

ЯДЕРНАЯ ФИЗИКА,
ЭЛЕМЕНТАРНЫЕ ЧАСТИЦЫ И ФИЗИКА ВЫСОКИХ ЭНЕРГИЙ

УДК 53.072

ОПТИМИЗАЦИЯ ОТБОРА РЕДКИХ СОБЫТИЙ
В ПРИСУТСТВИИ НЕОПРЕДЕЛЕННОГО
ПРЕОБЛАДАЮЩЕГО ФОНА

Р. А. Шоркин

Представлены результаты исследования влияния вклада малой доли сигнальных событий в фоновую выборку, используемую при обучении и оптимизации отбора на основе многомерного классификатора. Эффективность BDT (Boosted Decision Tree) классификаторов, обученных с применением фоновых выборок с включением неверно меченых сигнальных событий, сравнивалась с результатами BDT, обученных на чистых наборах данных. Продemonстрирована слабая зависимость мощности классификатора от абсолютного числа неправильно меченых сигнальных событий для данного относительного смещения фоновых и сигнальных событий (1000:1). Обнаружена обратная зависимость результатов оптимизации от возрастающего вклада неверно меченого сигнала.

Ключевые слова: анализ и статистические методы, методы анализа данных, применение деревьев решений, поиск гипотетических резонансов.

Введение. Исследования в области физики высоких энергий включают в себя процедуру эффективного отбора интересующих “сигнальных” событий из выборки данных, насыщенных сопутствующими “фоновыми” вкладами. Во множестве случаев использование многомерного классификатора позволяет создать набор высокоэффективных критериев разделения событий [1–3]. Однако существуют ситуации, характеризующиеся невозможностью обеспечения полного согласия между физическими процессами в

Национальный исследовательский технологический университет “МИСиС”, 119049 Россия, Москва, Ленинский пр-т, 4; e-mail: shorkin.ra@misis.ru.

основе классифицируемых событий и наборами данных, используемыми для обучения классификатора. Например, при использовании моделированных наборов данных указанные ситуации могут быть вызваны ограниченностью теоретических моделей. Это, в свою очередь, может привести к возникновению систематических эффектов, способных отрицательно повлиять на чувствительность измерений.

В данной работе рассмотрен случай прямого поиска гипотетической частицы X новой физики с неизвестной массой. Данный сценарий характерен для исследований универсальных экспериментов БАК ATLAS и CMS, а также планируемого эксперимента SHiP. Предполагается, что сечение рождения и относительная вероятность распада частицы X в детектируемое конечное состояние Y известны и моделируемы методом компьютерной симуляции. При этом сопутствующие процессы Стандартной модели, а также нефизичные фоновые вклады, связанные с реконструкцией событий в экспериментальной установке, являются чрезмерно множественными для генерации доступными вычислительными средствами. Более того, неопределенность в массе искомой частицы X делает невозможным выбор априори свободной от сигнала области спектра инвариантных масс набранных событий. Таким образом, задача выделения чистой выборки фоновых событий является нетривиальной, но в то же время актуальной: выборка фоновых событий используется для определения эффективности реконструкции событий и их отбора, а также в качестве классификатора для подавления фоновых вкладов в экспериментальном наборе данных.

Одним из возможных методов получения набора фоновых событий в описанных условиях является выборка малой доли полного набора экспериментальных данных. Число сигнальных событий в такой выборке прямо пропорционально её объему, поэтому достаточно малая выборка может включать только преобладающие фоновые события. Так, в работе [4], посвященной поиску гипотетических тяжелых нейтральных лептонов, авторы используют менее одной тысячной доли от полного набора событий протон-протонных соударений с энергией $\sqrt{s} = 13$ ТэВ интегральной светимостью 41.6 фб^{-1} , зарегистрированных детектором CMS в качестве фоновой выборки. К редким сигнальным событиям в преобладающей фоновой выборке также можно отнести события ШАЛ (широких атмосферных ливней), рожденных гамма-квантами сверхвысоких энергий, в выборке всех событий ШАЛ, регистрируемых гамма-астрономическими экспериментами. Отношение гамма-событий к событиям, вызванным космическими протонами, может варьироваться в пределах от 1:100 [5] до 1:10000 [6].

Альтернативным способом выборки фонового набора данных является метод построения нерассмотренных областей преобразованием соседних интервалов (Constructing Unobserved Regions by Transforming Adjacent Intervals, CURTAINS) [7].

В работе рассмотрен первый из описанных подходов. Представлено описание исследования, проведенного для случая эксперимента, набор данных которого включает в себя преобладающие события одного рода (фон) и редкие события второго рода (сигнал). Рассмотрены метрики эффективности классификаторов, обученных на наборе фоновых событий с ошибочно идентифицированными сигнальными событиями, в частности, для деревьев решений [8]. Проведено сравнение между метриками работы классификаторов, обученных на “чистых” и “смешанных” фоновых наборах.

Анализ поставленной проблемы. В основе рассматриваемого сценария находится предположение о существовании предсказанного из теоретических соображений или измеренного в предыдущих экспериментах верхнего предела на сечение рождения гипотетической частицы X и относительной вероятности её распада (branching fraction) в детектируемое конечное состояние Y . В общем случае, для шага k многоступенчатого алгоритма отбора (может включать в себя триггер эксперимента, реконструкцию событий, последовательное наложение критериев и пр.) верхний предел сигнального выхода в наборе данных может быть выражен согласно

$$N_S^k = L \cdot \sigma(X) \cdot Br(X \rightarrow Y) \cdot \prod_{i=1}^k \epsilon_i,$$

где L – интегральная светимость эксперимента, $\sigma(X)$ – верхний предел сечения рождения X в эксперименте, $Br(X \rightarrow Y)$ – относительная вероятность распада X в состояние Y , ϵ_i – условная эффективность реконструкции и отбора выполненных шагов алгоритма. Интегральная светимость измеряется в эксперименте, условные эффективности могут быть получены при моделировании сигнальных событий на основе теоретических предположений о сечениях редких процессов.

Для поставленной задачи обучения и оптимизации отбора ожидаемый выход сигнала N_S^k (S – signal) может быть рассмотрен, как консервативная оценка абсолютного вклада сигнальной компоненты в набор данных с преобладающими фоновыми событиями. Консервативная оценка относительного вклада сигнала в фоновый набор $r_{S/B}$ (B – background) может быть выражена, как отношение

$$r_{S/B} = \frac{N_S^k}{N^k},$$

где N^k – число событий в наборе данных на момент k -го шага отбора.

Абсолютный N_S^k и относительный $r_{S/B}$ вклады сигнальной компоненты являются внешними параметрами, влияющими на оптимальный размер фоновой выборки, используемой для обучения и оптимизации отбора. К внешним параметрам также относится разрешающая способность распределений измеренных характеристик сигнальных и фоновых событий. В качестве разрешающей способности (мощности) P_{sep} (separation power) данного параметра F событий примем отношение суммы весов правильно классифицированных S_{true} сигнальных событий к полной сумме весов S_{total} сигнальных событий при проведении границы отбора (cut) в точке равенства сигнального и фоновых распределений параметра F_{cut} . Для определенности рассмотрим случай нормального распределения признаков, где средние значения для сигнального и фоновых распределений соотносятся согласно $\mu_S > \mu_B$ (рис. 1). Тогда для P_{sep} можем записать

$$P_{\text{sep}} = \frac{(S_{\text{true}}|F > F_{\text{cut}})}{S_{\text{total}}}.$$

Для набора из N параметров разрешающая способность $\hat{P}_{\text{sep}}$ является средним значением разрешающих способностей составляющих параметров. Приведенное определение может быть проиллюстрировано примером распределений параметров с большей (0.68) и меньшей (0.55) разрешающей способностью, приведенным на рис. 1(а) и 1(б), соответственно.

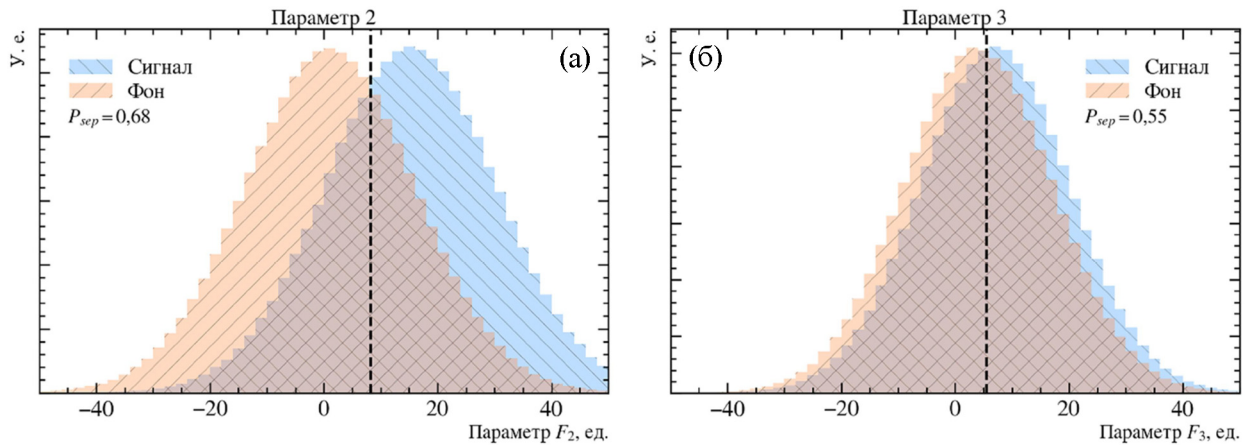


Рис. 1: Распределения параметров (а) F_2 и (б) F_3 для сигнальных (синий) и фоновых (оранжевый) событий. Пороговое значение F_{cut} обозначено пунктиром. Разрешающая способность P_{sep} составляет (а) 0.68 и (б) 0.55.

Наиболее простой подход для вычисления оптимального размера выборки предполагает вычисление такой доли от полного набора данных, что ожидаемый верхний предел

абсолютного вклада сигнальной компоненты N_S^k ограничен пренебрежимо малым значением, к примеру 0.1 событий. В рамках такого подхода возможно построение границ размеров выборок, масштабирующихся по $r_{S/B}$.

Подобный простейший подход не учитывает разрешающую способность характеристических параметров в основе классификации. Увеличение размеров выборки, используемой для обучения и оптимизации отбора, может положительно влиять на эффективность, несмотря на соответствующее возрастание доли сигнала фоновой выборке.

Компьютерное моделирование данных и обучение классификаторов. Для исследования влияния вклада неверно идентифицированных сигнальных событий в фоновую выборку на характеристики классификаторов были проведены обучение и оптимизация деревьев решений с бустингом (BDT, Boosted Decision Tree). Выборки для обучения и оптимизации включали в себя десять характеристических параметров (признаков), обозначаемых F^i ($i = 0, \dots, 9$). Примером параметров в эксперименте могут служить энергия частицы, ее поперечный импульс, псевдобыстрота и пр. Параметры сигнальных F_S^i и фоновых F_B^i событий были сгенерированы на основе соответствующих 10-мерных нормальных распределений.

Базовые распределения, соответствующие сигналу и фону, были созданы численными методами моделирования случайных величин с заданными наборами средних и дисперсий распределений для каждой из величин. Использовался алгоритм “random.multivariate_normal” библиотеки NumPy [9] языка программирования Python.

Разница между соответствующими компонентами векторов средних значений распределений характеристических параметров сигнальных и фоновых событий использовалась для контроля разрешающей способности набора $\hat{P}_{\text{sep}}$. Средние значения μ_B^i параметров F_B^i для фоновых событий были выбраны из равномерного распределения с пределами $(-8, 8)$. Средние значения μ_S^i соответствующих параметров F_S^i для сигнальных событий были определены посредством заданного сдвига $\mu_S^i = \mu_B^i \pm \Delta^i$.

Было рассмотрено две пары наборов сигнал-фон с различными средними разрешающими способностями. Для набора данных с большим из рассмотренных средним значением разрешающей способности ($\hat{P}_{\text{sep}} = 0.588$) значение Δ^i составляло от 2 до 12. Для набора данных с меньшей разрешающей способностью ($\hat{P}_{\text{sep}} = 0.541$) Δ^i составляло от 0 до 9. Исследование двух различных пар наборов позволяет произвести сравнение влияния малого вклада сигнальных событий в смешанный фоновый набор для классификаторов разной мощности.

Корреляционная матрица полученных в результате компьютерного моделирования распределений параметров представлена на рис. 2.

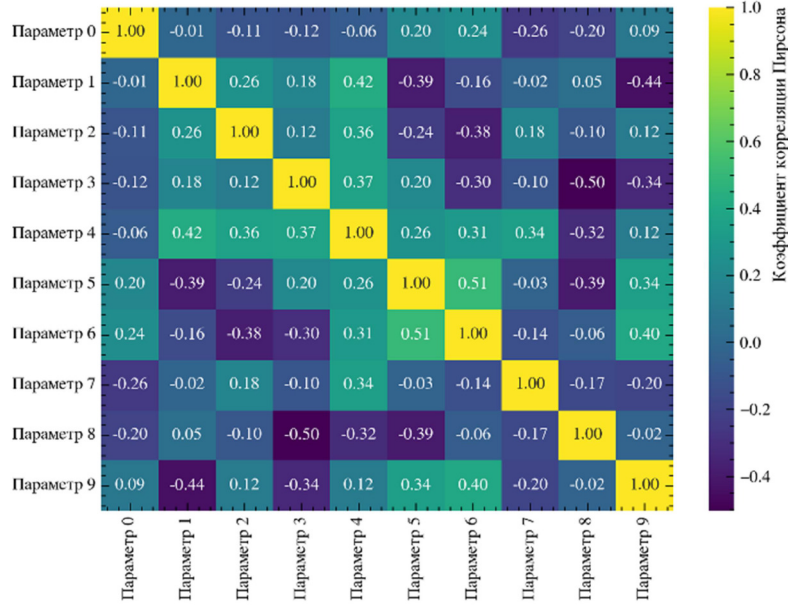


Рис. 2: Матрица корреляций характеристических параметров классификаторов.

Обучение исследуемых классификаторов контролировалось с помощью единого набора значений гиперпараметров. В частности, для классификаторов типа BDT использовались следующие параметры: максимальная глубина равнялась 3, скорость обучения равнялась 0.3, число циклов бустинга составляло 200.

Для реализации описанных выше условий поиска распадов гипотетической частицы X неизвестной массы была проведена процедура смешивания наборов внутри каждой пары.

- Было сгенерировано $N^k = 10^6$ фоновых и $N_S^k = 10^3$ сигнальных событий с десятью характеристическими параметрами на основе 10-мерного нормального распределения;
- $N_S^k = 10^3$ случайных фоновых событий из набора было заменено на сигнальные, что соответствует относительному вкладу $r_{S/B} = 0.1\%$. Данная примесь сигнальных событий в фоне не изменялась на протяжении всего исследования. Все события этого набора были помечены, как фон. Данный набор далее называется смешанным фоном;
- Для сравнительных исследований был сгенерирован набор из 10^6 фоновых событий. Данный набор далее называется чистым фоном;

- Для обучения, проверки и оптимизации отбора был сгенерирован набор из 10^6 сигнальных событий (далее чистый сигнал).

В результате процедуры было получено три набора смоделированных событий (выборка смешанного фона, выборка чистого фона и выборка чистого сигнала) для обоих рассматриваемых значений $\hat{P}_{\text{sep}}$.

Включение ошибочно идентифицированных сигнальных событий в смешанные фоновые наборы, используемые для обучения и оптимизации классификаторов, способно привести к систематическому сдвигу при вычислении оптимального критерия отбора, что, в свою очередь, оказывает прямое воздействие на чувствительность к сигналу и эффективность подавления фона. Для получения количественной оценки сдвига в эффективности классификаторов была проведена тренировка и оптимизация на чистом и смешанном наборах фоновых событий, а также чистом наборе сигнальных событий.

Стратегия формирования выборок для обучения и проверок классификаторов из сигнального и фонового набора данных сформулирована ниже. Описанный алгоритм использовался для двух комбинаций наборов с различной средней разрешающей способностью $\hat{P}_{\text{sep}}$.

BDT первого набора обучались на смеси фоновых и сигнальных событий в качестве первой выборки и чистых сигнальных событиях в качестве второй. В качестве первого шага были выбраны события из набора смешанного фона для обучения классификатора. Объем выборки варьировался в пределах от 0.1 до 5% от числа событий в полном наборе (10^6): $f_{\text{train}} = 0.1, 0.2, 0.5, 1, 2$ и 5%. Доля в 0.1% была выбрана для обеспечения минимального абсолютного вклада сигнальных событий в фоновую выборку (1 сигнальное событие на выборку). Остальные доли были выбраны для исследования влияния увеличенного абсолютного вклада неверно идентифицированного сигнала на работу классификатора.

Далее была выполнена выборка чистых сигнальных событий того же объема (0.1, 0.2, 0.5, 1, 2 и 5% от полного набора в 10^6) для использования при обучении.

Для тестирования и оптимизации BDT необходимо провести исследования на выборках, не имеющих пересечений с выборками для обучения. С этой целью для каждого классификатора, обученного на выборке размером f_{train} , была выполнена дополнительная независимая выборка объемом $f_{\text{test}} = 0.1, 0.2, 0.5, 1, 2$ и 5%. При этом объем выборки чистых сигнальных событий не варьировался и составлял 20×10^3 событий. События в наборах, использованных для проверки и оптимизации, получили веса на основе ожидаемых долей сигнальных и фоновых событий.

Для BDT второго набора была задействована аналогичная процедура с заменой смешанных фоновых наборов на чистые. Классификаторы второго набора использовались для определения реперных точек при определении сдвигов и систематических эффектов, проявляющихся при обучении и оптимизации на смешанных наборах.

Для каждой из двух комбинаций наборов (чистый фон, смешанный фон и чистый сигнал) с различной разрешающей способностью ($\hat{P}_{\text{sep}} = 0.588$ и $\hat{P}_{\text{sep}} = 0.541$) и для каждого объема выборок, используемых при обучении и оптимизации отбора (0.1, 0.2, 0.5, 1, 2 и 5% из 10^6), было получено по 500 различных BDT, каждое из которых было обучено и оптимизировано на уникальных выборках из наборов соответствующей комбинации. Обучение классификаторов типа BDT производилось с использованием пакета XGBoost языка программирования Python [10].

Сравнение классификаторов на основе параметра ROC AuC (площадь под кривой ошибок). В качестве количественной оценки предсказательной мощности обученного классификатора была рассмотрена площадь под кривой ошибок (Receiver Operating Characteristic Area under Curve, ROC AuC). Полученные значения площадей под кривыми ошибок для классификаторов на базе различных наборов параметров, обученных на смешанных и чистых фоновых наборах, представлены на рис. 3 синими и оранжевыми маркерами, соответственно. Для классификаторов с большей разрешающей способностью (рис. 3(а)) предсказательная мощность возрастает пропорционально относительному объему выборки (f_{train}), используемой для обучения, и достигает плато при $f_{\text{train}} = 2\%$. С другой стороны, для классификаторов с меньшей разрешающей способностью (рис. 3(б)) мощность достигает пикового значения при относительном объеме выборки, соответствующем 0.5% изначального набора данных. Это может быть вызвано как возрастанием абсолютного числа сигнальных событий в смешанном фоне, оказывающем отрицательное влияние на обучение более слабого классификатора, так и увеличением числа значений параметров, относящихся к хвостам распределений с широкой дисперсией. Последнее приводит к большему пересечению распределений параметров сигнальных и фоновых событий и, как следствие, к уменьшению способности классификатора к их различению. Количественный показатель предсказательной мощности классификаторов, обученных на смешанной фоновой выборке, ниже, чем у классификаторов, обученных на чистых наборах данных.

Для исследования возможного эффекта статистического объема контрольных выборок, используемых при построении кривых ошибок, на измеренное значение площади под кривой, была выполнена перекрестная проверка. Как можно увидеть из рис. 4,

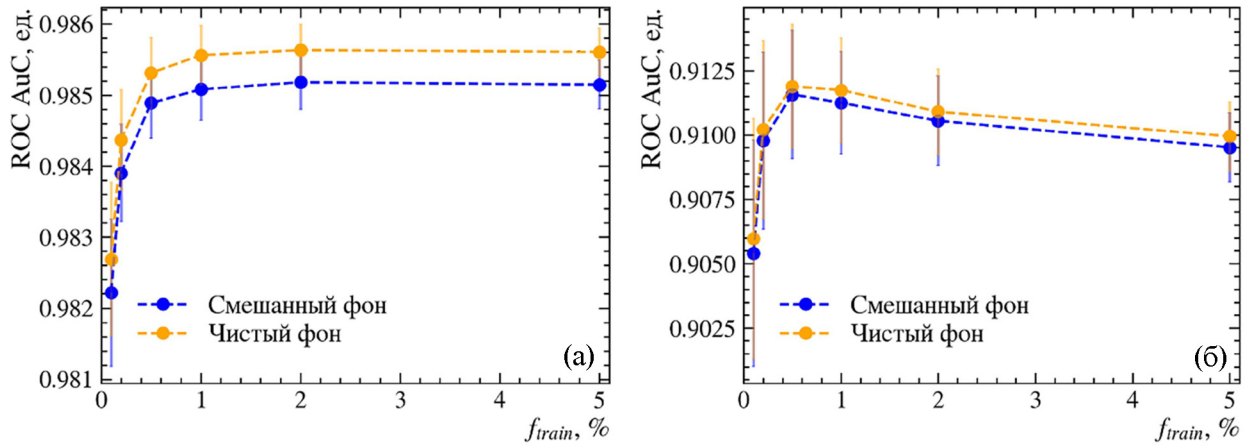


Рис. 3: Средние величины площади под кривой ошибок для классификаторов, обученных на наборах данных с (а) большей и (б) меньшей разрешающей способностью в зависимости от размеров обучающей выборки. Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанной выборке – синим.

полученные значения параметра предсказательной мощности классификатора не изменяются в зависимости от объема тестовой выборки для рассмотренных значений f_{test} .

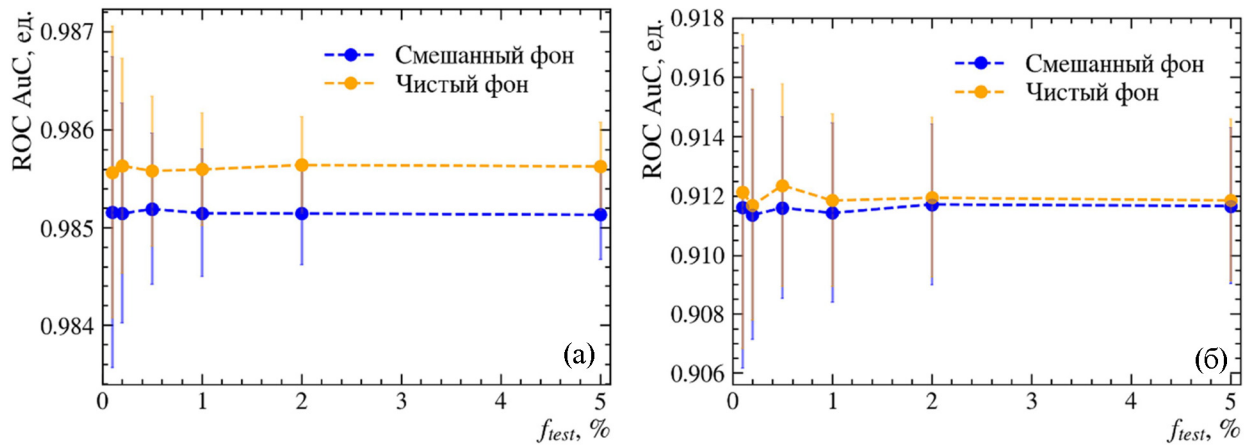


Рис. 4: Средние величины площади под кривой ошибок ($ROC AuC$) для классификаторов с (а) меньшей и (б) большей разрешающей способностью в зависимости от относительного объема проверочной выборки f_{test} . Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанной выборке – синим.

Критерий оптимальности классикатора и влияние объема тестовой f_{test} выборки на оптимизацию критериев отбора. Для оценки влияния ввода неверно идентифици-

рованных сигнальных событий в выборку, используемую в качестве фона при оптимизации отбора, была проведена процедура оптимизации классификаторов путем выбора границы между классами по оси предсказательной переменной $sBDT$ классификатора (BDT score) (рис. 5). Под предсказательной переменной понимается величина классифицирующей функции, соответствующая значениям вектора параметров данного события и характеризующая вероятность его отнесения к классу “сигнал”. Классификаторы оптимизировались на двух парах выборок: “чистый фон–чистый сигнал” и “смешанный фон–чистый сигнал”. Выполнялось сравнение характеристик их эффективности: добротности, чувствительности по отношению к сигналу (эффективность отбора сигнальных событий ϵ_S), коэффициента подавления фона (эффективность отбора фоновых событий ϵ_B).

Для пояснения используемого в исследовании метода оптимизации, рассмотрим описание базового сценария классификаторов, обученных на выборках объемом $f_{\text{train}} = 2\%$ от первоначального набора данных с большей разрешающей способностью. Объем базовой выборки был определен согласно оптимуму графика площадей под кривой ошибок (рис. 3).

Оптимизация отбора основана на нахождении максимума показателя добротности. В рассматриваемом случае в качестве показателя добротности была выбрана статистическая значимость сигнала $S/\sqrt{S+B}$, где S и B – суммы весов сигнальных и фоновых событий, соответственно, удовлетворяющих данному критерию по предсказательной переменной классификатора. Коэффициент добротности вычислялся для классификаторов, обученных и оптимизированных на чистых и на смешанных фоновых выборках.

Различие в эффективности оптимизации отбора может быть визуально наблюдаемо при сравнении графиков добротности классификаторов, оптимизированных на конкретных контрольных (независимо подобранных) выборках относительным объемом в 0.1% (рис. 5(а)) и 1% (рис. 5(б)) от изначального набора данных. Меньший объем выборки (0.1%), используемой при оптимизации, приводит к увеличению статистической погрешности при измерении показателей добротности. С другой стороны, соответствующее уменьшение в ожидаемом верхнем пределе количества неверно идентифицируемых сигнальных событий $N_S^k \cdot f_{\text{test}}$ (то есть абсолютного количества сигнальных событий) в смешанной фоновой выборке приводит к возрастанию степени согласия между добротностью классификаторов, обученных и оптимизированных на смешанных и чистых фоновых наборах. Напротив, в случае обучения и оптимизации на выборке большего объема (1%) наличие значительного абсолютного количества сигнальных событий в

смешанной выборке ухудшает добротность созданных критериев. Дополнительно можно видеть, что неуничтожимое загрязнение фоновой выборки сигнальными событиями приводит к существенной разнице между величинами наилучших $sBDT$.

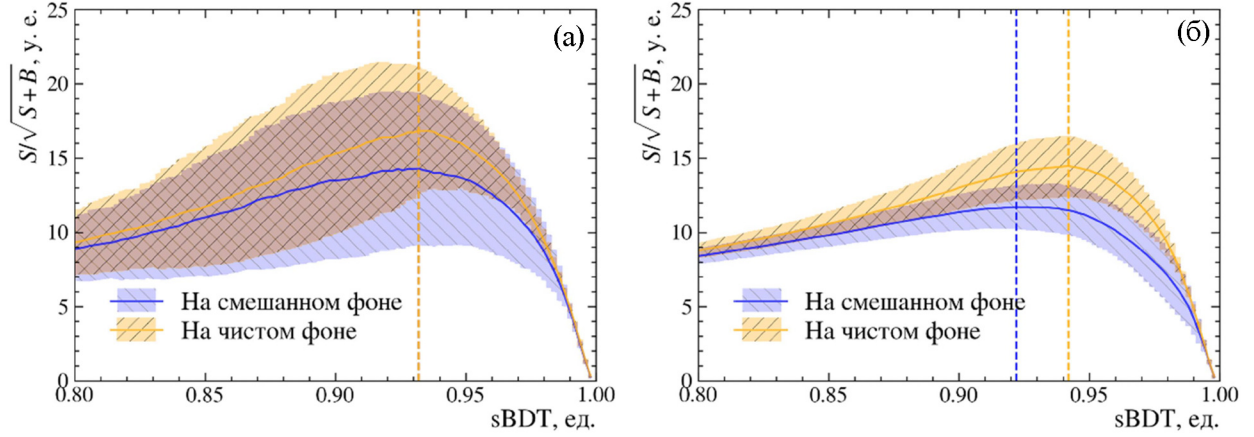


Рис. 5: Добротность классификаторов, обученных и оптимизированных на наборах “чистый сигнал и смешанный фон” (синий) и “чистый сигнал и чистый фон” (оранжевый). Объем обучающей выборки составляет $f_{train} = 2\%$ от полного набора данных с большей разрешающей способностью. Размер фоновой выборки, использованной при оптимизации, соответствует (а) 0.1% и (б) 1% от соответствующего набора. Пунктирные линии выделяют значения $sBDT$, соответствующие максимуму добротности $sBDT = sBDT_{optimal}$.

На следующем шаге рассмотрим влияние разных объемов тестовых выборок f_{test} на результаты оптимизации. Значения максимальных коэффициентов добротности классификаторов для всех рассматриваемых величин относительных объемов выборок f_{test} представлены на рис. 6. График слева соответствует случаю большей разрешающей способности набора параметров $\hat{P}_{sep} = 0.588$, график справа – меньшему значению разрешающей способности $\hat{P}_{sep} = 0.541$. Классификаторы с большей разрешающей способностью были обучены на выборках объемом $f_{train} = 2\%$ от изначального набора данных. Обучение классификаторов с меньшей разрешающей способностью производилось на выборках объемной долей в $f_{train} = 0.5\%$. Также, как и для частных случаев $f_{test} = 0.1, 1\%$, представленных на рис. 5, для всего рассматриваемого диапазона объемной доли f_{test} можно видеть, что разница добротностей классификаторов, обученных на чистых и смешанных выборках, минимальна для случаев малых тестовых выборок и увеличивается с ростом их объемов. Абсолютные значения добротности классификаторов с боль-

шей разрешающей способностью выше, чем классификаторов с меньшей разрешающей способностью для всего рассматриваемого диапазона f_{test} . Наблюдаемые максимальные значения показателя добротности уменьшаются пропорционально увеличению объема выборки для оптимизации и достигают плато при $f_{\text{test}} \approx 2\%$.

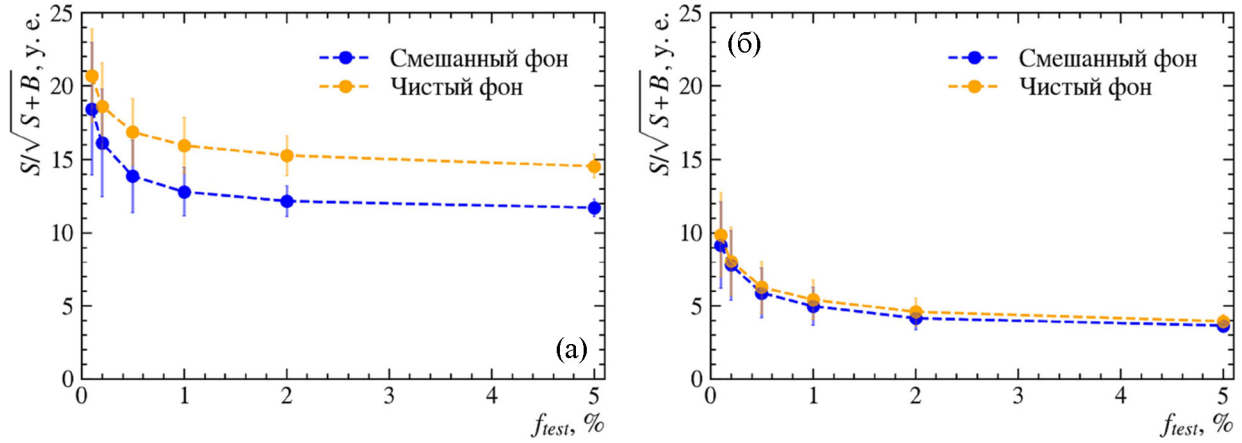


Рис. 6: Максимальные значения добротности классификаторов с (а) большей и (б) меньшей разрешающей способностью в зависимости от относительного объема проверочной выборки f_{test} . Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанном фоне – синим.

Дополнительно в ходе исследования была выполнена перекрестная проверка, осуществленная путем оптимизации классификаторов, обученных на фоновых выборках различных объемов. Результаты представлены на рис. 7. Наблюдаемые зависимости максимумов добротности от объема выборки коррелируют с зависимостью показателя предсказательной мощности классификатора, представленной на рис. 3.

Определим эффективность отбора с использованием критерия классификатора в виде

$$\epsilon_{S(B)} = \frac{S(B)(sBDT > \text{cut})}{S(B)_{\text{total}}}.$$

Значения предсказательной переменной классификатора $sBDT = sBDT_{\text{optimal}}$, соответствующие максимумам показателей добротности, были использованы для вычисления величин чувствительности модельного эксперимента к сигналу ϵ_S и подавлению фона ϵ_B .

На первом шаге расчет эффективностей проводился на основе чистых сигнальной и фоновой выборок с величиной $sBDT_{\text{optimal}}$ (т. е. для оптимизированного критерия), на

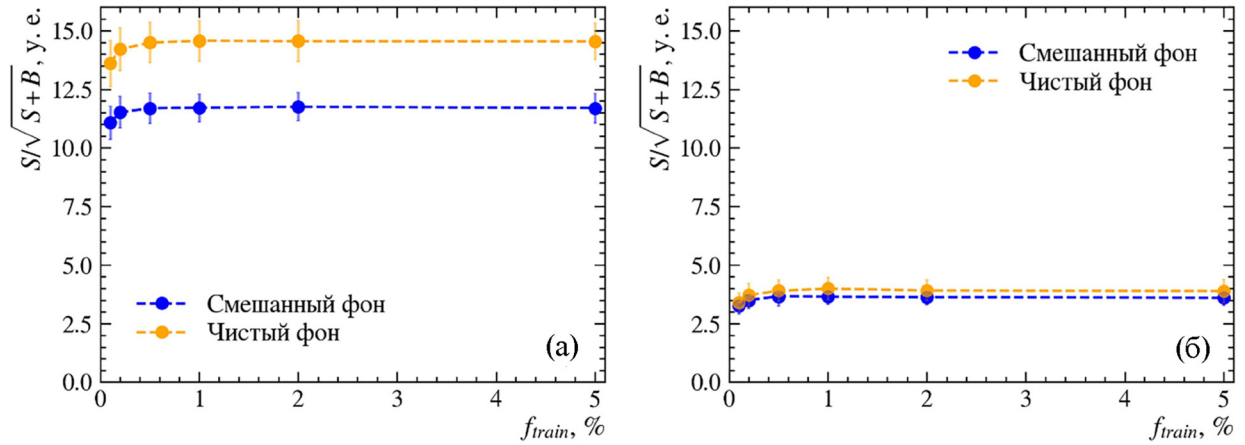


Рис. 7: Максимальные значения добротности классификаторов с (а) большей и (б) меньшей разрешающей способностью в зависимости от относительного объема выборки для обучения f_{train} . Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанном фоне – синим.

втором шаге – на основе чистой сигнальной и смешанной фоновой выборках. Рассчитанные таким образом значения $\epsilon_{S,B}$ позволяют оценить разницу между истинным (случай “чистый сигнал–чистый фон”) и наблюдаемым (“чистый сигнал–смешанный фон”) оптимальными критериями.

Различия в выборках приводят к различиям в эффективностях классификаторов по отношению к сигналу ϵ_S (рис. 8) и фону ϵ_B (рис. 9) для рассматриваемых значений разрешающей способности $\hat{P}_{sep}$. Оптимизация, произведенная на основе выборок меньшего объема, приводит к возрастанию статистической погрешности ϵ_S и ϵ_B , однако средняя эффективность при этом принимает сравнительно лучшие значения.

Для классификаторов с большей разрешающей способностью измеренная эффективность по отношению к сигналу ϵ_S (рис. 8(а)) максимальна в случае $f_{test} = 0.1\%$ и слабо изменяется в рассматриваемом диапазоне. Эффективность по отношению к сигналу ϵ_S классификаторов с меньшей разрешающей способностью (рис. 8(б)) медленно возрастает в области f_{test} от 1% до 5%.

Дополнительно следует заметить, что средняя эффективность по отношению к сигналу ϵ_S , полученная на основе выборок ограниченного объема (соответствующего $f_{test} \approx 0.2\%$ и ниже), демонстрирует высокую степень согласия между классификаторами, обученными на чистых и смешанных фоновых выборках.

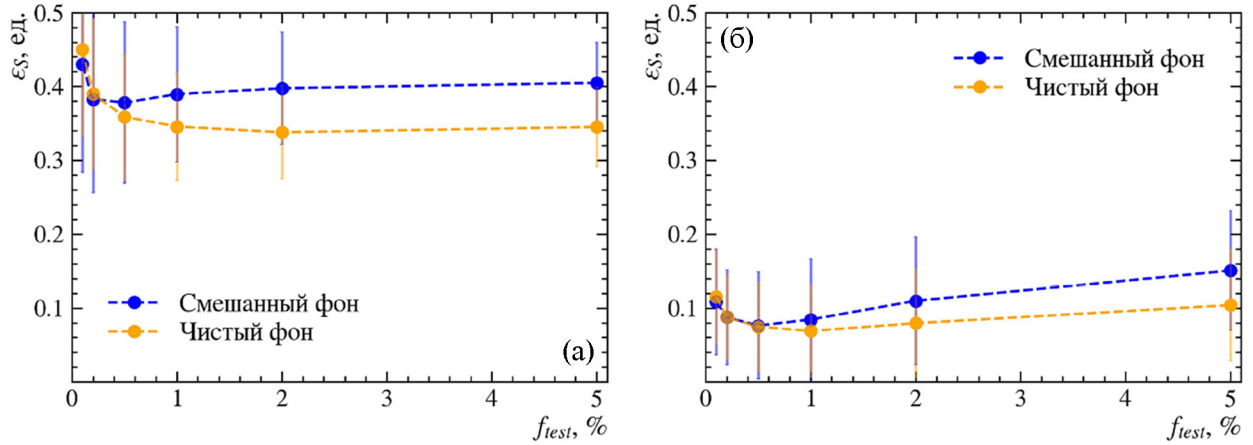


Рис. 8: Средняя эффективность ϵ_S оптимального критерия отбора на основе классификатора, оптимизированного на наборах данных с (а) большей ($\hat{P}_{sep} = 0.588$) и (б) меньшей ($\hat{P}_{sep} = 0.541$) разрешающей способностью в зависимости от относительного объема тестовой выборки f_{test} . Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанном фоне – синим.

Зависимость измеренной эффективности отбора по отношению к фону ϵ_B от относительного объема тестовой выборки f_{test} (рис. 9) демонстрирует схожий характер. Основное отличие связано с поведением эффективности ϵ_B , рассчитанной для случая обучения и оптимизации с использованием чистой фоновой выборки. В то время как в случае смешанного фона эффективность составляет $\epsilon_B \approx 0.1\%$ для относительного объема $f_{test} = 0.1\%$, при использовании выборки чистого фона того же объема эффективность ϵ_B приближается к нулю (полное подавление фона).

Это может быть объяснено наличием неверно идентифицированных сигнальных событий в смешанной фоновой выборке, используемой при оптимизации. Так, подобные сигнальные события, неверно помеченные как фон (относительный вклад которых составляет $r_{S/B} = 0.1\%$), могут быть отброшены только при более строгом (более высоком), неоптимальном пороговом значении предсказательной переменной $sBDT$. Следовательно, эти события могут быть рассмотрены как минимальный уровень фона. Наличие этого минимального фона вкупе с различиями, обусловленными обучением классификаторов на чистой и смешанной фоновой выборках, могут объяснить систематический сдвиг в пороговом значении предсказательной переменной $sBDT_{optimal}$, что, в свою очередь, приводит к изменению измеренных значений эффективностей ϵ_S и ϵ_B .

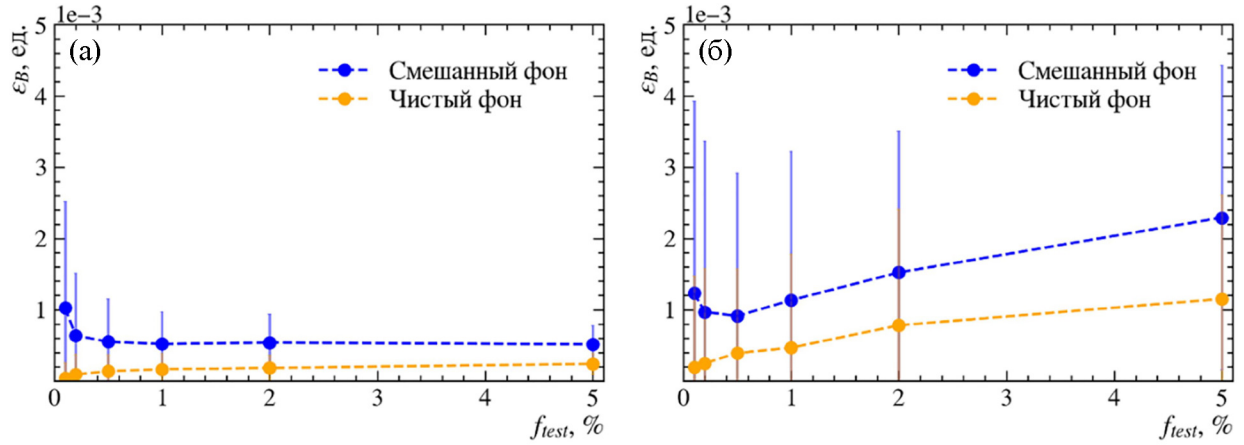


Рис. 9: Средняя эффективность ϵ_B оптимального критерия отбора на основе классификатора с (а) большей и (б) меньшей разрешающей способностью в зависимости от относительного объема тестовой выборки f_{test} . Значения для классификаторов, обученных на чистой фоновой выборке, обозначены оранжевым; на смешанном фоне – синим.

Заключение. В выполненном исследовании была рассмотрена проблема обучения и оптимизации классификатора для отбора редких событий в присутствии неопределенного преобладающего фона. Данная задача актуальна для случая поиска редких распадов гипотетической частицы X неизвестной массы в выборке событий, состоящей, в основном, из фоновых процессов. На данном этапе исследования рассматривался случай фиксированного относительного вклада сигнальных событий в смешанные фоновые выборки $r_{S/B} = 0.1\%$. В дальнейшем планируется рассмотреть эффекты, связанные с варьированием величины относительного вклада.

Результаты исследования показали, что использование выборки с пренебрежимо малым числом сигнальных событий (0.1% для рассмотренного набора данных, что соответствует 1 ожидаемому сигнальному событию в смешанной фоновой выборке) не является оптимальным. Увеличение относительного объема фоновой выборки, используемой при обучении классификатора, приводит к повышению его мощности, оценка которой производилась путем вычисления площади под кривой ошибок (ROC AuC) классификатора. Для классификаторов, обученных на наборах событий с параметрами с большей разрешающей способностью, увеличение объема тестовой выборки для обучения слабо влияет на метрики эффективности. Классификаторы с меньшей предсказательной мощностью обладают относительно малой разницей в метриках при обучении на чистых

и смешанных выборках фоновых событий. С другой стороны, при оптимизации отбора небольшие выборки выглядят более предпочтительными. Поскольку решение об объеме выборки для задач обучения и оптимизации производится на основе предположения о “худшем сценарии” – количество сигнальных событий в наборе соответствует оценке высшей границы – возможна итеративная процедура. После более точного измерения сечения процессов с искомым сигналом в анализе возможно произвести исследование с улучшенными оценками оптимальных объемов выборок.

Л И Т Е Р А Т У Р А

- [1] Lonnblad, C. Peterson, T. Rognvaldsson, Phys. Rev. Lett. **65**(11), 1321 (1990). DOI: 10.1103/PhysRevLett.65.1321.
- [2] G. Aad (on behalf of ATLAS collaboration), J. Phys. Conf. Ser. **523**(1), 012023 (2014). DOI: 10.1088/1742-6596/523/1/012023.
- [3] P. Abreu et al. (on behalf of DELPHI collaboration), Phys. Lett. B **295**(3-4), 383 (1992). DOI: 10.1016/0370-2693(92)91580-3.
- [4] A. Hrapetyan et al. (on behalf of CMS collaboration), JHEP **06**, 183 (2024). DOI: 10.1007/JHEP06(2024)183.
- [5] A. P. Kryukov, A. P. Demichev, V. A. Ilyin, Moscow Univ. Phys. **79** (Suppl 2), 690 (2024). DOI: 10.3103/S002713492470214X.
- [6] E. O. Gres, A. P. Kryukov, A. P. Demichev, et al., Moscow Univ. Phys. **78** (Suppl 1), 45 (2023). DOI: 10.3103/S002713492307010X.
- [7] A. Raine, S. Klein, D. Sengupta, T. Golling, Front. Big Data **6**, 899345 (2023). DOI: 10.3389/fdata.2023.899345.
- [8] Y. Coadou, *Artificial Intelligence for High Energy Physics*. Chapter 2: Boosted decision trees, World Scientific Publishing, 2022, pp. 9-58. DOI: 10.1142/9789811234033_0002.
- [9] C. R. Harris et al., Nature **585**, 357 (2020). DOI: 10.1038/s41586-020-2649-2.
- [10] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, 2016), pp. 785–794. DOI: 10.1145/2939672.2939785.

Поступила в редакцию 8 июля 2024 г.

После доработки 2 августа 2024 г.

Принята к публикации 5 августа 2024 г.