УДК 519.237, 524.1

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

Е. Б. Постников

(*НИИЯФ*)

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

Работа посвящена применению многомерной статистической теории распознавания образов к решению задач классификации первичных частиц космического излучения. В качестве иллюстрации на основе модельных данных решена практическая задача, не имеющая даже приближенного решения в рамках детерминистического подхода. Описан алгоритм разделения первичных частиц на два класса, использующий понятие байесовского линейного классификатора, а также очерчен широкий круг задач экспериментальной космофизики, для решения которых может быть применена предлагаемая методика.

Введение

При анализе и обработке как экспериментальных, так и модельных данных в физике космических лучей (КЛ) часто возникают однотипные задачи классификации. Например, на основе статистического анализа данных, несущих только косвенную информацию о первичной частице, таких, как ионизация в слоях толчковой установки или сигналы с матрицы стрипового детектора, часто требуется определить значение какой-либо дискретной характеристики первичной частицы, например заряда, массового числа, номера пада регистрирующей матрицы, через который прошла траектория частицы (эта величина отвечает за направление прихода частицы), 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 величина может принимать только два значения. Одним из наиболее простых и в то же время надежных методов классификации в случае двух классов является байесовское решающее правило, о применении которого к решению данного круга задач и будет рассказано в настоящей работе.

Преимущества многомерного подхода (в том числе и в рамках байесовской стратегии классификации) к решению космофизических задач были продемонстрированы еще в работе [2] применительно к компьютерной модели черенковского атмосферного телескопа. Впоследствии в ряде публикаций (напр., [3]) теория распознавания образов использовалась при обработке данных наблюдений широких атмосферных ливней (ШАЛ) для сепарации первичных частиц КЛ по массовому составу. При этом число переменных, на основе значений которых проводилась классификация, равнялось двум (это были полученные суммированием показаний ряда детекторов значения количества электронов и мюонов в каждом индивидуальном ШАЛ). Число же классов — количество наиболее распространенных в КЛ химических элементов — превышало два, в силу чего, а также по причине недостаточности статистики по модельным данным и большого значения энергетического порога исследуемых данных ($> 10^6$ ГэВ), была обнаружена сильная зависимость результатов от модели взаимодействия элементарных частиц с веществом, используемой для обучения распознающей методики.

В нашем случае речь идет об условиях моделирования ускорительного эксперимента, т. е. о более низком энергетическом интервале (сотни ГэВ), на котором не наблюдается расхождения между ядернофизическими моделями, и в этом случае статистика с легкостью может быть набрана в достаточном количестве. Кроме того, в предлагаемом алгоритме проблема некорректности вычислений и необходимости отбора из большого числа переменных нескольких самых информативных решается путем анализа спектра ковариационной матрицы данных [4], поэтому для построения методики классификации в настоящей работе были использованы все непосредственно измеряемые переменные. Наконец, простой алгоритм и положительные результаты были получены, в отличие от [3], без привлечения непараметрического подхода, что позволило избежать громоздких процедур оценивания функций плотности вероятности. Разработка и успешное применение представленного многомерного статистического алгоритма представляет собой продолжение внедрения современной многомерной статистической техники в обработку результатов космофизических экспериментов (работы [4–6], связь с которыми хорошо прослеживается в силу сходства общей идеологии подхода к постановке задач анализа и интерпретации данных).

Описание методики

Пусть наша регистрирующая аппаратура позволяет получать 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. Для применения статистических многомерных методик работы с данными следует объединить все эти измеряемые переменные в случайный вектор, который мы обозначим через $\boldsymbol{\xi}$. Каждый акт регистрации первичной частицы предоставляет в распоряжение исследователя очередную реализацию $\boldsymbol{\xi}$. Когда алгоритм решения задачи классификации первичных частиц будет построен, мы сможем в зависимости от значения конкретной реализации $\boldsymbol{\xi}$ относить первичную частицу к одному из двух заранее известных классов — классу I или классу II.

Согласно байесовскому решающему правилу, классификация производится на основе скалярной функции $t(\boldsymbol{\xi})$ — байесовского классификатора. Вид этой функции определяется 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 относится зафиксированная в этом событии первичная частица. Названный банк является обучающей выборкой. Он может быть получен, например, при помощи компьютерного моделирования с использованием хорошо известного в ядерной физике программного комплекса GEANT [7] симуляции по методу Монте-Карло процесса взаимодействия элементарных частиц с веществом.

После того как вид байесовского классификатора $t(\boldsymbol{\xi})$ определен, его значение вычисляется уже для любого события регистрации первичной частицы и сравнивается со значением порога ε, постоянной (неслучайной) величины, о вычислении которой будет рассказано ниже. Классификация происходит следующим образом: первичная частица относится к классу I, если $t(\boldsymbol{\xi}) < \varepsilon$, и к классу II, если $t(\boldsymbol{\xi}) > \varepsilon$.

Вычисление классификатора на контрольной выборке — банке данных, отличном от обучающего, но таком, для которого также точно известно, к какому из двух классов относится каждое содержащееся в банке событие, — позволяет оценить погрешность сформированной методики классификации. По этому банку определяется количество ошибочно классифицированных частиц, на самом деле априорно принадлежащих не тому классу, к которому они были отнесены методикой распознавания.

Расчетные формулы

Приведем формулы для вычисления байесовского классификатора уже с учетом конкретной реализации предлагаемого алгоритма [1, 8], в котором учтено то обстоятельство, что из физических соображений рассматриваемые нами классы всегда «неравноценны» (например, при анализе экспериментальных данных важнее как можно в более полном объеме избавиться от «фоновых» частиц, а вероятность отбросить при этом также и полезные события имеет несколько меньшее значение). Соответственно неравноценны для нас и погрешности ошибочного отнесения первичных частиц к первому и ко второму классу. На этом основании значение порога ε вычисляется не по жесткой формуле, диктуемой байесовским решающим правилом, а определяется экспериментально как оптимальное значение, при котором соотношение между ошибками классификации на классах I и II будет максимально соответствовать априорным представлениям исследователя.

В этом случае байесовский классификатор (в линейном приближении) будет определяться по следующей формуле:

$$t(\boldsymbol{\xi}) = (\mathbf{M}_2 - \mathbf{M}_1)^{\top} F^{-1} \boldsymbol{\xi}, \qquad (1)$$

где $\mathbf{M}_1$ и $\mathbf{M}_2$ — векторы математических ожиданий случайного вектора $\boldsymbol{\xi}$ по распределению первичных частиц только первого и только второго классов соответственно, F — ковариационная матрица случайного вектора $\boldsymbol{\xi}$ по распределению всех первичных частиц (смеси первого и второго классов), значок «$\top$» обозначает транспонированную матрицу (в данном случае — транспонированную матрицу-столбец, т. е. матрицу-строку), F^{-1} — матрица, обратная к матрице F.

На практике для вычисления классификатора необходимо оценить по обучающему банку данных все входящие в формулу (1) неизвестные величины — координаты векторов $\mathbf{M}_1$ и $\mathbf{M}_2$ и элементы матрицы F. Оценивание происходит по стандартной статистической схеме, позволяющей получить несмещенные оценки математических ожиданий, дисперсий и ковариаций [8].

Значение порога ε, согласно предлагаемому алгоритму, пробегает последовательно весь интервал значений $t(\boldsymbol{\xi})$ от минимального до максимального, что позволяет для каждого ε определить ошибку классификации на классах I и II и построить зависимость одной ошибки от другой.

Описание вычислительного эксперимента

Модельный вычислительный эксперимент проводился в рамках общей практической задачи оптимизации измерительной аппаратуры проекта NUCLEON [9]. Цель проекта состоит в создании компактной аппаратуры для регистрации КЛ (протонов и ядер) в широком энергетическом диапазоне. Энергию первичных частиц планируется определять из пространственной плотности $\rho(x, y)$ распределения потока вторичных частиц, рожденных в тонкой мишени (первый акт неупругого взаимодействия)

и размноженных в сверхтонкой толчковой установке — конвертере. Для измерения $\rho(x, y)$ в состав аппаратуры войдет кремниевый стриповый детектор. Величина I_i сигнала в каждом стрипе (с номером i) детектора пропорциональна ионизации в этом стрипе. Именно совокупность сигналов I_i со всех стрипов 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 $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} = \{I_1\ I_2\ \ldots\ I_m\}^\top.$$

Размерность вектора $\boldsymbol{\xi}$ имеет порядок 10^3, поэтому задача интерпретации результатов измерения аппаратуры NUCLEON является, как это отмечалось в [4–6], существенно многомерной.

Проблема классификации первичных частиц в рамках проекта возникла в связи с повышением точности энергетического разрешения аппаратуры. Дело в том, что некоторая доля из числа всех регистрируемых прибором первичных частиц может испытать взаимодействие не в мишени, а в конвертере. Пространственное распределение вторичных частиц в первом и во втором случае имеет разный характер. Поэтому точность обработки всего банка событий одинаковым алгоритмом является низкой в том случае, если доля фоновых «конвертерных» первичных частиц велика, а сами частицы достаточно массивны. Например, для протонов названный эффект не очень существенен, но становится весьма заметным уже для ядер He.

Таким образом, в рамках нашей задачи все события регистрации первичных частиц КЛ можно разделить на два класса в зависимости от места первого неупругого взаимодействия частицы с веществом прибора: к классу I относятся события со взаимодействием в мишени; к классу II — со взаимодействием в конвертере.

Для решения поставленной задачи мы использовали компьютерную модель аппаратуры, разработанную на основе программного комплекса GEANT 3.21 [7]. В ходе вычислительного эксперимента была разыграна статистическая выборка, имитирующая прохождение первичных частиц — ядер He, C и Ca — через апертуру установки.

Результаты вычислительного эксперимента

Результаты вычислительного эксперимента представлены в таблице. Разъясним обозначения переменных:

$P_{\mathrm{I}\to\mathrm{II}}$ — вероятность того, что первичная частица, которая на самом деле испытала взаимодействие в мишени (класс I), будет в результате классификации ошибочно принята за провзаимодействовавшую в конвертере (отнесена к классу II);

$P_{\mathrm{II}\to\mathrm{I}}$ — вероятность ошибочно классифицировать первичную частицу, испытавшую взаимодействие в конвертере, как провзаимодействовавшую в мишени.

Величина (в %) погрешностей $P_{\mathrm{II}\to\mathrm{I}}$ и $P_{\mathrm{I}\to\mathrm{II}}$ классификации первичных частиц КЛ по месту первого взаимодействия с веществом

Тип первичной частицы		Гелий			Углерод			Кальций		
Энергия, ГэВ		79	158	315	79	158	315	79	158	315
		$P_{\mathrm{I}\to\mathrm{II}}$								
$P_{\mathrm{II}\to\mathrm{I}}$	4	57	58	61	41	43	63	50	56	52
	8	41	44	46	32	32	49	42	45	40
	12	34	33	36	28	26	42	37	41	33
	16	27	23	31	24	24	37	32	37	30
	20	22	18	23	22	22	31	28	32	26

При анализе таблицы следует помнить, что наиболее важной для физического приложения рассматриваемого вычислительного примера является задача исключения из банка данных первичных частиц, испытавших взаимодействие в конвертере. 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 конвертеру, имеет для нас меньшее значение. Поэтому методика классификации должна в первую очередь минимизировать (либо не допускать превышения некоторого критического значения) ошибку $P_{\mathrm{II}\to\mathrm{I}}$. При заданном же ограничении на величину $P_{\mathrm{II}\to\mathrm{I}}$ методика должна минимизировать ошибку классификации «в обратном направлении» $P_{\mathrm{I}\to\mathrm{II}}$.

Из таблицы видно, что при незначительной величине погрешности наиболее существенного для нас типа $P_{\mathrm{II}\to\mathrm{I}}$ (порядка 10%) погрешность другого типа, $P_{\mathrm{I}\to\mathrm{II}}$, приводящая только к уменьшению общей статистики, составляет 30–40%, т. е. при использовании данной методики ограничение на величину $P_{\mathrm{II}\to\mathrm{I}}$, практически полностью отсеивающее «плохие» события (взаимодействие в конвертере), оставляет для анализа две трети всей статистики по «хорошим» событиям (мишень). Априорно задаваемому уровню $P_{\mathrm{II}\to\mathrm{I}}$ в 20% соответствует погрешность $P_{\mathrm{I}\to\mathrm{II}}$ также в районе 20%, что гарантирует нам 80% статистики по мишени.

На рис. 1 представлена зависимость $P_{\mathrm{I}\to\mathrm{II}}(P_{\mathrm{II}\to\mathrm{I}})$. 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 $S = \sum\limits_i c_i I_i$ (где коэффициенты c_i зависят от расстояния до оси пучка вторичных частиц [9]), использующегося в традиционной одномерной методике восстановления первичной энергии в проекте NUCLEON [9].

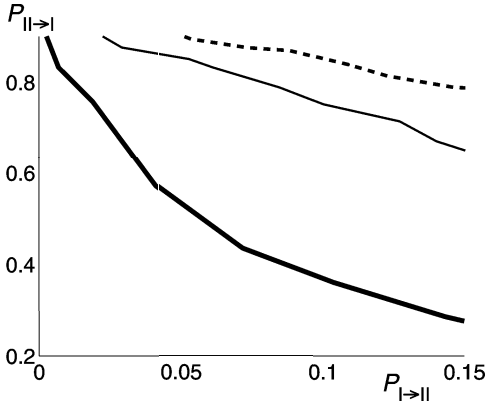

Рис. 1. Зависимость $P_{\mathrm{I}\to\mathrm{II}}(P_{\mathrm{II}\to\mathrm{I}})$ для различных алгоритмов классификации первичных частиц. Первичные частицы — ядра Не энергии 79 ГэВ на нуклон. Жирная сплошная кривая — многомерная методика распознавания; тонкая сплошная — классификация по значению одного параметра S [9]; пунктирная — по значению одного параметра N (множественность вторичных частиц)

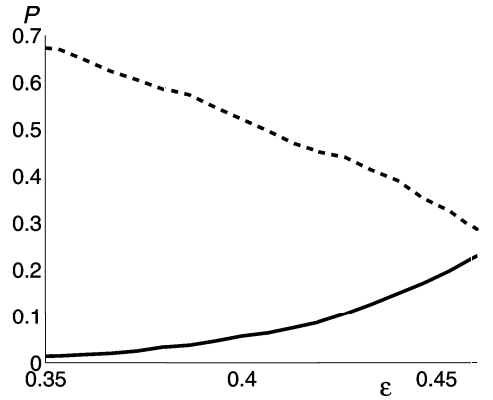

Рис. 2. Зависимость ошибок классификации от величины порога ε для ядер Ca при энергии 158 ГэВ на нуклон. Сплошная кривая — ошибка типа $P_{\mathrm{I}\to\mathrm{II}}$, пунктирная — ошибка типа $P_{\mathrm{I}\to\mathrm{II}}$. Значение порога ε приводится в долях интервала возможных значений байесовского классификатора $t(\boldsymbol{\xi})$ (1)

На рис. 2 приведены калибровочные кривые методики классификации в зависимости от значения порога ε.

На рис. 3 представлена наглядная интерпретация внутреннего механизма, который используется в предлагаемой методике классификации первичных частиц. Здесь изображены, во-первых, векторы математических ожиданий $\mathbf{M}_1$ и $\mathbf{M}_2$ измеряемых переменных, т. е. средние значения сигналов матрицы кремниевого стрипового детектора, для каждого стрипа; а во-вторых, в приведенном масштабе значения тех коэффициентов, на которые в процессе классификации будет домножаться сигнал в соответствующем стрипе [8]. Хорошо видно, что пространственное распределение сигнала от вторичных каскадов, порожденных взаимодействием первичных частиц в мишени, шире и имеет менее острый максимум, чем распределение, вызванное взаимодействием в конвертере [8].

Именно эти две особенности различия двух классов распределений эффективно учитываются в формуле (1) для байесовского классификатора, позволяя в результате провести сепарацию первичных частиц по местоположению точки первого взаимодействия: коэффициент байесовского классификатора как функция номера стрипа имеет локальный максимум в области центрального стрипа, в направлении которого был ориентирован пучок первичных частиц. При удалении от центрального стрипа коэффициент меняет знак, поскольку в среднем начинает преобладать уже не распределение каскада, относящегося к взаимодействию в конвертере, а распределение каскада от взаимодействия в мишени. С дальнейшим увеличением расстояния от центра пучка абсолютное значение коэффициента приближается к нулю, т. е. уменьшаются веса сигналов в соответствующих стрипах, что связано со все большим сближением формы двух распределений.

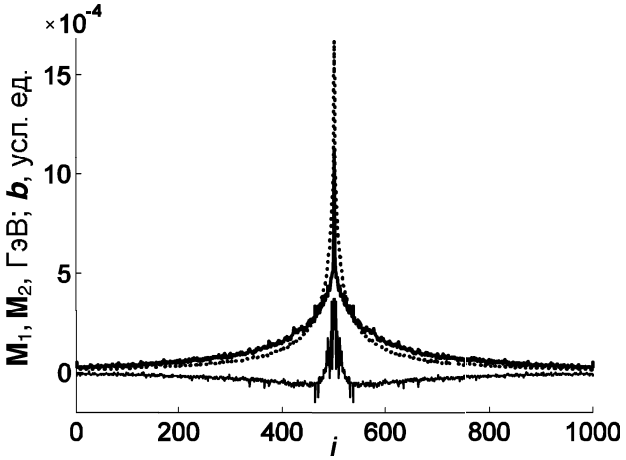

Рис. 3. Средние величины сигналов матрицы стрипового детектора от каскадов, порожденных взаимодействием первичных частиц (ядра C, 315 ГэВ/нуклон) с веществом мишени (координаты вектора $\mathbf{M}_1$, жирная сплошная кривая) и с веществом конвертера (координаты вектора $\mathbf{M}_2$, точечная кривая). i — номер стрипа; $i = 500$ соответствует центру каскада вторичных частиц. Тонкая кривая — величины коэффициентов $\mathbf{b}$ методики классификации [8] в приведенном масштабе

Заключение

Разработанная методика сепарации первичных частиц на два класса не только отличается 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 иллюстративной задачи — сепарации первичных частиц по местоположению точки первого неупругого взаимодействия, показала, что достаточно хо-

рошее качество классификации (суммарная взвешенная ошибка классификации не более 10% [8]) в исследованном диапазоне энергий наблюдается для ядер как легких (He), так и тяжелых (Ca) элементов. В отличие от методики определения первичной энергии в том же проекте научной аппаратуры NUCLEON одномерные алгоритмы обработки данных для решения задачи сепарации оказались несостоятельными.

Полученные в работе выводы позволяют надеяться, что методика окажется полезной при решении многих прикладных задач космофизики, где необходимо разделение частиц первичного космического излучения на два класса на основе измерений большого числа физических переменных.

Литература

1. *Фукунага К.* Введение в статистическую теорию распознавания образов. М., 1979.
2. *Aharonian F.A., Chilingaryan A.A., Plyasheshnikov A.K., Konopelko A.K.* // Preprint YerPhI. 1171(48). Yerevan, 1989.
3. *Antoni T., Apel W.D., Badea F.* et al. // Astroparticle Physics. 2002. **16**, N 3. P. 245.
4. *Подорожный Д.М., Постников Е.Б., Свешникова Л.Г.* // Ядерная физика. 2005. **68**, № 1. С. 51.
5. *Подорожный Д.М., Постников Е.Б., Свешникова Л.Г., Турундаевский А.Н.* // Препринт НИИЯФ МГУ. 2003-12/725. М., 2003.
6. *Постников Е.Б., Башинджагян Г.Л., Короткова Н.А.* и др. // Изв. РАН. Сер. физ. 2002. **66**, № 11. С. 1634.
7. GEANT User's Guide, CERN DD/EE/83/1. Geneva, 1983.
8. *Постников Е.Б.* // Препринт НИИЯФ МГУ. 2004-23/762. М., 2004.
9. *Короткова Н.А., Подорожный Д.М., Постников Е.Б.* и др. // Ядерная физика. 2002. **65**, № 5. С. 884.

Поступила в редакцию
30.11.04