КРАТКИЕ СООБЩЕНИЯ

УДК 004.032.26 DOI: 10.14529/mmp230109

РАЗРАБОТКА АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОИСКА НОВЫХ РАСПАДОВ B_c^+ МЕЗОНОВ С ЧАРМОНИЕМ И МНОГОЧАСТИЧНЫМИ АДРОННЫМИ СОСТОЯНИЯМИ

А.В. Егорычев[1]***, Д.Ю. Перейма***[1]
[1]НИЦ «Курчатовский институт», г. Москва, Российская Федерация

В статье представлен процесс реализации алгоритма машинного обучения для классификации событий в физике высоких энергий. Приведены результаты тестирования классификатора на основе градиентного ускоренного дерева решений для улучшения эффективности отбора редких распадов B_c^+ мезонов с чармонием и многочастичными адронными состояниями. Разработка алгоритма выполнялась с применением пакета для многомерного анализа данных. Обучение классификатора основано на использовании данных математического моделирования и экспериментальных данных, набранных детектором LHCb на Большом адронном коллайдере в период с 2011 по 2018 гг.

Ключевые слова: многомерный анализ; машинное обучение; анализ данных; дерево решений; прелестные адроны; чармоний.

Введение

Машинное обучение (МО) – одна из форм искусственного интеллекта, которая позволяет компьютерным системам классифицировать те или иные явления на основе входных статистических данных без использования предварительных инструкций. Такой класс методов основан на выявлении общих закономерностей в анализируемых данных, а не на заранее описанных правилах для решения поставленных задач. Нейронные сети и прочие алгоритмы МО широко применяются в различных сферах анализа и классификации данных как в научно-исследовательских, так и в коммерческих сферах. В физике высоких энергий методы МО зачастую применяются для классификации сигнальных событий, представляющих интерес для дальнейшего изучения и анализа, и выделения их от фоновых, являющихся непременным атрибутом любого научного эксперимента.

Обучение – это процесс, с помощью которого алгоритм определяет различные признаки, присутствующие во входных данных, и пытается выявить в них общие характеристики. Такая процедура предполагает наличие известных условий для независимых переменных и известной модулируемой величины – зависимой переменной. В процессе обучения алгоритм находит взаимосвязь в данных, чтобы в последующем, основываясь на выделенных закономерностях и приобретенном «опыте», оценить прогнозы для модулируемой величины при новых условиях.

В представленной работе приводятся результаты разработки алгоритма МО для отбора событий от распадов B_c^+ мезонов на чармоний (J/ψ мезон) и многочастичные адронные состояния (каоны и пионы). Оценка эффективности отбора событий проводилась с использованием двух каналов распада $B_c^+ \to J/\psi K^+K^-\pi^+\pi^-\pi^+$ и $B_c^+ \to J/\psi 4\pi^+3\pi^-$, где J/ψ мезон восстанавливался в моде распада на пару противоположно заряженных мюонов. Сравнение эффективностей выделения сигнальных распадов проводилось между двумя конкурирующими способами: методом МО, построенном на классификации событий с помощью обучаемых деревьев решений

(Gradient Boosted Decision Trees, BDTG [1]), и традиционным методом анализа на основе ограничений. Традиционный метод базируется на применении порога к определенной переменной. Такой способ оставляет только те события, которые удовлетворяют условию ограничения. Если для отбора интересующих событий используют несколько переменных, то порог применяется к каждой из них.

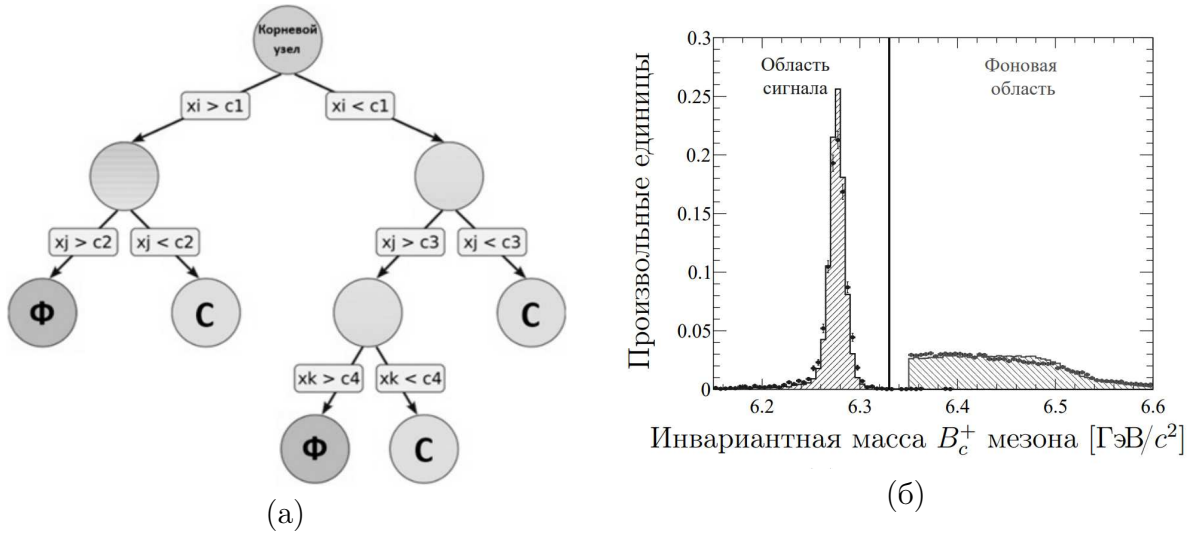
1. Обучение классификатора BDTG

Как правило, в физике высоких энергий классификатор BDTG используется для выделения редких сигналов из большого количества фоновых событий или применяется для идентификации физических объектов в спектрометре. На практике это достигается двумя способами. В одном случае применяется пороговое значение на переменную отклика классификатора. Во втором – используется форма отклика в качестве разделительной переменной для анализа. В ходе выполнения представленного исследования рассматривается первый метод.

Использование многомерного классификатора BDTG для разделения двух классов событий – сигнальных и фоновых, включает несколько этапов. На этапе обучения в используемых наборах данных, для которых известна принадлежность событий к тому или иному классу, должны быть определены т. н. дискриминирующие переменные – характеристики событий, которые значимо отличаются между двумя классами. На данном этапе строятся деревья решений (рис. 1 (а)), в которых каждая ветвь дерева отвечает условию на некоторое значение одной из дискриминирующих переменных, а конечные узлы дерева (листья) – весам, определяющим вероятность отнесения события к тому или иному типу. Обучением в данном случае является процесс, который определяет критерии классификации для всех событий в каждом узле. Такая процедура повторяется до тех пор, пока не будет сформировано все дерево и определены веса, при которых достигается наилучшее разделение между сигналом и фоном. Затем на этапе применения обученного классификатора к новым данным, в которых события не классифицированы, веса, получаемые из деревьев решений, позволяют сформировать одну переменную t, называемую откликом классификатора. Она строится таким образом, что ее распределения для обоих классов сигнальных и фоновых событий перекрываются минимально. Далее, применяя требование к переменной t для подавления фоновых событий, можно достичь лучшего отношения сигнала к фону (значимости сигнала), чем устанавливая пороговые значения на отдельные дискриминирующие переменные. В представленном анализе обучение классификатора BDTG [1] выполнялось с помощью инструмента для многомерного анализа данных (TMVA) [2], реализованного в пакете программ для анализа и обработки экспериментальных данных ROOT [3]. В физике высоких энергий зачастую невозможно использовать исследуемые данные в обучении, так как истинные метки для изучаемых процессов неизвестны. Поэтому в обучении, как правило, используются данные, полученные методом Монте-Карло – специальными симуляциями, которые реализуют современное понимание фундаментальных законов и принципов [4, 5].

Для использования классификатора BDTG на практике важно помнить, что прежде чем применять его к исследуемым данным, необходимо иметь хорошее согласие между экспериментальными данными и физической моделью, которая используется для их описания (модель распада, образование и рождение частиц, эффективность реконструкции и т. д.). Математическое моделирование данных в эксперименте LHCb можно разделить на две стадии: моделирование распадов частиц и моделирование детектора и взаимодействия частиц с ним. Первым этапом математического моделирования является генерация протон-протонных столкновений. Такая процедура выполняется с использованием пакета программ Pythia [6]. Моделирование

распадов частиц, содержащих тяжелые кварки, выполняется с использованием программы EvtGen [7]. Излучение фотонов в конечном состоянии добавляется в модель с помощью библиотеки Photos [8]. Взаимодействие частиц с детектором моделируется с использованием пакета программного обеспечения Geant4 [5]. Впоследствии, моделируемый сигнал проходит через все этапы реконструкции, что и физические данные.

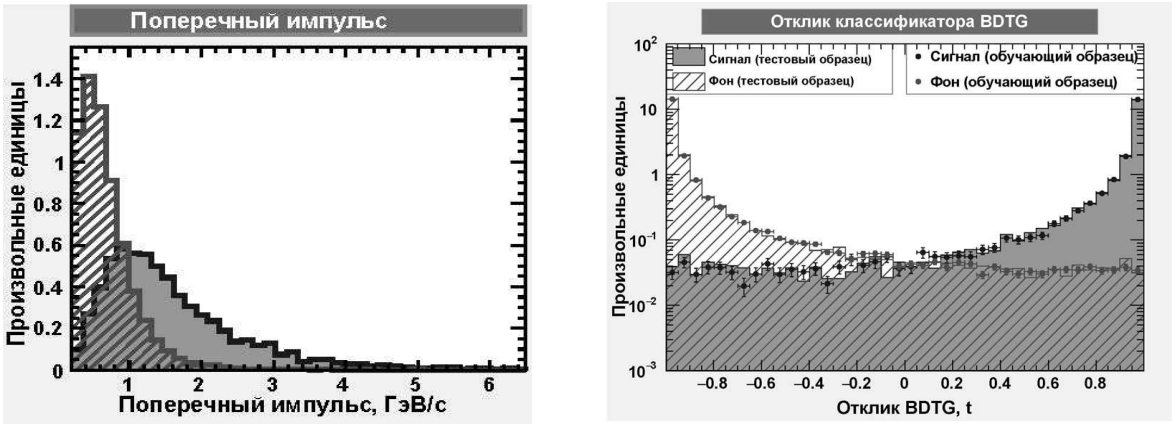


(а) (б)

Рис. 1. Схематический вид структуры дерева решений (а). Сигнальный и фоновый компоненты, используемые в обучении классификатора BDTG (б). Заштрихованные гистограммы обозначают входные данные для канала $B_c^+ \to J/\psi\, K^+K^-\pi^+\pi^-\pi^+$. Точками отображены данные для распада $B_c^+ \to J/\psi 4\pi^+3\pi^-$. Вертикальная линия разделяет сигнальную и фоновую область

Для классификации событий использовались два набора данных, описывающих сигнальный и фоновый компоненты изучаемых процессов. В качестве сигнального компонента отбирались истинные распады B_c^+ мезонов, полученные с помощью математического моделирования данных методом Монте-Карло. Для моделирования поведения фонового компонента использовались события из экспериментальных данных, расположенные в контрольном интервале за пределами сигнальной области исследуемых распадов. Распределения, иллюстрирующие исходные данные для обоих компонентов, используемых в обучении, показаны на рис. 1 (б). Количество событий для каждого канала распада, поступивших на вход классификатора, составило около 10000 для сигнального компонента и 65000 – для фонового. С помощью пакета TMVA входные данные разделялись на две выборки (обучающую и тестовую). Процедура обучения производилась по основным кинематическим и геометрическим переменным исследуемых распадов. Всего в обучении было задействовано девять переменных для канала $B_c^+ \to J/\psi\, K^+K^-\pi^+\pi^-\pi^+$ и одиннадцать переменных в канале $B_c^+ \to J/\psi 4\pi^+3\pi^-$. В качестве примера на рис. 2 (а) показана зависимость поперечного импульса каона для событий с участием B_c^+ мезонов и без них. Список дискриминирующих переменных, используемых в обучении, приведен в таблице.

Результат обучения сохранялся в табличные файлы формата XML. Тестовая выборка использовалась для промежуточной проверки результатов обучения. Далее полученный отклик классификатора извлекался из соответствующих баз данных и проецировался на выборку физических данных для последующего выделения сигнала и анализа. В разработанном алгоритме дерево решений состоит из 100 узлов, при этом



(а) (б)

Рис. 2. Распределение по поперечному импульсу заряженного каона (а) и распределение величины отклика классификатора BDTG (б). Показаны гистограммы для фонового компонента (заштрихованная гистограмма) и сигнального (закрашенная гистограмма). Представленные данные соответствуют распаду $B_c^+ \to J/\psi K^+K^-\pi^+\pi^-\pi^+$

Таблица

Список дискриминирующих переменных, используемых в многомерном классификаторе BDTG. В последней колонке указан порядковый номер переменной в порядке убывания ее важности в процедуре классификации

Переменная	Описание	Ранжирование важности
$\chi^2_{\text{DTF}}/\text{ndf}(B_c^+)$	параметр качества аппроксимации дерева распада	9
$c\tau(B_c^+)$	время жизни B_c^+ мезона	8
$p_{\text{T}}(\pi^{\pm})$	поперечные импульсы пионов	7
$\chi^2_{\text{IP}}(B_c^+)$	параметр, обозначающий, что B_c^+ мезон образован в первичной вершине протон-протонного соударения	6
$p_{\text{T}}(K^{\pm})$	поперечные импульсы каонов	5
$\chi^2_{\text{vtx}}(B_c^+)$	параметр качества аппроксимации вторичной вершины распада B_c^+ мезона	4
$p_{\text{T}}(J/\psi)$	поперечный импульс J/ψ мезона	3
$y(B_c^+)$	быстрота B_c^+ мезона	2
$\chi^2_{\text{IP}}(\text{tracks})$	параметр, обозначающий, что все дочерние треки образованы во вторичной вершине распада B_c^+ мезона	1

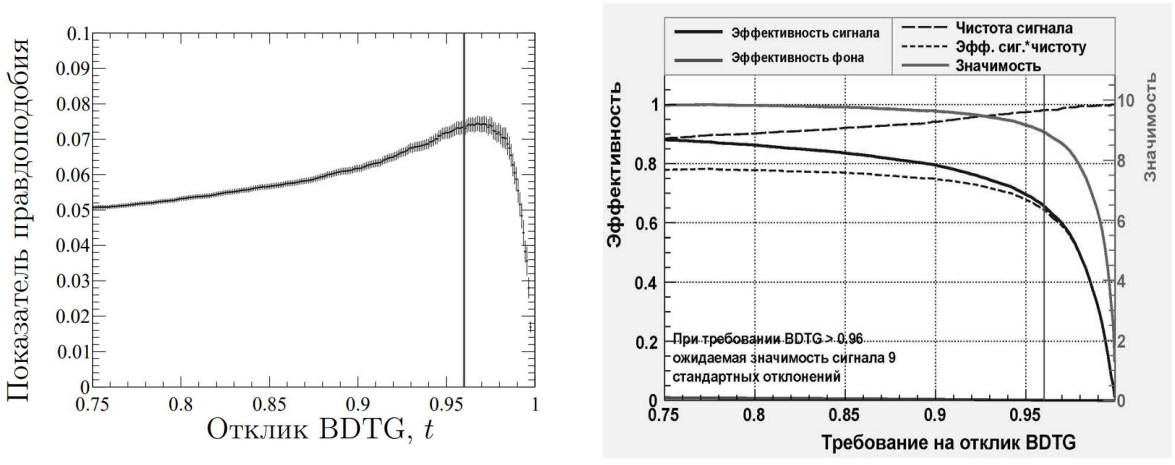
доля обучающих событий в каждом узле составляет около 2,5% от общего объема входных данных. Количество возможных разбиений узлов для поиска оптимального разделения сигнала от фона во входных данных равно 20. Обучение алгоритма проводилось таким образом, чтобы при запуске функции классификатора для тестируемого события для известного набора обучающих переменных, на выходе выдавалось значение-дискриминатор в диапазоне от -1 до 1. Распределения величины отклика BDTG классификатора для сигнальных и фоновых событий в тестовом и обучающем образцах данных представлено на рис. 2 (б). При этом чем ближе значение дискриминатора к единице, тем «ближе» изучаемое событие к искомому классу.

2. Поиск оптимальной рабочей точки на отклик классификатора и результаты

После обучения классификатора важным этапом анализа является поиск оптимального требования на отклик классификатора. Процедура оптимизации отбора сигнальных событий состояла в построении функции отклика, производящей наилучшее разделение сигнальных и фоновых событий, и в поиске оптимального требования отбора по переменной классификатора t, которое бы обеспечивало максимальную значимость сигнала в исследуемом образце данных. Оценка эффективности выделения сигнала осуществлялась с использованием следующей функции:

$$F(t) = \frac{\epsilon(t)}{\alpha/2 + \sqrt{B(t)}},$$

где ϵ – эффективность классификатора, определенная из данных математического моделирования, $\alpha = 5$ – требуемая значимость сигнала и B – количество фоновых событий в сигнальной области B_c^+ мезона, оцененное из экспериментальных данных. Выбор порогового значения $t = 0,75$ осуществлялся путем поиска максимальной производительности программного кода и минимизации временных затрат на процессорную обработку статистических данных. Зависимость показателя правдоподобия от отклика классификатора BDTG для канала $B_c^+ \to J/\psi K^+K^-\pi^+\pi^-\pi^+$ показана на рис. 3 (а). Поскольку точки в области максимума правдоподобия согласуются в пределах погрешностей, то для обеспечения высокой эффективности выделения сигнала, оптимальное требование отбора по переменной классификатора t установлено левее максимума оптимизационной кривой. Графическая характеристика качества классификатора, полученная из тестовых данных с помощью встроенных инструментов пакета TMVA, представлена на рис. 3 (б). Как видно из представленных зависимостей, положение оптимальной рабочей точки совпадает для двух рассмотренных методов оценки. При этом ожидаемая значимость сигнала в этом канале распада должна быть около девяти стандартных отклонений.



(а) (б)

Рис. 3. Вид оптимизационной кривой BDTG классификатора для распада $B_c^+ \to J/\psi K^+K^-\pi^+\pi^-\pi^+$ из экспериментальных (а) и тестовых (б) данных. Вертикальная сплошная линия иллюстрирует оптимальную рабочую точку

После нахождения оптимального требования на отклик классификатора BDTG и применения соответствующего ограничения для выделения сигнала были получены

распределения по инвариантной массе для изучаемых каналов. Чтобы определить количество сигнальных событий и статистическую значимость исследуемых распадов, производилась подгонка соответствующих распределений. Модель аппроксимации при этом состояла из суммы двух компонентов: сигнал и фон. Сигнальный компонент в распределении описывался с помощью модифицированной функции Гаусса [9]. Фоновый компонент аппроксимировался с помощью полинома первого порядка. Проекции распределений по инвариантной массе отобранных $B_c^+ \to J/\psi 4\pi^+ 3\pi^-$ и $B_c^+ \to J/\psi K^+ K^- \pi^+ \pi^- \pi^+$ кандидатов с наложенной функцией подгонки показаны на рис. 4. Выходы сигналов, оцененные из подгонки каждого канала, составили 19 ± 5 и 88 ± 12, где погрешности только статистические. Статистические значимости обнаруженных распадов (без учета возможных систематических эффектов) равны 5,6 и 10,7 стандартных отклонений соответственно. При этом значимость сигнала в канале $B_c^+ \to J/\psi K^+ K^- \pi^+ \pi^- \pi^+$ согласуется с ожидаемым значением (см. рис. 3 (б)). Полученные результаты в пределах погрешностей совпадают со значениями, оцененными из анализа без использования пакета TMVA [10]. Разработанный алгоритм классификации событий демонстрирует высокую эффективность выделения сигнала от редких распадов B_c^+ мезонов с многочастичными адронными состояниями, сравнимую с традиционными методами анализа. Таким образом, данный алгоритм можно использовать не только как основной метод выделения сигнальных распадов, но и как надежный инструмент для независимой проверки результатов.

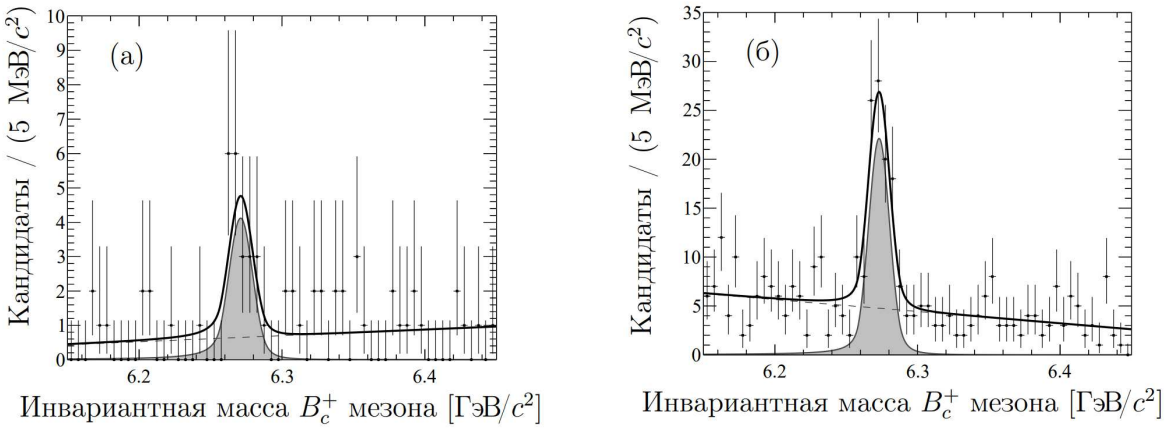


Рис. 4. Проекции распределений по инвариантной массе отобранных $B_c^+ \to J/\psi 4\pi^+ 3\pi^-$ (а) и $B_c^+ \to J/\psi K^+ K^- \pi^+ \pi^- \pi^+$ (б) кандидатов. Функция подгонки отображена сплошной линией, проходящей через точки экспериментальных данных. Закрашенная область соответствует наблюдаемому сигналу

Заключение

С использованием пакета для многомерного анализа данных TMVA был разработан алгоритм классификации для поиска и выделения сигнальных событий от редких распадов B_c^+ мезонов с чармонием и многочастичными адронными состояниями. Реализация алгоритма основана на методе классификации с помощью градиентного ускоренного дерева решений. Эффективность классификатора отбора была протестирована на данных для двух редких каналов распада $B_c^+ \to J/\psi K^+ K^- \pi^+ \pi^- \pi^+$ и $B_c^+ \to J/\psi 4\pi^+ 3\pi^-$. В ходе исследования было установлено, что при оптимальном требовании на отклик классификатора наблюдаемые сигналы от изучаемых распадов и их значимости находятся в согласии с результатами традиционного анализа, пред-

ставленного в работе [10]. Благодаря проведенному исследованию впервые удалось экспериментально обнаружить новый канал распада $B_c^+ \to J/\psi\, K^+K^-\pi^+\pi^-\pi^+$. Также выполнено экспериментальное наблюдение канала $B_c^+ \to J/\psi\, 4\pi^+3\pi^-$, что является первым свидетельством существования распада тяжелого B_c^+ мезона в конечное состояние с девятью частицами. Полученные результаты важны для последующих исследований по оптимизации и выделению редких распадов частиц, содержащих тяжелые кварки. Таким образом, реализованный алгоритм в дальнейшем может найти широкое применение в других исследованиях по поиску редких распадов тяжелых B_c^+ мезонов.

Работа проводилась при финансовой поддержке Совета по грантам Президента Российской Федерации для государственной поддержки молодых российских ученых и по государственной поддержке ведущих научных школ Российской Федерации. Номер гранта: МК-894.2022.1.2.

Литература/References

1. Breiman, L. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen et al. – Belmont: Wadsworth International Group California, 1984.
2. Höcker, A. TMVA – Toolkit for Multivariate Data Analysis / A. Höcker, J. Stelzer, F. Tegenfeldt et al. – URL: http://cds.cern.ch/record/1019880 (дата обращения: 01.09.2022)
3. Oficial Website of the ROOT Package root.cern.ch. – URL: https://root.cern.ch/root/htmldoc/guides/users-guide/ROOTUsersGuideA4.pdf (дата обращения: 01.09.2022)
4. Clemencic, M. The LHCb Simulation Application, Gauss: Design, Evolution and Experience / M. Clemencic // Journal of Physics. – 2011. – № 331. – Article ID: 032023. – 7 p.
5. Agostinelli, S. Geant4 – a Simulation Toolkit / S. Agostinelli, J. Allison, K. Amako et al // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. – 2003. – V. 506, № 3. – P. 250–303.
6. Sjöstrand, T. A Brief Introduction to PYTHIA 8.1 / T. Sjöstrand, S. Mrenna, P. Skands // Computer Physics Communications. – 2008. – V. 178. – P. 852–867.
7. Lange, D.J. The EvtGen Particle Decay Simulation Package / D.J. Lange // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. – 2001. – V. 462. – P. 152–155.
8. Golonka, P. Photos Monte Carlo: a Precision Tool for QED Corrections in Z and W Decays / P. Golonka, Z. Was // The European Physical Journal. – 2006. – V. 45. – P. 97–107.
9. Skwarnicki, T. A Study of the Radiative Cascade Transitions between the Υ' and Υ Resonances / T. Skwarnicki. – URL: https://inspirehep.net/literature/230779 (дата обращения: 01.09.2022)
10. Aaij, R. Study of the B_c^+ Meson Decays to Charmonia Plus Multihadron Final States / R. Aaij. – URL: https://arxiv.org/abs/2208.08660 (дата обращения: 01.09.2022)

Артем Викторович Егорычев, инженер, лаборатория нейтринной физики, Курчатовский институт (г. Москва, Российская Федерация), Artem.Egorychev@cern.ch.

Дмитрий Юрьевич Перейма, кандидат физико-математических наук, старший научный сотрудник, лаборатория нейтринной физики, Курчатовский институт (г. Москва, Российская Федерация), Dmitrii.Pereima@cern.ch.

Поступила в редакцию 28 декабря 2022 г.

MSC 68T07 **DOI: 10.14529/mmp230109**

DEVELOPMENT OF A MACHINE LEARNING ALGORITHM FOR THE SEARCHES OF THE NEW B_c^+ MESON DECAYS WITH CHARMONIUM AND MULTIHADRON FINAL STATES

A. V. Egorychev[1]***, D. Yu. Pereima***[1]
NRC "Kurchatov Institute", Moscow, Russian Federation
E-mail: Artem.Egorychev@cern.ch, Dmitrii.Pereima@cern.ch

The paper describes the process of implementation of machine learning algorithm for the classification of the events in high energy physics. The results of testing a classifier based on gradient boosted decision tree to improve the selection efficiency of the rare B_c^+ meson decays with charmonium and multihadron final states are presented. The development of the algorithm is performed using a toolkit for multivariate data analysis. The training of the classifier is based on the simulated data and experimental data, collected by the LHCb detector at the Large Hadron Collider in the period from 2011 to 2018.

Keywords: multivariate analysis; machine learning; data analysis; decision tree; beauty hadrons; charmonium.

Received December 28, 2022