org. Access mode: `https://www.tensorflow.org/`.

[95] PyTorch: An Imperative Style, High-Performance Deep Learning Library / Adam Paszke, Sam Gross, Francisco Massa et al. // Advances in Neural Information Processing Systems 32 / Ed. by H. Wallach, H. Larochelle, A. Beygelzimer et al. — Curran Associates, Inc., 2019. — P. 8026–8037.

[96] Scholkopf Bernhard, Smola Alexander J. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. — MIT press, 2001.

[97] Boser Bernhard E., Guyon Isabelle M., Vapnik Vladimir N. A Training Algorithm for Optimal Margin Classifiers // Proceedings of the Fifth Annual Workshop on Computational Learning Theory. — COLT '92. — New York, NY, USA : Association for Computing Machinery, 1992. — P. 144–152. — Access mode: https://doi.org/10.1145/130385.130401.

[98] Kim Taesup, Bengio Yoshua. Deep Directed Generative Models with Energy-Based Probability Estimation // arXiv preprint arXiv:1606.03439. — 2016.

[99] Deep learning for inferring cause of data anomalies / V. Azzolini, M. Borisyak, G. Cerminara et al. // Journal of Physics: Conference Series. — 2018. — sep. — Vol. 1085. — P. 042015.

[100] Sjöstrand Torbjörn, Mrenna Stephen, Skands Peter. PYTHIA 6.4 physics and manual // Journal of High Energy Physics. — 2006. — may. — Vol. 2006, no. 05. — P. 026.

[101] An introduction to PYTHIA 8.2 / Torbjörn Sjöstrand, Stefan Ask, Jesper R Christiansen et al. // Computer physics communications. — 2015. — Vol. 191. — P. 159–177.

[102] Recent developments in Geant4 / J. Allison, K. Amako, J. Apostolakis et al. // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. — 2016. — Vol. 835. — P. 186 – 225.

[103] Ganin Yaroslav, Lempitsky Victor. Unsupervised Domain Adaptation by Backpropagation // Proceedings of the 32nd International Conference on Machine Learning / Ed. by Francis Bach, David Blei. — Vol. 37 of Proceedings of Machine Learning Research. — Lille, France : PMLR, 2015. — 07–09 Jul. — P. 1180–1189. — Access mode: http://proceedings.mlr.press/v37/ganin15.html.

[104] Louppe Gilles, Kagan Michael, Cranmer Kyle. Learning to pivot with adversarial networks // Advances in neural information processing systems. — 2017. — P. 981–990.

[105] Borisyak M., Kazeev N. Machine Learning on data with sPlot background subtraction // Journal of Instrumentation. — 2019. — aug. — Vol. 14, no. 08. — P. P08020–P08020.

[106] Lizotte Daniel James. Practical bayesian optimization. — University of Alberta, 2008.

[107] Rasmussen Carl Edward. Gaussian processes in machine learning // Summer School on Machine Learning / Springer. — 2003. — P. 63–71.