Содержание

1. Введение

В Институте ядерной физики им. Г. И. Будкера СО РАН проводятся эксперименты на электрон-позитронном коллайдере ВЭПП-2000. Коллайдер имеет два места встречи пучков, в которых установлены детекторы КМД-3 (криогенный магнитный детектор) и СНД (сферический нейтральный детектор). В программу проекта ВЭПП-2000 включено измерение сечений процессов, наблюдаемых при столкновениях электронов и позитронов в диапазоне энергий до 2 ГэВ. В число изучаемых процессов входят процессы с фотонами в конечном состоянии [1].

Для регистрации фотонов в детекторе КМД-3 используется жидкоксеноновый (LXe) калориметр. Фотон, попадая в калориметр, порождает электромагнитный ливень из вторичных электронов, позитронов и фотонов. Частицы ливня регистрируются сигнальными полосками, расположенными на слоях калориметра. Сработавшие сигнальные полоски группируются в кластеры.

При столкновениях электронов e^- и позитронов e^+ большое сечение рождения имеют процессы, содержащие нейтральный пион π^0 в конечном состоянии. Одним из таких процессов является $e^-e^+ \to \pi^0\gamma$. Нейтральный пион с вероятностью 99% распадается на два фотона γ. Минимальный угол между фотонами при более высокой энергии пи-мезона становится меньше, и при значении $E_{\pi^0} = 1$ ГэВ этот угол составляет 0,27 радиана. Электромагнитные ливни таких фотонов в калориметре могут перекрываться, и появляется вероятность принять два слившихся кластера, образованных двумя фотонами, за одиночный кластер от одного фотона. Чтобы увеличить статистику событий, повысить точность измерения сечения и других параметров процесса, необходимо разделять такие кластеры [2].

Таким образом, требуется разработать алгоритм различения однофотонных и двухфотонных кластеров в LXe калориметре детектора КМД-3. Для решения этой задачи выполнены следующие шаги:

- описание детектора КМД-3 и LXe калориметра,
- подготовка данных для обработки,
- обзор библиотеки TMVA 4 и метода BDT,
- реализация алгоритма классификации с помощью метода BDT,
- анализ задачи разделения кластеров.

Результаты данной работы планируется сравнить с результатами, полученными в работах [2, 3].

2. Детектор КМД-3 и LXe калориметр

На рис. 1 представлена схема криогенного магнитного детектора КМД-3.

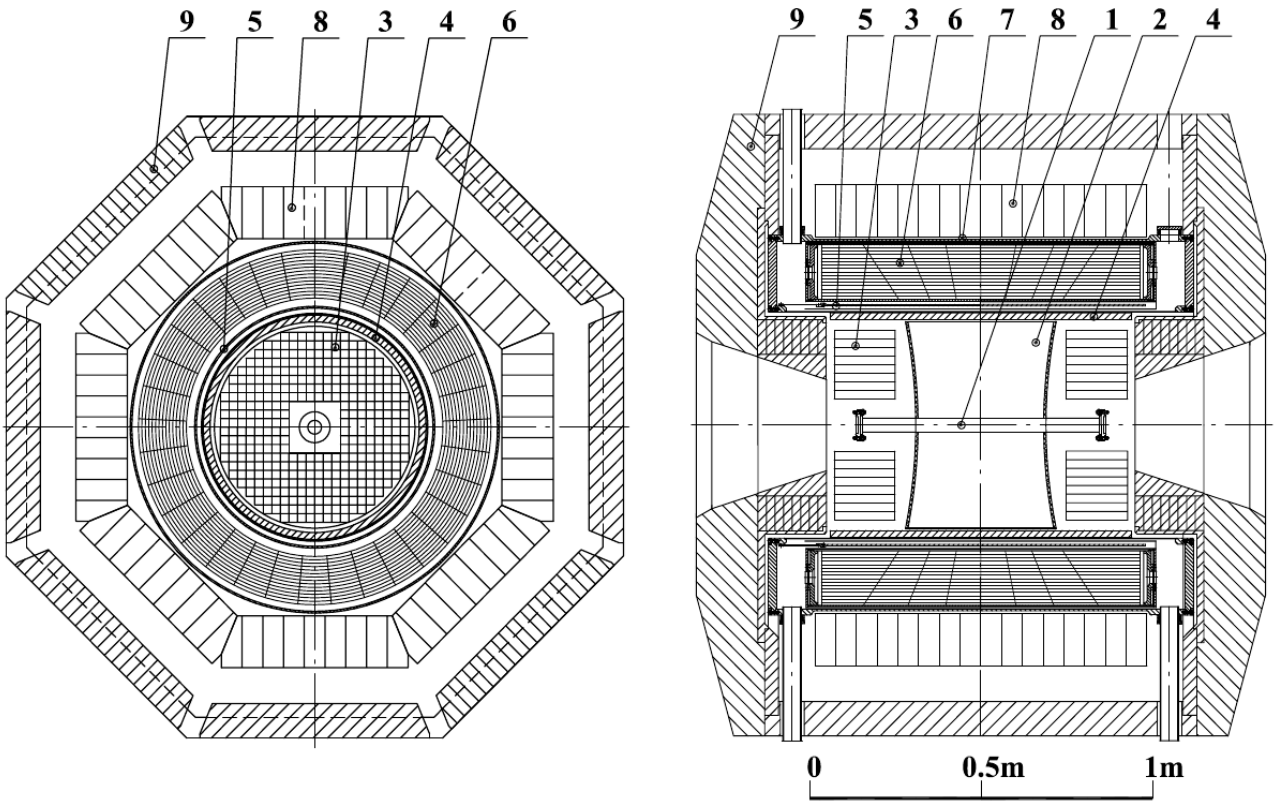

Рис. 1. Схема детектора КМД-3: 1 – вакуумная камера, 2 - дрейфовая камера, 3 – торцевой BGO калориметр, 4 – Z-камера, 5 – сверхпроводящий соленоид, 6 – цилиндрический жидкоксеноновый калориметр, 7 – времяпролётная система, 8 – цилиндрический CsI калориметр, 9 – ярмо магнита [4].

Цилиндрический электромагнитный калориметр детектора регистрирует фотоны и электроны, которые вылетают под большими углами к оси пучков (от 39° до 141°), и охватывает телесный угол $0,8 \cdot 4\pi$. Он состоит из двух

соосных подсистем: жидкоксенонового LXe калориметра и кристаллического CsI калориметра.

Жидкоксеноновый калориметр образован семью соосными катодами и восемью анодами, образующими систему ионизационных камер. Структура электродов LXe калориметра представлена на рис. 2.

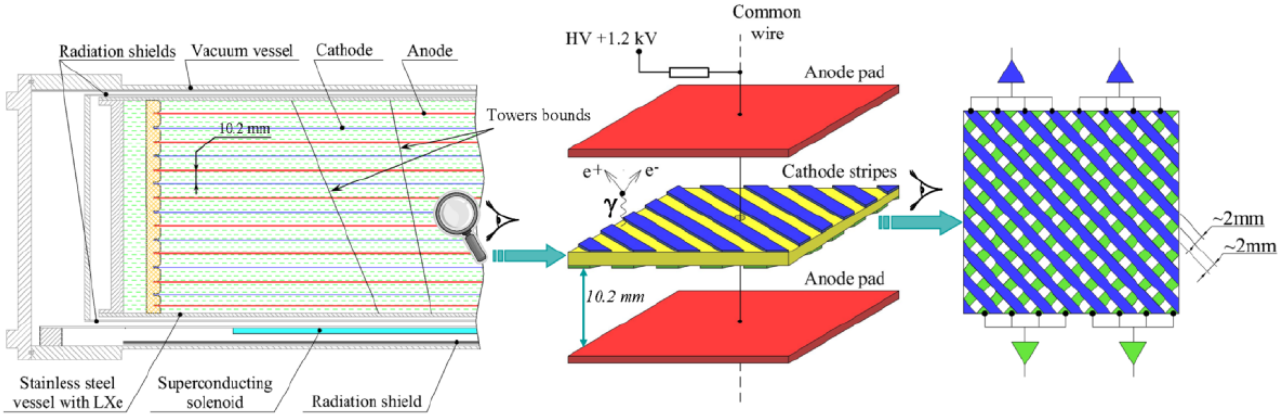

Рис. 2. Структура электродов жидкоксенонового калориметра [4].

Поверхность каждого анода разделена на 264 прямоугольные площадки – восемь вдоль оси пучков Z, тридцать три в перпендикулярной ей плоскости R, φ. Площадки с одинаковым положением по Z и φ соединены проводником в «башню». Такая «башня» ориентирована в место встречи пучков. Аноды используются для измерения выделенной энергии.

Каждый катод разбит на 2112 полосок. Полоски объединены в группы по четыре, и с каждой такой группы сигнал снимается по одному каналу. На противоположных сторонах катода полоски ориентированы взаимно перпендикулярно и находятся под углом 45° к оси пучков Z. Такая структура катода используется для восстановления координат кластера [4].

Регистрация фотонов производится путём сбора информации о частицах электромагнитного ливня, порождённого при взаимодействии фотона с материалом калориметра. LXe калориметр позволяет измерять координаты с погрешностью от 0,7 мм для первого слоя до 1,4 мм для шестого слоя [5, 6]. Точность измерения энергии составляет примерно 8,5% для нейтральных пимезонов [7].

3. Формат выходных данных с полосок и их обработка

В дальнейшей работе обучение и тестирование конечного алгоритма будет проводиться на результатах моделирования процесса $e^+ e^- \to \pi^0 \gamma$, так как данный процесс содержит оба типа событий (с кластером от одного фотона либо с объединённым кластером от двух фотонов), а также зависит только от калориметрической системы детектора.

В результате оцифровки данных с полосковой части LXe калориметра сохраняются номера каждой сработавшей полоски и амплитуды их сигналов. После обработки собранной информации проводится реконструкция полосковых кластеров. Полосковым кластером на одном из слоёв калориметра называется набор соседних полосок с одной стороны катода. Алгоритм реконструкции кластеров заключается в следующем:

- находятся полоски, амплитуда сигнала которых превысила верхний порог срабатывания,
- к найденным затравочным полоскам присоединяются соседние полоски, амплитуда которых превысила нижний порог срабатывания.

В рамках данной работы полоски, не принадлежащие ни к одному из полосковых кластеров, не будут использоваться далее.

Получившиеся данные сохраняются в формате классов CmdLXeStripHit и CmdLXeStripCluster, которые были специально разработаны командой детектора КМД-3. В классе CmdLXeStripHit содержится такая информация, как номер полоски, номер слоя, номер полоски на слое, направление полоски, амплитуда и ошибка измерения амплитуды, а также параметры спирали, которые задают геометрическое положение полоски. Класс CmdLXeStripCluster содержит следующие параметры: номер кластера и номера полосок, которые принадлежат этому кластеру. Таким образом, отбирались полоски, которые принадлежат к одному из полосковых кластеров, после чего записывались координаты точек всех попарных пересечений спиралей найденных полосок,

расположенных с разных сторон катода. Для удобства были созданы структуры strips и cross_pos, которые содержат информацию об индивидуальных свойствах полосок (номер, амплитуда, номер слоя и т. д.) и номера пересекающихся полосок с координатой их пересечения, соответственно.

Так как в данной работе будет использоваться алгоритм машинного обучения с учителем, то потребуется чёткое разделение кластеров на однофотонные и двухфотонные. Таким образом, для каждого события помимо вышеперечисленных данных сохраняются данные с генератора, такие как углы вылета, энергия и тип моделируемых частиц.

4. Библиотека TMVA и метод BDT

Библиотека TMVA (The Toolkit for Multivariate Analysis) предоставляет среду для параллельного вычисления и применения методов многомерной классификации и, начиная с версии TMVA 4, многомерной регрессии. В данной библиотеке представлены только алгоритмы машинного обучения с учителем. Они используют тренировочные события с известным результатом, чтобы определить отображение, которое описывает либо критерии классификации, либо регрессионное приближение для функции результата. Среда TMVA разработана в основном для нужд физики высоких энергий, но может применяться и в других областях [8].

В данной работе используется метод BDT (boosted decision trees), основанный на нахождении множества деревьев решений с присвоенными им весами. Дерево решений – это классификатор, построенный в виде двоичного дерева. Пример такого дерева изображён на рис. 3.

Задача алгоритма, использующего дерево решений, заключается в том, чтобы пройти это дерево, начиная с корневого узла, и закончить на одном из листьев дерева. В каждом узле находится критерий выбора дочернего узла по значению одной из переменных. В результате алгоритм относит событие к категории сигнальных или фоновых.

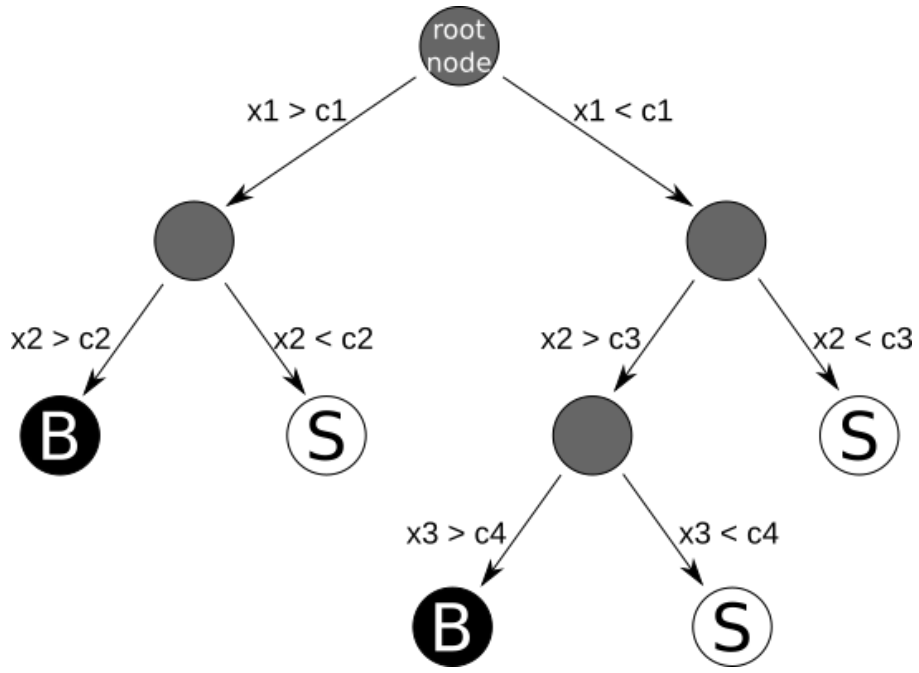

Рис. 3. Пример дерева решений. Решения принимаются на основе переменных $x1$, $x2$, $x3$, критериями являются величины $c1$, $c2$, $c3$, $c4$.

Тренировка дерева производится поэтапно, начиная с корневого узла. Выбирается критерий, наиболее эффективно делящий события на фоновые и сигнальные, то есть дающий наибольшую долю правильно классифицированных событий в обеих категориях. Такая же процедура выбора критерия выполняется на каждом из двух образовавшихся подмножеств событий и так далее до определённой глубины дерева. Таким образом, фазовое пространство оказывается поделено на сегменты, помеченные как сигнал или фон в зависимости от того, событий какого типа больше всего попало в данный сегмент при тренировке алгоритма.

Достоинства дерева решений – простота описания и интерпретации, а также возможность разбивать фазовое пространство на большое число гиперкубов, обозначаемых как фон или сигнал – в отличие от анализа на основе совокупности пороговых значений, которая выделяет лишь один гиперкуб. Недостатком дерева решений является его чувствительность к статистическим флуктуациям в тренировочном наборе событий. Например, две переменные, одинаково влияющие на классификацию событий, под действием флуктуаций могут идти в разном порядке следования сравнений, что изме-

нит и всю дальнейшую структуру дерева, в итоге воздействуя на результат классификации.

С целью уменьшения флуктуаций метод BDT использует бустинг деревьев. Для этого данные делятся на некоторое количество ансамблей N, и на каждом ансамбле тренируется по одному дереву решений. В итоге получается N деревьев, каждое из которых имеет свой уровень разделения и даёт свой отклик на одно и то же событие. Деревьям присваиваются веса в зависимости от качества классификации. Итоговая величина r отклика на событие вычисляется как средневзвешенное по всем деревьям:

$$r = \sum_i \omega_i r_i,$$

где ω_i – вес i-го дерева, r_i – отклик i-го дерева [8].

5. Алгоритм классификации с помощью BDT

Перед началом работ по реализации алгоритма разделения фотонных кластеров было решено выполнить проверку эффективности решения задачи классификации с помощью BDT на тестовых данных. Был разработан специальный генератор данных, который создавал близкие распределения трёх скоррелированных данных. Генератор тренировочного набора данных на выходе даёт заданное количество событий, имеющих три параметра: массу, рост и возраст. Каждому событию присваивается тип – сигнальное или фоновое событие (рис. 4).

Для реализации алгоритма подключена библиотека TMVA 4, выбран метод классификации BDT. В качестве критерия разделения при тренировке дерева можно применять $GiniIndex = p(1-p)$, где $p = \frac{s}{s+b}$ – «чистота» набора, s и b – соответственно, число сигнальных и фоновых событий в наборе; статистическую значимость $\frac{S}{\sqrt{S+B}}$, где $S = \frac{s_+}{s}$ и $B = \frac{b_+}{b}$ – корректность распознавания, s_+ и b_+ – число событий, правильно распознанных как сигнал или фон; ошибку классификации $(1 - \max(p, 1-p))$ и прочие [8]. В реали-

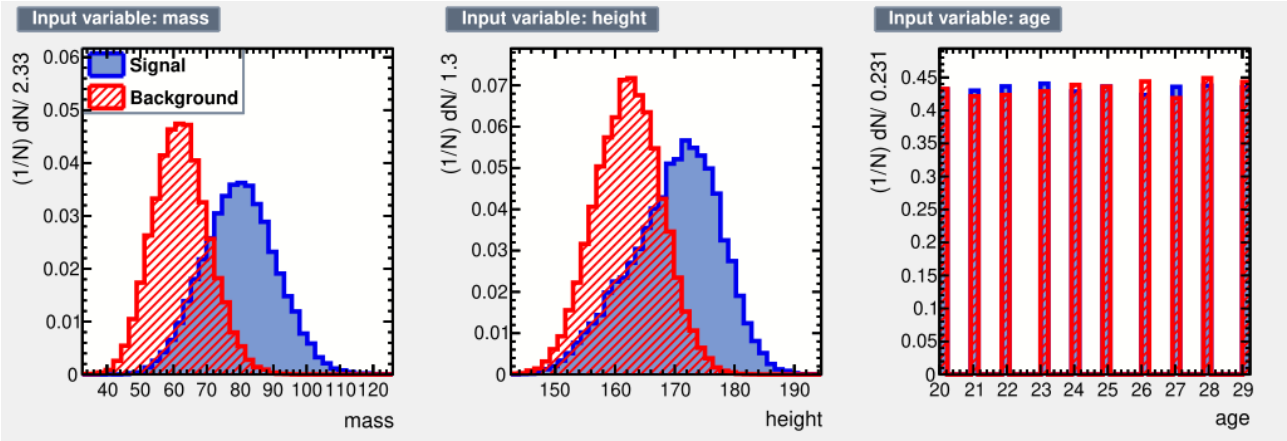

Рис. 4. Распределения входных данных. Сгенерировано 100000 событий. Переменная mass – масса частицы, height – амплитуда сигнала, age – номер слоя.

зованном алгоритме критерием разделения является минимальное значение $GiniIndex$. Максимальная глубина дерева равна трём, и для уменьшения статистических флуктуаций минимальная доля от всех событий на один узел выбрана равной 2,5%. Во всём ансамбле 100000 событий, на них тренируются 850 деревьев.

Характеристики алгоритма после обучения представлены на рис. 5 и рис. 6. Кривая ROC (receiver operating characteristic), сопоставляющая долю событий, отнесённых к фону, и долю распознанного сигнала, отображает качество алгоритма: лучше тот алгоритм, который при том же уровне разделения сохраняет больше всего сигнальных событий.

Статистическая значимость $\frac{S}{\sqrt{S+B}}$ показывает, насколько правильно распознаются события при данном пороговом значении отклика. На рис. 6 изображен график значимости, а также величины корректности распознавания сигнала S и фона B. По максимальному значению значимости определено оптимальное значение порога -0,02.

6. Алгоритм разделения кластеров

Задача алгоритма – классифицировать события, в которых образованный кластер включает в себя один фотон (сигнал) или два фотона (фон), то есть определить событие как сигнальное или фоновое, имея координаты

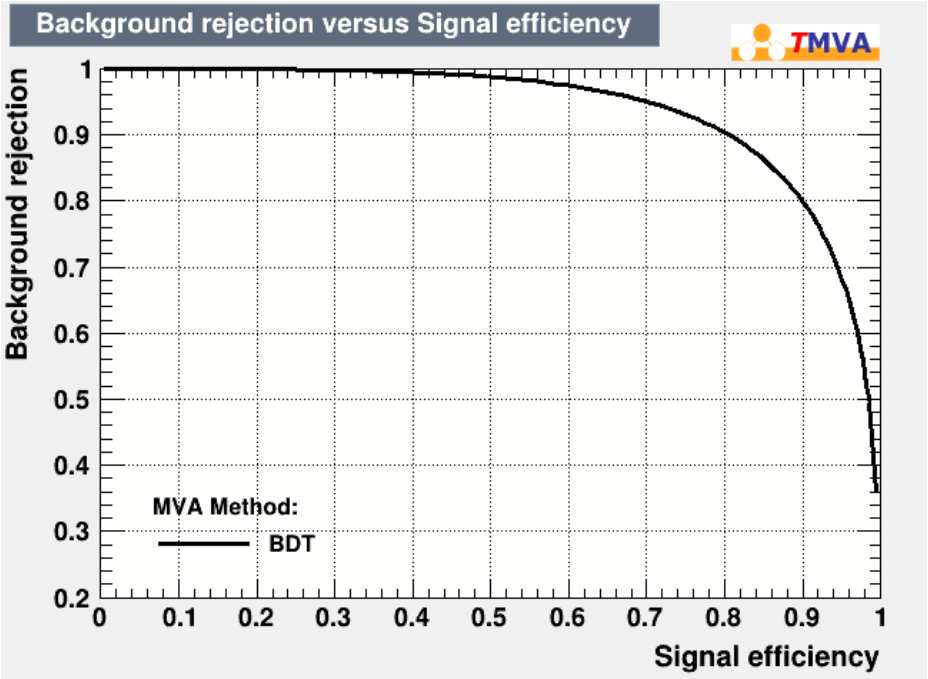

Рис. 5. Кривая ROC. По горизонтальной оси отложена эффективность по сигналу, по вертикальной оси – эффективность по фону.

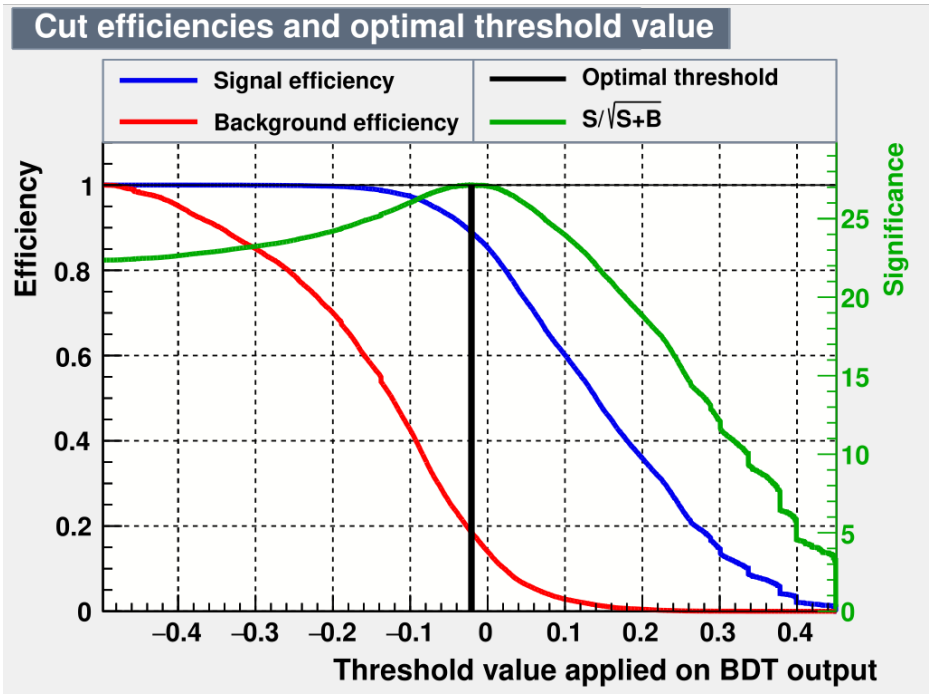

Рис. 6. График статистической значимости. Вертикальной линией обозначен выбранный критерий деления событий.

совокупности точек и связанные с ними значения энергии.

Для решения этой задачи в среде C++/ROOT с использованием библиотеки TMVA 4 выполнены следующие шаги:

1. реализация и проверка алгоритма разделения тестовых данных при помощи BDT,

2. подготовка данных моделирования процесса $e^+ e^- \to \pi^0 \gamma$ для дальнейших обучения и проверки алгоритма классификации.

Для подготовки данных моделирования требуется выделить события с однофотонными и двухфотонными кластерами. Для этого будут использоваться параметры моделируемых частиц. Наблюдая наличие кластеров вблизи моделируемых углов фотонов, можно оценить, перекрываются ли эти кластеры. Если фотоны разлетелись по углам далеко друг от друга и кластер от одного из фотонов не собрался, то в данной ситуации единственный кластер будет считаться сигнальным. В случае, когда фотоны летят близко, кластер единственный и находится в окрестности фотонов, это фоновое событие. На таком разделении будет тренироваться алгоритм BDT. Так же рассматривается способ определения количества фотонов в реконструированном кластере при помощи сравнения энергии фотона из генератора с полным энерговыделением в башенном кластере.

Фотон с энергией выше 10 MeV в основном взаимодействует с веществом через рождение электрон-позитронных пар в поле ядра. После рождения пары, электрон и позитрон испытывают потери от тормозного излучения (двигаясь с ускорением в поле ядра) и ионизации атомов. Таким образом, вторичные частицы (электроны, позитроны и фотоны) движутся далее вглубь материала, вызывая электромагнитный каскад или ливень. Увеличение числа заряженных частиц происходит до тех пор, пока их энергия не сравняется с критической E_c, далее ливень движется в основном за счёт γ-квантов с минимальным поглощением. Многократное рассеяние приводит к тому, что разброс по углам на каждом шаге рассеяния накапливается, и среднеквадратичный поперечный размер ливня, называемый радиусом Мольера R_m, равен

$$R_m = \frac{21}{E_c\ [MeV]} X_0,$$

где X_0 – радиационная длина, равная толщине, на которой энергия электронов падает в e раз. Продольный профиль ливня имеет форму конуса (рис. 7),

при этом энерговыделение в центре значительно больше, чем на краях. Если объять ливень цилиндром с радиусом $2R_m$, то внутри этого цилиндра будет содержаться 95% всей энергии ливня [9, 10].

Таким образом, при реализации алгоритма разделения кластеров планируется использовать такую информацию об электромагнитном ливне, как размеры кластера на конкретном слое, поперечное и продольное распределение энерговыделения, относительное смещение центров момента инерции между слоями и т.д.

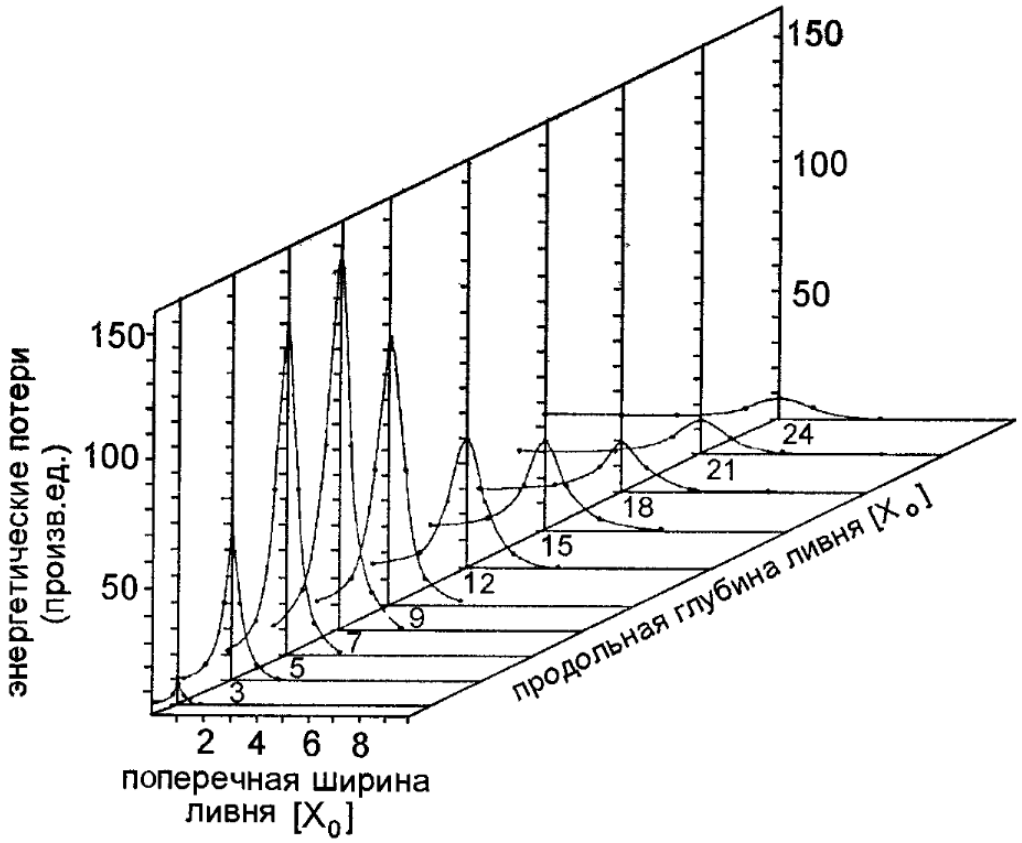

Рис. 7. Форма электромагнитного ливня. Размеры отложены в единицах радиационной длины X_0 [9].

7. Результаты

Полученные в результате обучения алгоритма деревья были применены к другому набору тестовых данных, сгенерированных тем же способом, что и тренировочный набор. В результате события классифицировались на сигнальные и фоновые. Сравнение результатов классификации с изначально

заданными типами событий представлено гистограммой на рис. 8.

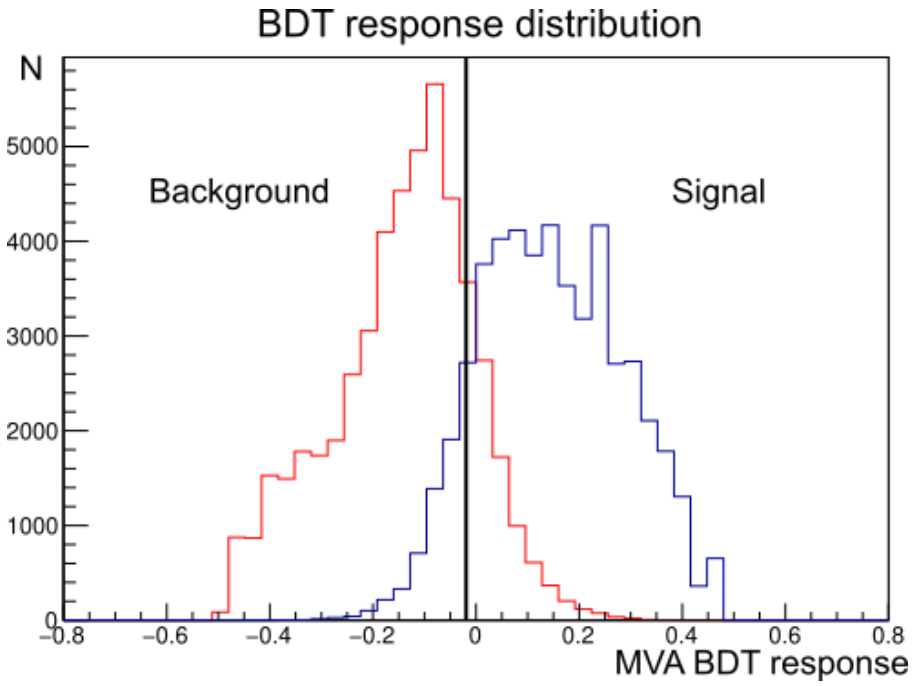

Рис. 8. Распределение отклика BDT на проверочный набор событий. Синим цветом обозначен сигнал, красным – фон. Вертикальной чёрной линией выделен порог отклика. Если отклик на событие ниже порога, оно распознаётся как фоновое, иначе – как сигнальное.

Из всех событий 50,4% классифицировались как сигнал, из них 85,3% правильно, а остальные 14,7% на самом деле являлись фоновыми. Из 49,6% распознанных фоновых событий действительно фоновыми были 86,1%, а остальные 13,9% при генерации задавались как сигнал. В итоге доля ложных распознаваний среди всех событий составляет 14,3%.

Таким образом, метод BDT является перспективным инструментом в разработке алгоритмов разделения, так как в этой задаче требуется анализ по многим переменным, а данный метод разработан на многомерную классификацию событий. Результатами работы являются оценка эффективности BDT для различения близких распределений, подготовка данных моделирования к дальнейшей работе и план реализации алгоритма разделения кластеров.

8. Заключение

В данной работе произведён анализ задачи разделения кластеров в электромагнитном жидкоксеноновом калориметре криогенного магнитного детек-

тора КМД-3. Проведена проверка работы алгоритма на основе метода BDT, реализованного в библиотеке TMVA 4, для тестового набора данных, и выявлена его эффективность. Разработан план реализации алгоритма классификации событий, в которых кластер порождён одним или двумя фотонами.

Достоинствами метода BDT являются возможность выделить много сегментов фазового пространства путём создания деревьев решений и уменьшить статистические флуктуации отклика при помощи бустинга. В дальнейшем планируется реализовать алгоритм разделения кластеров на подтипы в зависимости от количества образующих ливней, а также разработать методику реконструкции перекрывающихся кластеров.

Список литературы

[1] Yu. M. Shatunov *et al.* Project of a new electron positron collider VEPP-2000. *Conf. Proc.*, C0006262:439–441, 2000.

[2] Лакомов А. Е. Разделение близких фотонов в калориметре КМД-3. Выпускная квалификационная работа, НГУ, Новосибирск, 2019.

[3] Рабусов А. В. Изучение процесса аннигиляции e+e- - - -> eta gamma, eta - - -> pi+pi-pi0 с детектором КМД-3 на коллайдере ВЭПП-2000. Магистерская диссертация, НГУ, Новосибирск, 2016.

[4] Шебалин В. Е. *Реконструкция фотонов и энергетическая калибровка цилиндрического калориметра детектора КМД-3.* Диссертация на соискание учёной степени кандидата физико-математических наук, ИЯФ СО РАН, Новосибирск, 2016.

[5] Denis Epifanov. Electromagnetic calorimeters of the CMD-3 detector. *Journal of Physics: Conference Series*, 293:012009, April 2011.

[6] A. V. Anisyonkov *et al.* Liquid xenon calorimeter for a CMD-3 detector. *Nuclear Instruments and Methods in Physics Research Section A:*

Accelerators, Spectrometers, Detectors and Associated Equipment, 598:266–267, January 2009.

[7] A. V. Anisenkov *et al.* Status of the Liquid Xenon calorimeter of the CMD-3 detector. *Journal of Instrumentation*, 9(08):C08024–C08024, August 2014.

[8] A. Hoecker *et al. TMVA 4. Toolkit for Multivariate Data Analysis with ROOT. Users Guide.* Regents of CERN et al., 2005-2009.

[9] К. Групен. *Детекторы элементарных частиц.* Сибирский хронограф, Новосибирск, 1999.

[10] А. П. Онучин. *Экспериментальные методы ядерной физики.* Изд-во НГТУ, Новосибирск, 2010.