Е.Б.Постников

ПРИМЕНЕНИЕ МНОГОМЕРНЫХ МЕТОДОВ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ К РЕШЕНИЮ ЗАДАЧ КЛАССИФИКАЦИИ ЧАСТИЦ ПЕРВИЧНОГО КОСМИЧЕСКОГО ИЗЛУЧЕНИЯ

Препринт НИИЯФ МГУ - 2004-23/762

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. М.В. ЛОМОНОСОВА

НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
ЯДЕРНОЙ ФИЗИКИ им. Д.В. СКОБЕЛЬЦЫНА

Е.Б.ПОСТНИКОВ

ПРИМЕНЕНИЕ МНОГОМЕРНЫХ МЕТОДОВ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ К РЕШЕНИЮ ЗАДАЧ КЛАССИФИКАЦИИ ЧАСТИЦ ПЕРВИЧНОГО КОСМИЧЕСКОГО ИЗЛУЧЕНИЯ

Препринт НИИЯФ МГУ - 2004-23/762

УДК 519.237, 524.1
ББК 22.38

Evgeniy Postnikov
E-mail: postn@eas.sinp.msu.ru

Application of Pattern Recognition Multidimensional Technique to the Classification of Primary Cosmic Rays Particles

Preprint of SINP MSU - 2004-23/762

Abstract

The subject of the paper is an application of the pattern recognition theory based on multidimensional statistical approach to solving the classification problem appearing in cosmic rays physics. The paper fully describes the algorithm of the primary cosmic particles separation into 2 classes, developed on basis of the Bayesian linear classifier. As an illustration the practical computational problem of the primary particles separation by location of their first interaction point within the device (spectrometer of the NUCLEON scientific project) is solved with the aid of the suggested algorithm using model data. As distinct from the primary energy determination problem in the above-mentioned NUCLEON project, the practical computational problem under consideration cannot be solved by any method, even approximate one, within the bounds of the deterministic (not statistical) approach.

In addition, a wide range of the experimental cosmophysics problems to be solved by the technique suggested is briefly outlined.

Е.Б. Постников

Применение многомерных методов теории распознавания образов к решению задач классификации частиц первичного космического излучения

Препринт НИИЯФ МГУ - 2004-23/762

Аннотация

Работа посвящена применению базирующейся на многомерном статистическом подходе теории распознавания образов к решению задач классификации первичных частиц космического излучения. Подробно описывается разработанный на основе байесовского линейного классификатора алгоритм разделения частиц первичного космического излучения на два класса. В качестве иллюстрации предложенным методом на основе модельных данных решена практическая задача разделения первичных частиц по местоположению точки первого взаимодействия в блоке установки измерительной аппаратуры спектрометра (проект NUCLEON). В отличие от задачи определения первичной энергии в том же проекте NUCLEON, данная практическая задача не может быть решена никакими, даже приближенными, методами в рамках детерминистического, а не статистического, подхода.

В работе также очерчен весьма широкий круг задач экспериментальной космофизики, для решения которых может быть применена предлагаемая методика.

Введение

При анализе и обработке экспериментальных либо модельных данных в физике космических лучей (КЛ) часто возникают довольно однотипные задачи *классификации*. Например, на основе статистического анализа данных, несущих только косвенную информацию о первичной частице, таких, как ион-потери в слоях толчковой установки или сигналы с матрицы стрипового детектора, часто требуется определить значение какой-либо дискретной характеристики первичной частицы. Возможность отнесения подобных задач к задачам классификации обусловлена тем, что определение истинного значения дискретной физической характеристики χ первичной частицы среди дискретного множества возможных значений: $\chi \in \{\chi_1, \chi_2, \ldots, \chi_m\}$ полностью эквивалентно классификации этой частицы как принадлежащей к одному из m типов или классов: *первому* классу – классу частиц со значением $\chi=\chi_1$; *второму* классу - классу частиц, для которых $\chi=\chi_2$;... , *m-му* классу – классу частиц с $\chi=\chi_m$.

В качестве поддающейся классификации физической характеристики первичной частицы может выступать любая ее дискретная характеристика, например: заряд; магнитная жесткость; номер сектора пространства, откуда прибыла частица, или блока измерительной аппаратуры, в котором произошло взаимодействие частицы, и т.д.

Как известно [1], применяемая для решения задач классификации классическая теория распознавания образов обеспечивает наивысшую точность классификации в том случае, если требуется разделить исследуемые объекты всего по *двум классам*, т.е. когда

дискретная случайная величина χ может принимать только два значения: $\chi=\chi_1$ или $\chi=\chi_2$. Одним из наиболее простых и в то же время надежных (а именно, обеспечивающим минимальное значение вероятности ошибки [1]) методов классификации в случае двух классов является *байесовское решающее правило*, о применении которого к решению космофизических задач и будет рассказано в данной работе.

Отметим, что в экспериментальной космофизике круг задач, к которым применима простая классификация объектов по двум классам, очень широк. Чаще всего именно к одному из двух возможных типов необходимо отнести первичную частицу по результатам проведенных измерений. Например, для того чтобы выделить среди всех сортов первичных частиц именно *протоны*, можно ввести два формальных класса объектов-частиц: *протоны* и «*не протоны*». Подобные два класса следует вводить таким образом, чтобы задача их распознавания представляла собой физически наиболее сложную задачу, для решения которой целесообразно привлечь мощный аппарат теории распознавания образов. Как правило, в любой практической ситуации исследователь легко может найти такие два интересующие его класса. Полная же «идентификация» первичных частиц, т.е. разделение их по всем возможным в данной ситуации классам или, что то же самое, по всем возможным значениям случайной величины χ, как правило, успешно решается более простыми средствами и способами, традиционно используемыми в физике космических лучей (например, для разделения частиц по зарядам существует детектор заряда, позволяющий с приемлемой точностью классифицировать частицы

по *модулю* их электрического заряда. Поэтому наиболее сложными задачами анализа и интерпретации получаемых подобным детектором и другими регистраторами в составе измерительной аппаратуры данных являются такие задачи, как отделение протонов от электронов, т.е. частиц одинакового по модулю заряда, и т.д.)

Если те данные, которые имеются в нашем распоряжении для решения подобных задач экспериментальной физики космических лучей, являются *многомерными*, т.е. измерительная аппаратура регистрирует значения *нескольких* параметров для каждой первичной частицы, для этого случая можно предложить воспользоваться многомерной методикой распознавания образов. В настоящей работе будет подробно представлена идеология, техника и конкретные результаты применения для решения космофизических задач классификации *методики байесовского решающего правила* [1]. Успешное применение этого многомерного статистического алгоритма в рассматриваемой области экспериментальных физических исследований представляет собой продолжение внедрения в обработку результатов космофизических экспериментов современной многомерной статистической техники, начатого в работах [2-4] с использования для определения первичной энергии и спектрального индекса статистического метода линейного наилучшего в среднеквадратичном оценивания случайных векторов. Преемственность этих исследований хорошо прослеживается в силу сходства общей идеологии подхода к постановке задач анализа и интерпретации данных.

Описание методики

Вводная часть описания

Пусть наша регистрирующая аппаратура может измерять о каждой первичной частице многомерную информацию, т.е. значения не одного, а нескольких физических параметров. Следуя принятой в [4] терминологии, назовем эти величины *измеряемыми переменными*. По зафиксированным аппаратурой значениям измеряемых переменных, во-первых, какой-либо методикой определяется сама первичная энергия; а во-вторых, требуется решить задачу классификации частиц как принадлежащих к одному из двух заранее известных типов. Примером подобной схемы измерений может служить любая аппаратура для регистрации КЛ, в составе которой присутствует стриповый детектор или другой многоканальный регистратор.

Для применения статистических многомерных методик работы с данными следует объединить все эти измеряемые переменные в *случайный вектор*, который мы обозначим через ξ. Каждый акт регистрации первичной частицы нашей аппаратурой предоставляет в наше распоряжение очередную реализацию случайного вектора ξ. Когда алгоритм решения задачи классификации первичных частиц будет построен, мы сможем, в зависимости от значения конкретной реализации вектора ξ, относить первичную частицу к одному из двух заранее известных классов, которые мы обозначим *класс I* и *класс II*.

При решении задачи *определения первичной энергии* E по регистрируемым значениям измеряемых переменных реализовывалась статистическая схема

линейного оценивания энергии E, трактуемой как случайная величина, по измерению случайного вектора ξ. Для решения принципиально иной задачи *классификации первичных частиц* формируется другая статистическая схема, оперирующая, однако, теми же понятиями и использующая в своем алгоритме те же статистические объекты, что и названная выше задача, решение которой подробно описано в работах [2-4].

Согласно байесовскому решающему правилу, классификация производится на основе скалярной функции от случайного вектора измеряемых переменных ξ, называемой *байесовским классификатором* и обозначаемой $t(\xi)$. Вид этой функции может быть выявлен из статистического анализа данных по измеряемым переменным, экспериментальным либо модельным, но обязательно уже классифицированным, т.е. по такому банку данных, для каждого события из которого точно известно, к какому из двух классов, классу I или классу II, относится зафиксированная в этом событии первичная частица. Именно такое требование к банку данных, на основе анализа которого предполагается сконструировать методику классификации, позволяет произвести процедуру «обучения» методики и «научить» ее распознавать тип частицы, используя информацию о различиях между статистическими распределениями частиц первого и второго классов, извлеченную методикой распознавания образов из банка данных. Поэтому названный банк данных называется *обучающей выборкой*. Он может быть получен, например, при помощи компьютерного моделирования с использованием хорошо известного в ядерной физике и физике космических

лучей программного комплекса GEANT [6] симуляции по методу Монте-Карло процесса взаимодействия элементарных частиц с веществом.

После того, как вид байесовского классификатора $t(\xi)$ определен, его значение может быть вычислено для любого события регистрации первичной частицы рассматриваемой измерительной аппаратурой, как для априори известного, так и для априори неизвестного класса частицы. В каждом событии вычисленное значение $t(\xi)$ сравнивается со значением *порога* ε, постоянной (неслучайной) величины, значение которой также определяется по обучающему банку данных. Классификация происходит стандартным образом: если значение $t(\xi)$ оказывается большим порога, первичную частицу следует отнести к первому классу; если меньше - ко второму классу. Случай строго равенства $t(\xi)=\varepsilon$ можно произвольно отнести к первому либо второму классу; он непринципиален, т.к. вероятность его ничтожно мала.

Вычисление классификатора на банке данных, отличном от обучающего, но таком, для которого, как и для обучающего, точно известно, к какому из двух классов относится каждое содержащееся в банке событие, позволяет оценить погрешность сформированной методики классификации. Такой позволяющий достоверно оценить погрешность методики банк данных называется *контрольной выборкой*. Для оценивания величины погрешности метода классификации по этому банку делается предположение о том, что классификация частиц из контрольной выборки неизвестна, и проводится их классификация по обрисованному выше алгоритму. После этого определяется количество ошибочно классифицированных частиц,

на самом деле априорно принадлежащих не тому классу, к которому они были отнесены методикой распознавания. Долей ошибочно классифицированных частиц среди общего числа частиц в каждом из двух классов определяются величины каждой из двух возможных *ошибок классификации*: ошибочной классификации частицы, на самом деле относящейся к классу I; и ошибочной классификации частицы, истинным классом которой является класс II. Если контрольная выборка, аналогично обучающей, представляет собой выборку из модельных, а не экспериментальных данных, то в случае достаточно надежной модели и однородного во времени физического процесса (статистические параметры теоретико-вероятностного описания которого не изменяются с течением времени), естественно ожидать, что при применении разработанной и протестированной на контрольной выборке методики классификации уже к реальным экспериментальным данным погрешности классификации будут иметь такую же величину.

Вычислительные формулы

Конкретный вид байесовского классификатора (в нормальном приближении) определяется по следующей формуле:

$$t(\xi) = 0{,}5(\xi-M_1)^T F_1^{-1} (\xi-M_1) - 0{,}5(\xi-M_2)^T F_2^{-1} (\xi-M_2) + \\ + 0{,}5\ln(||F_1|| / ||F_2||); \qquad (1)$$

где:

- M_1 и M_2 – векторы математических ожиданий случайного вектора ξ по распределению первичных частиц только первого и только второго классов соответственно;

- F_1 и F_2 – ковариационные матрицы случайного вектора ξ по распределению первичных частиц первого класса и распределению первичных частиц второго класса соответственно;
- значок «Т» обозначает транспонированную матрицу (в данном случае – транспонированную матрицу-столбец, то есть матрицу-строку);
- двойные прямые скобки обозначают определитель матрицы;
- «-1» в правом верхнем углу над символом матрицы обозначает обратную матрицу.

Байесовский классификатор в таком виде является нелинейной (квадратичной) функцией от измеряемых переменных. Он учитывает различие двух многомерных распределений измеряемых переменных, соответствующих классу I и классу II, *не только по их средним значениям* (векторам математических ожиданий M_1 и M_2), *но и по значениям их дисперсий и корреляций*, входящих в выражения для матриц F_1 и F_2.

Еще более простое и уже линейное по измеряемым переменным – координатам вектора ξ – выражение получается для байесовского классификатора в случае, если для разделения первичных частиц по двум классам предполагается использовать *только* различие статистических распределений первого и второго классов *по средним значениям* измеряемых переменных. В этом случае мы делаем предположение, что величины дисперсий измеряемых переменных и корреляционные связи между ними одинаковы для всех классов, различаются же эти классы только значениями векторов M_1 и M_2. Обозначив общую ковариационную матрицу случайного вектора ξ

через F, перепишем упрощенное выражение для байесовского классификатора в виде:

$$t(\xi) = (M_2 - M_1)^T F^{-1} \xi + 0{,}5(M_1{}^T F^{-1} M_1 - M_2{}^T F^{-1} M_2). \quad (2)$$

Выражение для порога ε в обоих случаях имеет один и тот же вид:

$$\varepsilon = \varepsilon_0 = \ln(P_1/P_2), \quad (3)$$

где P_1 и P_2 – вероятности появления среди всего потока первичных частиц на входе измерительной аппаратуры частиц первого и второго классов соответственно.

Отметим, что при увеличении или уменьшении значения порога по сравнению с величиной $\ln(P_1/P_2)$ суммарная ошибка классификации будет увеличиваться, однако при этом также будет изменяться соотношение между вероятностями ошибочного отнесения к другому классу первичных частиц из класса I и первичных частиц из класса II. Поэтому, если из физических соображений рассматриваемые нами классы «неравноценны» (например, уже упоминавшиеся протоны и «не протоны»), соответственно, неравноценны для нас и погрешности ошибочного отнесения первичных частиц к первому и второму классу (например, более важной для нас может являться задача исключения из банка данных всех «не протонов», чем задача идентификации максимального количества протонов, в число которых могут попасть и ошибочно классифицированные частицы других типов).

На этом основании в работе будет использоваться экспериментальный метод определения значения порога ε как того оптимального значения, при котором соотношение между ошибками классификации на классе I и на классе II будет максимально со-

ответствовать априорным представлениям исследователя. Значение порога ε согласно предлагаемому алгоритму пробегает последовательно весь интервал значений классификатора $t(\xi)$ от минимального до максимального, что позволяет для каждого ε определить ошибку классификации на классе I и на классе II и построить зависимость одной ошибки от другой.

Экспериментальный метод определения порога ε позволяет также упростить формулы для вычисления классификатора (1) и (2). Поскольку значение $t(\xi)$ будет теперь сравниваться не с жестко определенной величиной порога (3), а с произвольной константой, пробегающей весь диапазон изменения $t(\xi)$, то из формул (1) и (2) можно убрать все постоянные (неслучайные) слагаемые, т.е. все члены, не зависящие от ξ. В результате получим упрощенную формулу для вычисления классификатора с учетом различия ковариационных матриц обоих классов:

$$t(\xi) = \xi^T (F_1^{-1} - F_2^{-1})\xi - 2(F_1^{-1}M_1 - F_2^{-1}M_2)\xi; \qquad (4)$$

и упрощенную формулу для $t(\xi)$ в предположении равенства ковариационных матриц разных классов:

$$t(\xi) = (M_2 - M_1)^T F^{-1} \xi. \qquad (5)$$

После процедуры вычисления значения классификатора $t(\xi)$ и задания начального значения порога ε для каждого конкретного события из банка контрольной выборки принятие решения об отнесении рассматриваемого события к первому или второму классу производится следующим образом:

- если $t(\xi) < \varepsilon$ - первичная частица относится к классу I;
- если $t(\xi) > \varepsilon$ - первичная частица относится к классу II. (6)

При заданном значении ε по этому алгоритму классифицируем все частицы из рассматриваемого банка данных и определяем ошибки классификации. Далее задаем следующее значение ε и повторяем всю процедуру. После того, как будет исследован весь диапазон возможных значений ε, мы получим зависимости ошибок классификации от ε, а также зависимость, выражающую соотношение между ошибкой по отнесению первичных частиц к первому классу от ошибки по отнесению ко второму классу. Анализ полученных зависимостей позволит определить оптимальные значения порога ε, которые и будут использоваться при классификации первичных частиц уже по экспериментальным, а не модельным, данным.

Для наглядности в качестве результатов применения методики распознавания образов для решения задачи классификации первичных космических частиц можно предложить построение таблицы, аналогичной построенной ниже таблице 1, либо кривой зависимости соотношения между обеими ошибками классификации от значения порога ε (пример построения подобной кривой также приводится ниже, на рисунке 1). Используя данную кривую в качестве «калибровочной», исследователь сможет выбрать оптимальное пороговое значение, исходя из общей погрешности классификации и соотношения между ошибками классификации по первому и по второму классу частиц.

Оценивание статистических характеристик распределений

Для того чтобы применить реализующие идентификацию первичных частиц методику байесовского

решающего правила формулы (3)-(5) на практике, необходимо оценить по обучающему банку данных все входящие в эти формулы неизвестные величины, а именно, координаты векторов M_1 и M_2 и элементы матриц F_1 и F_2. Оценивание происходит по стандартной статистической схеме, позволяющей получить несмещенные оценки математических ожиданий, дисперсий и ковариаций. Для наглядности приведем формулы строго определения оцениваемых величин и формулы оценивания их значений по конечному банку данных.

Согласно определениям вектора математического ожидания и ковариационной матрицы, имеем:

$$(M_i)_{теор} = M_i\xi,\ i=1,2;$$

$$(F_i)_{теор} = M_i(\xi - M_i\xi)(\xi - M_i\xi)^T,\ i=1,2;$$

$$(F)_{теор} = M(\xi - M\xi)(\xi - M\xi)^T; \qquad (7)$$

где:

- M_i обозначает математическое ожидание по распределению первичных частиц только i-го класса;
- M – математическое ожидание по распределению всех первичных частиц (смеси обоих классов).

Формулы для несмещенного оценивания векторов и матриц моментов распределения измеряемых переменных:

$$M_i = \langle\xi\rangle_i,\ i=1,2;$$

$$F_i = \frac{N_i}{N_i - 1}\langle(\xi - \langle\xi\rangle_i)(\xi - \langle\xi\rangle_i)^T\rangle_i,\ i=1,2;$$

$$F = \frac{N}{N-1}\langle(\xi - \langle\xi\rangle)(\xi - \langle\xi\rangle)^T\rangle; \qquad (8)$$

где:

- угловые скобки с индексом символизируют среднее арифметическое значение по части обучающей выборки, относящейся к классу с номером i;
- угловые скобки без индекса – среднее арифметическое по всей обучающей выборке (смеси обоих классов);
- N_i – объем части обучающей выборки из i-го класса;
- N – объем всей обучающей выборки ($N = N_1 + N_2$).

Множители $\frac{N_i}{N_i - 1}$ и $\frac{N}{N-1}$ вводятся в формулы для выборочных оценок матриц ковариаций по той причине, что *свойству несмещенности оценок* – равенства математического ожидания оценки как случайной функции от выборки истинному значению оцениваемого момента распределения – удовлетворяют не обычные арифметические средние, а суммы из N слагаемых с весовым коэффициентом N–1, например:

$$F = \frac{1}{N-1}\sum_{j=1}^{N}(\xi - \langle\xi\rangle)(\xi - \langle\xi\rangle)^T = \frac{N}{N-1}\frac{1}{N}\sum_{j=1}^{N}(\xi - \langle\xi\rangle)(\xi - \langle\xi\rangle)^T. \qquad (9)$$

Хотя на больших выборках отличие сумм с данным весовым коэффициентом от арифметических средних значений несущественно, тем не менее, именно в виде (8) оценки всех входящих в вычислительные формулы (4), (5) моментов распределения измеряемых переменных обладают этим важным свойством несмещенности.

Наконец, оценки значений вероятностей P_1 и P_2 из выражения (3) для порога ε вычисляются по данным файла обучающей выборки очевидным образом

как доли числа частиц 1 и 2 класса среди суммарного количества частиц в выборке:

$P_i = N_i / N$, i=1,2.

Вычислительный эксперимент

Физическая задача

Приведем результаты модельного вычислительного эксперимента, осуществленного для решения одной из задач в рамках общей практической задачи оптимизации измерительной аппаратуры проекта NUCLEON – спектрометра в составе этой аппаратуры [5]. Цель названного проекта состоит в создании компактной аппаратуры относительно небольшого веса для регистрации КЛ (протонов и ядер) в широком энергетическом диапазоне от 10^{11} до 10^{15} эВ. С использованием аппаратуры NUCLEON энергия первичных частиц определяется на основе анализа пространственной плотности распределения потока вторичных частиц, рожденных в тонкой *мишени* (первый акт неупругого взаимодействия) и размноженных в сверхтонкой толчковой установке – *конверторе*.

Для измерения пространственной плотности вторичных частиц в аппаратуре NUCLEON предназначены две перпендикулярно друг по отношению к другу ориентированные матрицы кремниевых стриповых детекторов. Величина I_i сигнала в каждом стрипе (с номером i) детектора пропорциональна ионизационным потерям в этом стрипе. В данном случае именно совокупность сигналов I_i со всех стрипов обеих матриц детектора, одновременно измеряемых в акте регистрации одного события прохождения первичной частицы КЛ через аппаратуру, и составляет в

нашей статистической интерпретации измерительной схемы случайный вектор измеряемых переменных ξ. Количество одновременно регистрируемых измеряемых переменных (т.е. размерность вектора ξ) в нашей задаче очень велико – порядка нескольких тысяч, поэтому задача интерпретации результатов измерения аппаратуры NUCLEON является, как это отмечалось в [2-4], существенно многомерной.

Перед мишенью на пути следования первичной частицы через апертуру прибора в аппаратуре NUCLEON предполагается расположить один или несколько детекторов заряда, которые позволят успешно решать задачу разделения регистрируемых частиц по значению заряда частиц. Эксперименты с компьютерной моделью аппаратуры подтверждают данное предположение [7]. Тем не менее, проблема классификации первичных частиц в измерительной модели аппаратуры NUCLEON возникла в связи с повышением точности энергетического разрешения спектрометра. Дело в том, что некоторая доля из числа всех предполагающихся к регистрации прибором первичных частиц может провзаимодействовать не в мишени, а в выполняющем функцию сверхтонкой толчковой установки конверторе.

Поскольку типы веществ (мишень состоит из углерода, конвертор – из вольфрама), толщина мишени и конвертора и расстояние их до плоскости стрипового детектора различны, а кроме того, начавший развиваться только в конверторе вторичный каскад не успеет размножиться в такой же степени, как каскад, образовавшийся в мишени, - по всем этим причинам пространственное распределение вторичных частиц от взаимодействий в мишени и в конвер-

торе будет иметь разный характер. Поэтому обработка одинаковым алгоритмом, как одномерным методом, использующим эмпирический параметр S [5], так и многомерной методикой [2-4], всего банка событий прохождения первичных частиц через аппаратуру NUCLEON может сопровождаться большой погрешностью в том случае, если доля провзаимодействовавших в конверторе первичных частиц велика, а сами частицы достаточно массивны – например, для протонов названный эффект не очень существенен (поскольку для первичных протонов разница в характере распределений вторичных каскадов, порожденных ими в мишени и конверторе, не так существенна), но становится весьма заметным уже для ядер гелия. На практике подобная ситуация возникла в связи с обработкой результатов тестового эксперимента прототипа аппаратуры NUCLEON на ускорителе элементарных частиц в CERN (Швейцария): в силу настроек тестовой аппаратуры доля провзаимодействовавших в конверторе, а не в мишени, первичных частиц была очень значительна (около 85%), что явилось одной из возможных причин ухудшения точности энергетических измерений с ростом заряда и массы первичной частицы.

Описание вычислительного эксперимента

Описываемой в данной работе методикой была решена вычислительная задача по разделению первичных частиц, подлежащих регистрации аппаратурой NUCLEON, на два класса:

- класс I – первичные частицы, первый акт взаимодействия которых с веществом произошел в мишени измерительной установки;

- класс II – первичные частицы, не провзаимодействовавшие в мишени и претерпевшие первый акт взаимодействия в конверторе установки.

Для решения поставленной задачи была использована компьютерная модель спектрометра NUCLEON, разработанная на основе моделирующего взаимодействие элементарных частиц с веществом программного комплекса GEANT 3.21 [6]. В ходе описываемого вычислительного эксперимента была разыграна статистическая выборка, имитирующая прохождение первичных частиц – ядер гелия (40 тысяч событий), углерода (30 тысяч событий) и кальция (20 тысяч событий) – через апертуру установки. Около четверти этих событий, согласно принятым условиям триггерного отбора, претерпели упругое взаимодействие с веществом в пределах измерительной установки; остальные прошли всю геометрию прибора без взаимодействия и тем самым интереса не представляли. Доля провзаимодействовавших в конверторе измерительной установки первичных частиц составила около 85% от общего числа первичных частиц, претерпевших неупругое взаимодействие в пределах установки. Соответственно, доля первичных частиц, взаимодействие которых произошло в мишени, составила приблизительно 15%.

Моделировался случай вертикального падения пучка первичных частиц на центр плоскостей стриповых детекторов. С целью наибольшего соответствия параметров модели условиям практической задачи обработки результатов эксперимента с прототипом аппаратуры на ускорителе элементарных частиц, энергия поступающих на вход компьютерной модели спектрометра первичных частиц выбиралась одина-

ковой для всего пучка (так называемый монохроматический пучок по энергиям, или монопучок) и равной 158 ГэВ на нуклон, что совпадало с энергией частиц экспериментального пучка на ускорителе; либо 79 и 315 ГэВ/нуклон (половина экспериментальной энергии и двойное ее значение соответственно). Геометрия моделируемой установки была максимально приближена к геометрии прототипа, и с этой же целью в обработку вычислительной методикой были приняты только данные по сигналам с одной из двух матриц кремниевых стриповых детекторов, поскольку только одна из двух матриц надежно работала в ускорительном эксперименте с прототипом (толщина матрицы 300 мкм, ширина каждого стрипа 50 мкм, общее количество стрипов на одной матрице – около одной тысячи).

Результаты вычислительного эксперимента

Результаты вычислительного эксперимента по разделению первичных частиц (ядер гелия, углерода и кальция) по месту первого взаимодействия – мишень либо конвертор, представлены ниже в таблице 1. Дадим разъяснения использующихся в этой таблице *обозначений переменных*:

- $P_{К\ среди\ М}$ – это вероятность того, что первичная частица, которая на самом деле претерпела взаимодействие в мишени, будет в результате классификации ошибочно принята за провзаимодействовавшую в конверторе. Данная вероятность в вычислительном эксперименте оценивается как доля частиц, подвергнувшихся описанной выше ошибочной классификации, среди всего числа частиц, испытавших взаимодействие в мишени;

- $P_{М\ среди\ К}$ – вероятность того, что первичная частица, которая на самом деле претерпела взаимодействие в конверторе, будет ошибочно классифицирована как провзаимодействовавшая в мишени. Аналогично $P_{К\ среди\ М}$, величина $P_{М\ среди\ К}$ также оценивается как доля соответствующим образом классифицированных частиц среди всего числа частиц, испытавших взаимодействие в конверторе;
- порог $\varepsilon_{отн}$ представляет собой отношение значения ε, подставляемого в формулу для классификации первичных космических частиц (6), к фиксированному значению ε_0 (3) порога для байесовского решающего правила,

$$\varepsilon_{отн} = \varepsilon / \varepsilon_0.$$

Для каждого значения порога имеем свою вычислительную формулу для классификации, сопровождающуюся своим значением погрешности. Из полученной в вычислительном эксперименте таблицы, аналогичной таблице 1, исследователь может определить оптимальное значение ε, исходя из собственных априорных представлений как о величине каждой из погрешностей $P_{М\ среди\ К}$ и $P_{К\ среди\ М}$, так и о соотношении между ними.

При анализе таблицы 1 следует помнить, что наиболее важной для физического приложения рассматриваемого вычислительного примера является задача исключения из банка данных первичных частиц, претерпевших взаимодействие в конверторе. Ошибка, связанная с отбрасыванием при этом из общей статистики тех событий, в которых первичная частица на самом деле провзаимодействовала в мишени, но взаимодействие было отнесено к конвертору, из априорных физических соображений имеет

для нас меньшее значение. Обращаясь к структуре таблицы, сказанное означает, что хорошая методика классификации должна в первую очередь минимизировать (либо не допускать превышения некоторого критического значения) величину ошибки $P_{M\ среди\ K}$ (строки таблицы, выделенные серым цветом). При заданном же ограничении на величину $P_{M\ среди\ K}$ методика должна минимизировать ошибку классификации «в обратном направлении» $P_{K\ среди\ M}$, значения которой приведены в строках, располагающихся ниже выделенных.

Таблица 1. Величина **погрешностей** $P_{M\ среди\ K}$ и $P_{K\ среди\ M}$ **классификации** первичных частиц КЛ по местоположению точки акта первого взаимодействия с веществом: в пределах *мишени* или в пределах *конвертора* установки спектрометра

Тип первичной частицы	Энергия, ГэВ/нуклон	Тип погрешности	Параметр $\varepsilon_{отн}$							
			0.5 ... 0.3	0.25	0.2	0.15	0.10	0.05	-0.05	
Гелий	79	$P_{M\ среди\ K}$	0	0.4	3	7	13	19	27	45
		$P_{K\ среди\ M}$	100	88	69	43	30	22	19	9
	158	$P_{M\ среди\ K}$	4	7	9	10	13	15	19	26
		$P_{K\ среди\ M}$	61	48	42	36	32	26	20	12
	315	$P_{M\ среди\ K}$	4	7	8	9	10	13	15	19
		$P_{K\ среди\ M}$	61	49	46	41	38	36	32	24
Углерод	79	$P_{M\ среди\ K}$	3	9	11	15	19	22	26	32
		$P_{K\ среди\ M}$	46	31	29	24	21	20	19	16
	158	$P_{M\ среди\ K}$	4	6	7	8	9	10	11	15
		$P_{K\ среди\ M}$	46	37	34	33	31	28	27	25
	315	$P_{M\ среди\ K}$	6	9	10	11	13	15	17	20
		$P_{K\ среди\ M}$	57	47	44	42	41	38	36	31

Кальций	79	$P_{M\ среди\ K}$	3	5	6	7	7	8	9	12
		$P_{K\ среди\ M}$	58	50	47	46	45	43	41	38
	158	$P_{M\ среди\ K}$	4	7	8	9	10	11	13	16
		$P_{K\ среди\ M}$	57	49	46	45	45	43	41	37
	315	$P_{M\ среди\ K}$	3	6	6	7	8	9	10	14
		$P_{K\ среди\ M}$	55	47	45	43	41	38	36	32

Из таблицы видно, что при величине погрешности наиболее существенного для нас типа $P_{M\ среди\ K}$ порядка 10-15% погрешность другого типа $P_{K\ среди\ M}$, приводящая только к уменьшению общей статистики, составляет около 30-35%, т.е. при использовании данной методики ограничение на величину $P_{M\ среди\ K}$ в 10-15% оставляет для анализа событий регистрации первичных частиц, претерпевших взаимодействие в мишени, 65-70% всей статистики по таким событиям. Кроме того, видно, что оптимальное для любого заданного уровня $P_{M\ среди\ K}$ значение порога $\varepsilon_{отн}$, как одного из параметров методики, немного варьируется для разных ядер и разных энергий, однако даже при одном и том же фиксированном значении $\varepsilon_{отн}$ и фиксированном уровне $P_{M\ среди\ K}$ погрешность $P_{K\ среди\ M}$ вполне удовлетворительна. Например, если мы хотим, чтобы ни в одном из случаев $P_{M\ среди\ K}$ не превышала 15%, мы можем выставить значение порога $\varepsilon_{отн}$ = 0.2, при этом $P_{K\ среди\ M}$ ни для одного из проанализированных типов ядер и значений первичной энергии не превышает 46%. Вопрос о зависимости точности классификации и оптимальных параметров классифицирующей методики от энергии требует дополнительного исследования на статистических выборках, являющихся уже не монохромными по энергиям, как в нашем случае, а представляющих собой пучки пер-

вичных частиц разных энергий, например, равномерного распределения энергий в логарифмическом диапазоне.

Отметим также, что *полная*, или *суммарная*, ошибка классификации, которую можно определить как вероятность хотя бы какой-нибудь из двух возможных ошибок $P_{К\ среди\ М}$, или $P_{М\ среди\ К}$, не равна обычной алгебраической сумме этих двух величин. Выражение для суммарной ошибки классификации можно вывести из формулы полной вероятности, поскольку два типа «событий-ошибок» – отнесение взаимодействия в мишени к взаимодействию в конверторе и наоборот – составляют полную группу несовместных событий (взаимодействие первичной частицы могло произойти только либо в мишени, либо в конверторе, если мы рассматриваем только провзаимодействовавшие первичные частицы). Поэтому полная ошибка $P_{сумм}$ равна:

$P_{сумм}$ = P(ошибки|конвертор) P(конвертор) + + P(ошибки|мишень) P(мишень),

где:

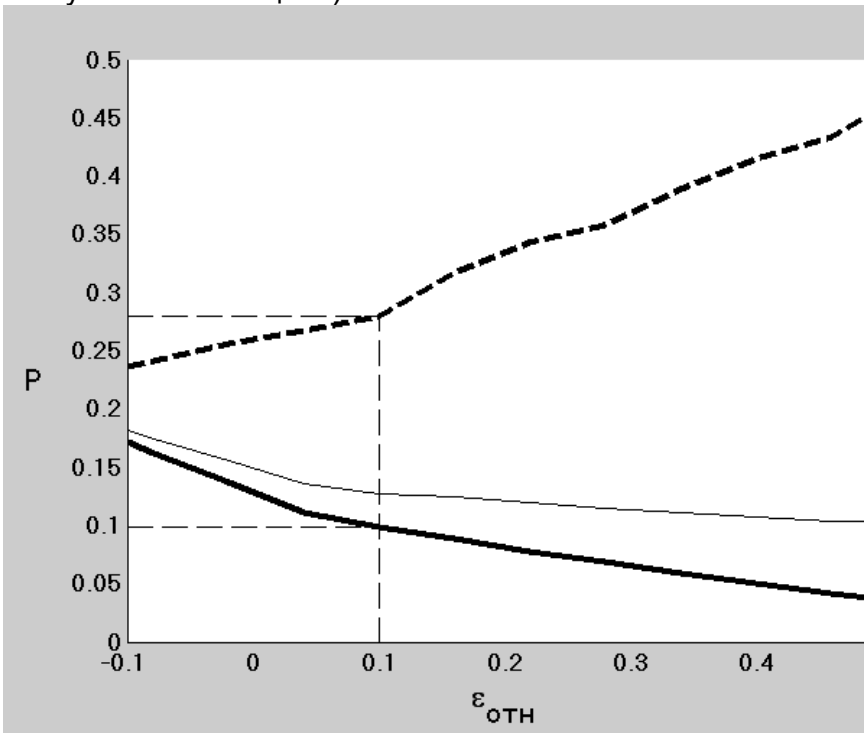
- P(ошибки|конвертор) – вероятность любой ошибки классификации при условии, что на самом деле взаимодействие произошло в конверторе. Очевидно, что P(ошибки|конвертор) = $P_{М\ среди\ К}$;
- P(ошибки|мишень) – вероятность любой ошибки классификации при условии, что взаимодействие произошло в мишени. P(ошибки|мишень) = $P_{К\ среди\ М}$;
- P(конвертор) и P(мишень) – вероятности взаимодействия в конверторе и мишени соответственно. Если первый (из двух возможных) класс первичных частиц – это частицы, провзаимодействовавшие в мишени, а второй – в конверторе, то:
  - P(мишень)=P_1,
  - P(конвертор)=P_2.

Окончательно,

$$P_{сумм} = P_{К\ среди\ М}\ P_1 + P_{М\ среди\ К}\ P_2. \qquad (10)$$

На рисунке 1 приводится пример того, как данные из таблицы, подобной таблице 1, могут быть представлены в виде калибровочной кривой.

Рисунок 1. Зависимость **ошибок классификации** $P_{М\ среди\ К}$ и $P_{К\ среди\ М}$ **от величины порога** $\varepsilon_{отн}$ для ядер углерода при энергии 158 ГэВ на нуклон (построена на основе графы «Углерод, энергия 158 ГэВ на нуклон» таблицы 1)



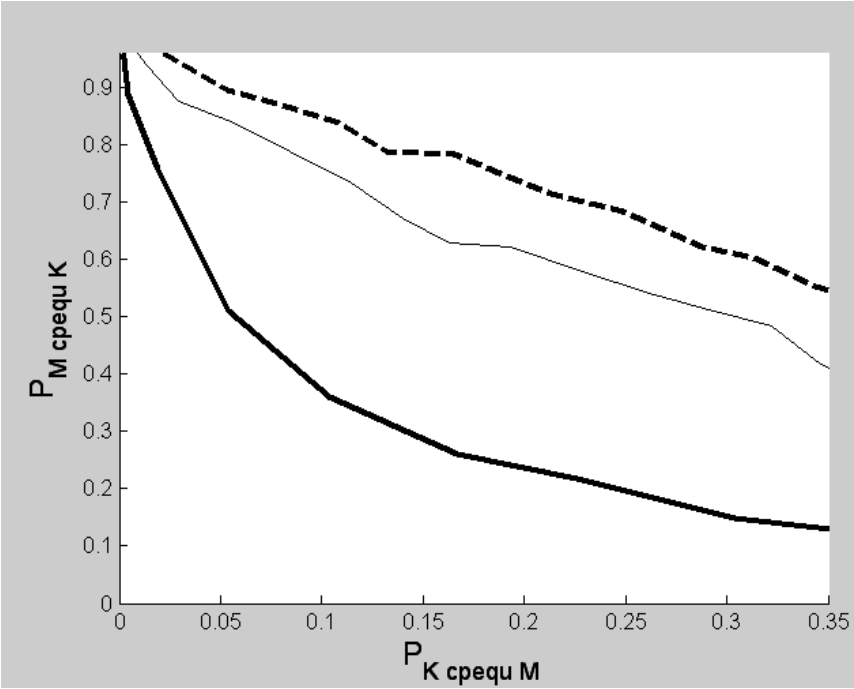
На рисунке:

- толстая пунктирная кривая – ошибка типа $P_{К\ среди\ М}$ (отнесение к взаимодействию в конверторе событий, взаимодействие в которых на самом деле произошло в мишени);
- толстая сплошная кривая – ошибка типа $P_{М\ среди\ К}$ (отнесение к взаимодействию в мишени событий со взаимодействием в конверторе);
- тонкие пунктирные линии демонстрируют графическую процедуру выбора оптимального значения порога $\varepsilon_{отн}$. Нижняя горизонтальная линия изображает величину ошибки $P_{М\ среди\ К}$ в 10%, вертикальная линия определяет соответствующий этой ошибке порог ε, верхняя горизонтальная линия указывает на величину имеющей при этом место ошибки $P_{К\ среди\ М}$ (в данном случае 28%);
- тонкая сплошная кривая изображает величину суммарной ошибки классификации (10). Суммарная ошибка невелика и лишь немного превышает ошибку $P_{М\ среди\ К}$, т.к. значительная часть (P_2 = 85%) общего числа частиц во входном пучке претерпела первое неупругое взаимодействие в конверторе, а не в мишени.

Рисунок 2 представляет зависимость $P_{К\ среди\ М}$ ($P_{М\ среди\ К}$). Для сравнения на этом же рисунке приведены аналогичным образом посчитанные погрешности классификации частиц по значению либо одного параметра $N=\sum I_i$ (суммарный сигнал детектора при регистрации одной первичной частицы, пропорциональный множественности порожденных ею вторичных частиц), либо одного параметра $S = \sum_i c_i I_i$ (где коэффициенты c_i зависят от расстояния до оси пучка

вторичных частиц [5]), использующегося в традиционной одномерной методике восстановления первичной энергии в проекте NUCLEON [5]. Погрешности обоих одномерных алгоритмов гораздо больше погрешности многомерной методики классификации, так как распределения и того, и другого одномерных параметров в силу больших статистических флуктуаций практически не зависят от местоположения точки первого взаимодействия первичной частицы и потому, как видно из рисунка, непригодны для классификации частиц.

Рисунок 2. Соотношение между **ошибками классификации** $P_{К\ среди\ М}$ и $P_{М\ среди\ К}$ для **различных алгоритмов классификации** первичных частиц. Первичные частицы – ядра He энергии 79 ГэВ на нуклон



На рисунке:

- толстая сплошная линия – многомерная методика распознавания;
- тонкая сплошная линия – классификация по значению одного параметра *S*;
- пунктирная линия – классификация по значению множественности вторичных частиц

Наглядная иллюстрация использующегося в методике принципа классификации

Наглядная интерпретация того внутреннего механизма, который используется в предлагаемой методике для классификации первичных частиц, а также демонстрация физической причины высокой точности разделения первичных частиц по области локализации их первого упругого взаимодействия, которая наблюдается в исследованном случае, может быть получена графически следующим образом. Изобразим на рисунке 3 векторы математических ожиданий M_1 и M_2 измеряемых переменных, т.е. средние значения сигналов матрицы кремниевого стрипового детектора, для каждого стрипа. Формулу (2) для вычисления байесовского классификатора представим в таком виде:

$$t(\xi) = \mathbf{b}^T\xi + c, \qquad (11)$$

где:

- $\mathbf{b}^T=(M_2-M_1)^T F^{-1}$ – матрица-строка (**b** – матрица-вектор) постоянных коэффициентов;
- $c=0{,}5(M_1{}^T F^{-1}M_1 - M_2{}^T F^{-1}M_2)$ – скалярная постоянная.

Поскольку сама процедура классификации (6) заключается в сравнении выражения (11) для t(ξ) с постоянной величиной (здесь под «постоянной» понимается «одинаковая для всех стрипов матрицы детектора») – порогом ε, то, очевидно, предлагаемая методика классификации будет тем чувствительнее, чем точнее значения коэффициентов $\mathbf{b}^T$ будут отражать различие между векторами M_1 и M_2 (напомним, тестируемая в данном вычислительном эксперименте методика не учитывает различие дисперсий и корреляций двух распределений, ограничиваясь только раз-

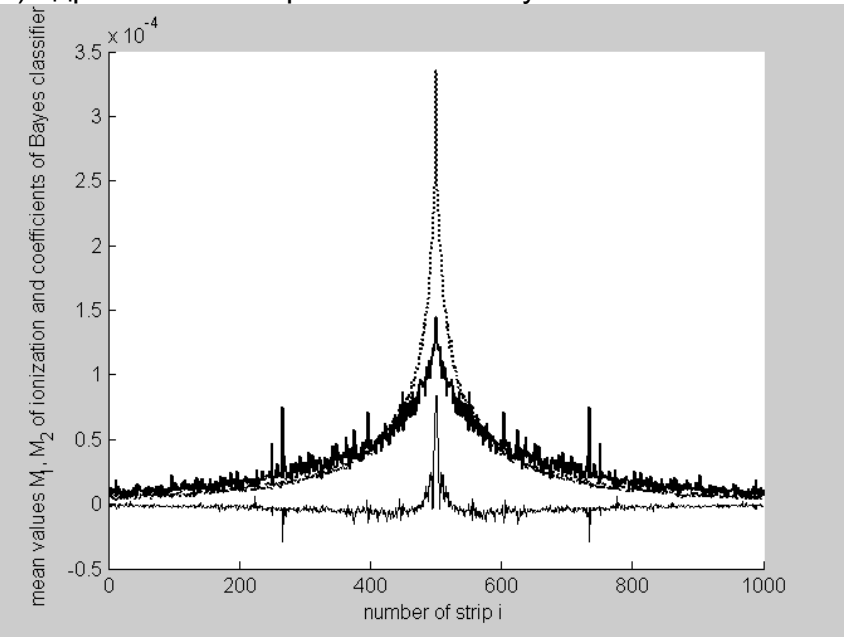
ницей их средних значений). Следовательно, весьма полезным представляется изобразить на том же графике (рисунок 3) значения коэффициентов $\mathbf{b}^T$ (т.е. тех коэффициентов, на которые в силу (11) в ходе классификации будет домножаться сигнал в соответствующем стрипе) в приведенном масштабе, для того чтобы ясно представлять себе не только визуальный характер пространственных распределений сигнала в стриповом детекторе от первичных частиц первого и второго классов, но и те особенности этих распределений, которые позволяют относить событие к первому или второму классу согласно предлагаемой методике.

На рисунке 3 хорошо видно, что пространственное распределение сигнала от вторичных каскадов, порожденных взаимодействем первичных частиц в мишени, имеет более широкую форму и характеризуется меньшей величиной сигнала, чем распределение, вызванное взаимодействием в конверторе. Это объясняется тем, что, во-первых, конвертор располагается ближе к плоскости детектора, чем мишень; а во-вторых, если взаимодействие произошло в мишени, то вторичный каскад гамма-квантов, который дополнительно размножается в веществе конвертора путем распада гамма-квантов на электронно-позитронные пары, образуется еще в веществе мишени и успевает до размножения в конверторе разойтись на большее расстояние. Достигающий в итоге плоскости детектора электронно-позитронный каскад в случае, если взаимодействие первичной частицы с веществом произошло в конверторе, а не в мишени, будет у́же каскада, образовавшегося от взаимодействия в мишени, но выше его в максимуме. Именно эти две

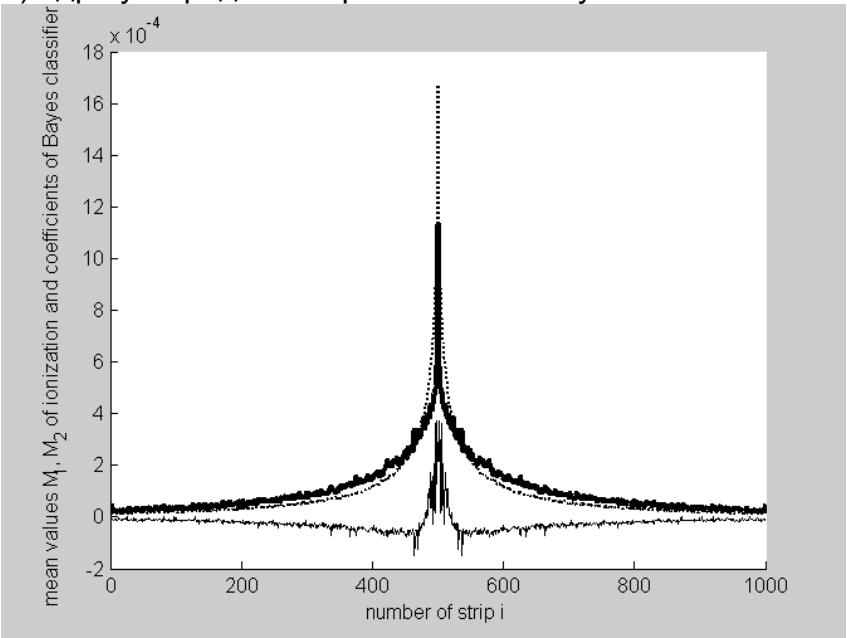
особенности различия двух классов распределений эффективно учитываются в формуле (11) для байесовского классификатора, позволяя в результате провести сепарацию первичных частиц по местоположению точки первого взаимодействия.

Рисунок 3. **Средние величины сигналов** матрицы стрипового детектора от каскадов, порожденных взаимодействием первичных частиц с веществом **конвертора** (координаты вектора M_1) и с веществом **мишени** (координаты вектора M_2). На этом же рисунке в приведенном таким образом, чтобы кривые не перекрывали друг друга, масштабе отложены величины **коэффициентов** $\mathbf{b}^T$ методики классификации из формулы (11)

а) ядра гелия с энергией 158 ГэВ/нуклон



б) ядра углерода с энергией 315 ГэВ/нуклон



На рисунках:

- толстая сплошная линия – M_1 (средние значения сигнала от взаимодействия в мишени);
- точечная линия – M_2 (средние значения сигнала от взаимодействия в конверторе);
- тонкая линия – коэффициенты $\mathbf{b}^T$ байесовского классификатора (формула (11));

Эффективность учета статистических различий пространственных распределений обоих классов событий хорошо демонстрируется на рисунке: мы видим, что коэффициенты байесовского классификатора имеют локальный максимум в области центрального стрипа, в направлении которого был ориен-

тирован пучок первичных частиц. При удалении от центрального стрипа коэффициенты меняют знак, поскольку уже не распределение каскада, относящегося к взаимодействию в конверторе, а распределение каскада от взаимодействия в мишени начинает в среднем преобладать. Также хорошо отражается с помощью графика тот факт, что абсолютные значения коэффициентов приближаются к нулю, т.е. уменьшаются веса сигналов в соответствующих стрипах, с дальнейшим увеличением расстояния от центра пучка, что связано со всё большим сближением формы двух распределений.

Заключение

Разработанная методика сепарации первичных частиц на два класса, как было продемонстрировано, отличается не только гибкостью и эффективностью, но и допускает весьма наглядную интерпретацию. Она проста в реализации (например, в компьютерной среде MATLAB) и, в силу возможности варьирования значения порога в вычислительной формуле, позволяет исследователю самому выбрать вариант методики, устраивающий его по таким параметрам, как величины ошибок отнесения первичных частиц к каждому из двух классов и соотношение между этими ошибками.

Серия вычислительных экспериментов с модельными данными, проведенная для решения конкретной использованной в качестве иллюстративного материала задачи – сепарации первичных частиц по местоположению точки первого неупругого взаимодействия – показала, что достаточно хорошее качество классификации (суммарная ошибка классифика-

ции (10) не более 10%) в исследованном диапазоне энергий наблюдается для ядер как легких (гелий), так и тяжелых (кальций) химических элементов. Заметим, что, в отличие от методики определения первичной энергии в том же проекте научной аппаратуры NUCLEON, никакого «альтернативного» многомерной статистической методике алгоритма сепарации первичных частиц по области первого взаимодействия на сегодняшний день не существует.

Все эти полученные в работе практические выводы позволяют надеяться, что данная методика окажется действенной и полезной при решении многих прикладных задач космофизики, где необходимо разделение частиц первичного космического излучения на два класса на основе измерений большого числа физических переменных, в каждой из которых присутствует необходимая для формирования алгоритма классификации статистическая информация.

Литература

1. Фукунага К., Введение в статистическую теорию распознавания образов. Москва: Наука, 1979.
2. Подорожный Д.М., Постников Е.Б., Свешникова Л.Г., Турундаевский А.Н. Препринт НИИЯФ МГУ. 2003-12/725. Москва. 2003.
3. Е.Б. Постников, Г.Л. Башинджагян, Н.А. Короткова и др., Изв. РАН. Сер. физ. 66, 1634 (2002).
4. Д.М.Подорожный, Е.Б.Постников, Л.Г.Свешникова. ЯФ 68, 51 (2005).
5. Н.А.Короткова, Д.М.Подорожный, Е.Б.Постников и др., ЯФ 65, 884 (2002).
6. GEANT User's Guide, CERN DD/EE/83/1 (Geneva, 1983).
7. A.Turundaevsky, D.Podorozhnyi, E.Postnikov et all, The KLEM-NUCLEON Instrument Detailed Simulation. Proc. 28 ICRC. Tsukuba. Japan. 2003. V.OG.1.5. P.2213-16

Содержание

Евгений Борисович Постников

Применение многомерных методов теории распознавания образов к решению задач классификации частиц первичного космического излучения

Препринт НИИЯФ МГУ - 2004-23/762

Работа поступила в ОНТИ 06. 12.2004 г.

Издательство УНЦ ДО

ИД № 00545 от 06.12.1999

117246, Москва, ул. Обручева, 55А, УНЦ ДО
т\ф (095) 718-6966, -7767, -7785 (комм.)
e-mail: ***izdat@abiturcenter.ru***
http://www.abiturbook.ru/izdat

Заказное. Подписано в печать 06.12.2004 г. Формат 60x90/16
Бумага офсетная №2. Усл.п.л. 2,31
Тираж 25 экз. Заказ № 722

Отпечатано в Мини-типографии УНЦ ДО
http://abiturcenter.ru/print/
в полном соответствии с качеством
предоставленного оригинал-макета