Moscow University Physics Bulletin
Vol. 60, No. 6, pp. 37–42, 2005

Vestnik Moskovskogo Universiteta. Fizika
UDC 519.237, 524.1

USING MULTIDIMENSIONAL PATTERN RECOGNITION TECHNIQUES IN CLASSIFYING PRIMARY COSMIC-RAY PARTICLES

E. B. Postnikov

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

The application of statistical multidimensional pattern recognition theory to solving problems associated with the classification of primary cosmic-ray particles is studied. As an illustration, a practical problem that has not even an approximate solution within the deterministic approach is solved on the basis of model data. An algorithm that is used to divide the primary particles into two classes and employs the concept of a Bayes linear classifier is described, and a broad spectrum of problems of experimental space physics is specified whose solution can be achieved by using the proposed method.

INTRODUCTION

Very similar classification problems often arise in the analysis and processing of experimental and model data in cosmic-ray physics. For instance, when statistically analyzing data that carry only indirect information about a primary particle, such as ionization in the layers of a shock device or signals from the matrix of a stripped detector, there is often the need to determine the value of a discrete characteristic of a primary particle, e. g., the charge and the mass number of the incoming particle, and the number of the pad in the recording matrix through which the particle has passed (this number characterizes the direction of incidence of the particle) or of the measuring unit in which the particle interacted, etc.

As is known [1], classical pattern recognition theory used to solve classification problems guarantees the highest accuracy when the investigated objects have to be divided only into two classes, i. e., when a discrete quantity can take on only two values. One of the simplest and, at the same time, reliable classification methods in the case of two classes uses the Bayes decision rule. The present paper is a study of how this rule can be applied to problems associated with the classification of primary cosmic-ray particles.

The advantages of a multidimensional approach (including those within the Bayes classification strategy) to the solution of space-physics problems were demonstrated in 1989 by Aharonian et al. [2] as applied to a computer model of a Cherenkov atmospheric telescope. Later, researchers (see, e. g., [3]) used pattern recognition theory to process data from observations of extensive air showers (EAS) for separating primary cosmic-ray particles according to their mass composition. Here the number of variables used in the classification was two (these were the numbers of electrons and muons in each individual EAS obtained by summing the readings from a number of detectors). The number of classes, or the number of chemical elements that were most abundant in the cosmic rays, was greater than two, in view of which, and also because of insufficient statistics on the model data and the high energy threshold of the investigated data ($> 10^6$ GeV), a strong dependence of the results of the model of the interaction between elementary particles and matter was revealed and used in training the recognition method.

In our case we are concerned with simulating an accelerator experiment, i.e., a lower energy range (hundreds of gigaelectronvolts) within which no discrepancy between nuclear-physics models is observed, and in this case sufficient statistics can easily be gathered. More than that, in the proposed algorithm the problem of meaningfulness of calculations and of the need to select from a large number of variables only the most informative ones is solved by analyzing the covariance data matrix [4], with the result that we used all the directly measured variables to build our classification method. Finally, a simple algorithm and positive results were obtained, in contrast to those of [3], without adopting a nonparametric approach, which allowed avoiding cumbersome procedures of estimating probability density functions. The development and successful application of the proposed multidimensional statistical algorithm continue the implementation of modern multidimensional statistical techniques in the processing of the results of space-physics experiments (see [4–6], the relation with which is clearly traced due to the similarity of the general ideology of the approach and the statement of problems of analyzing and interpreting data).

THE METHOD

Let us assume that our recording instruments allow extracting multidimensional information about each primary particle, which means that the values of several physical parameters (instead of one) are obtained. Following the terminology adopted in [4], we call these quantities measured variables. To apply data-processing multidimensional statistical techniques, we must combine all measured variables in a random vector, which we denote by $\boldsymbol{\xi}$. Each act of recording a primary particle supplies the researcher with the next implementation of $\boldsymbol{\xi}$. When the algorithm of solving the primary-particle classification is completed, we will be able, depending on the value of the specific implementation of $\boldsymbol{\xi}$, to refer the primary particle to one of the two preset classes, class I or class II.

According to the Bayes decision rule, classification is done according to the scalar function $t(\boldsymbol{\xi})$, known as the Bayes classifier. The form of this function is determined by statistically analyzing the data on the measured variables, either experimental or model, but necessarily classified, i.e., from a data bank in which for each event we know exactly to which of the two classes the primary particle fixed by this event belongs. This data bank is known as the training sample and can be obtained, for instance, by using the GEANT software package (well-known in nuclear physics) that employs the Monte Carlo method to simulate the interaction between elementary particles and matter [7].

After the form of the Bayes classifier $t(\boldsymbol{\xi})$ has been found, its value is calculated for any event of recording a primary particle and the result is compared to the threshold ε, a constant (nonrandom) quantity, whose calculation will be described later in the paper. Classification is done in the following way: a primary particle belongs to class I if $t(\boldsymbol{\xi}) < \varepsilon$ and to class II if $t(\boldsymbol{\xi}) > \varepsilon$.

Calculation of the classifier on a control sample, which is a data bank that differs from the training one but is such that it is exactly known to which of the two classes an event in the bank belongs, makes it possible to estimate the error of the classification method. The bank is used to determine the number of erroneously classified particles that a priori are known to belong not to the class to which they were referred by the recognition method.

THE CALCULATION FORMULAS

Below we give the formulas for calculating the Bayes classifier that allow for a specific implementation of the proposed algorithm [1, 8], which takes into account the fact that the classes we examine are always, on physical grounds, of unequal status (for instance, in analyzing experimental data it is most important to get rid as fully as possible of "background" particles, while the probability of getting rid of useful events in the process is somewhat lower). Accordingly, the errors in referring the primary particles to the wrong class are also of unequal status. Because of all this, the value of the threshold ε is calculated not by the rigid formula prescribed by the Bayes decision rule but is determined through experiments as the optimal value at which the relation between the classification errors for classes I and II agrees as best as possible with the a priori ideas of the researcher.

In this case the Bayes classifier (in the linear approximation) is given by the following formula:

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

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are the expectation vectors of the random vector ξ over the distribution of primary particles of only the first class and only the second class, respectively; F is the covariance matrix of the random vector ξ over the distribution of all primary particles (a mixture of the first and second classes), the superscript $\top$ denotes a transposed matrix (in our case a transposed column matrix, or a row matrix); and F^{-1} is the inverse of F.

Actually, to calculate the classifier one must use the training data bank to estimate all the unknown quantities in (1), i. e., the coordinates of the vectors $\mathbf{M}_1$ and $\mathbf{M}_2$ and the elements of matrix F. Estimates are made according to the standard statistical scheme, which makes it possible to obtain unbiased estimates of the dispersions, variances, and covariances [8].

According to the proposed algorithm, the value of the threshold ε consecutively runs through the entire interval of values of $t(\xi)$ from the minimal value to the maximal one, which makes it possible to determine for each ε the classification error on classes I and II and to build the dependence of one estimate on the other.

DESCRIPTION OF THE COMPUTATIONAL EXPERIMENT

A model computational experiment was carried out within the general practical problem of optimizing the measuring devices of the NUCLEON project [9]. The goal of this project is to build compact devices for recording cosmic rays (protons and nuclei) within a broad energy range. The idea is to determine the energy of the primary particles by measuring the spatial density $\rho(x, y)$ of the distribution of the flux of secondary particles that are produced inside a thin target (the first act of inelastic interaction) and then multiply in an ultrathin shock device, the converter. A silicon stripped detector is included in the measuring unit so that $\rho(x, y)$ can be measured. The magnitude I_i of the signal in each strip (with number i) of the detector is proportional to the degree of ionization in this strip. It is the sum of the signals I_i from all strips of the detector measured simultaneously in the act of recording a single event of the passage of a cosmic-ray particle through the measuring unit that amounts, in our statistical interpretation of the measuring scheme, the random vector of the measured variables

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

The number of dimensions of vector ξ is of order 10^3; therefore, the problem of interpreting the results of measurement done by the NUCLEON complex is, as noted in [4–6], essentially multidimensional.

The problem of classifying primary particles within the project emerged in connection with the increase in the accuracy of the energy resolution of the equipment. The thing is that a fraction of all primary particles recorded by the device may interact not in the target but in the converter. The spatial distributions of secondary particles in the first and in the second case are different. Hence the accuracy of processing the entire data bank of events by the same algorithm is low when the fraction of background "converter" primary particles is large and the particles are massive. For instance, the effect is insignificant when protons are involved but becomes significant already in the case of helium nuclei.

Thus, within our problem, all events in which primary cosmic-ray particles are recorded can be divided into two classes depending on the place where a particle first interacts inelastically with the substance of the device: events with interaction in the target belong to class I, while events with interaction in the convertor belong to class II.

To solve the problem, we used a computer model of the device, a model developed around the GEANT 3.21 software package [7]. In the course of the computational experiment, a statistical sampling imitating the passage of primary particles, helium, carbon, and calcium nuclei, through the device's aperture is drawn.

RESULTS OF THE COMPUTATIONAL EXPERIMENT

The results of the computational experiment are listed in Table 1. Here is the notation used in the table:

$P_{\mathrm{I}\to\mathrm{II}}$ is the probability of a primary particle that actually interacted in the target (class I) being erroneously classified as a particle that interacted in the converter (class II); and

Table 1

The Probabilities $P_{II\to I}$ and $P_{I\to II}$ (in %) of Classifying Primary Cosmic-Ray Particles by the Place of Their First Interaction with Matter

Type of primary particle		Helium			Carbon			Calcium		
Energy, GeV		79	158	315	79	158	315	79	158	315
		$P_{I\to II}$								
$P_{II\to I}$	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

$P_{II\to I}$ is the probability of a primary particle that interacted in the converter being erroneously classified as a particle that interacted in the target.

When analyzing the data in the table, one must bear in mind that the most important (for physical applications) computational example is the problem of removing from the data bank the data on primary particles that have interacted in the converter. The error associated with the removal from the general statistics of those events in which a particle interacted in the target but was erroneously referred to class II is of less importance to us. Hence, the classification method must, as its first objective, minimize the $P_{II\to I}$ error (or not allow a certain critical value to be surpassed). After the limit on $P_{II\to I}$ has been fixed, the method must minimize the classification error in the "opposite direction", $P_{I\to II}$.

Table 1 shows that when the error of the most important type $P_{II\to I}$ is insignificant (of order 10%), the error of the other type, $P_{I\to II}$, which only reduces the general statistics, amounts to 30–40%, i. e., with this method the limit on $P_{II\to I}$ that almost completely extinguishes "bad" events (interaction in the converter) leaves for analysis two-thirds of all statistics in "good" events (target). If we a priori fix $P_{II\to I}$ at 20%, the $P_{I\to II}$ error is also about 20%, which guarantees 80% for the target.

Figure 1 shows how $P_{I\to II}$ depends on $P_{II\to I}$. For the sake of comparison, the same figure shows, in a similar manner, the calculated errors of particle classification either according to the value of a single parameter $N = \sum_i I_i$ (the total signal from a detector when only one primary particle is recorded, which is proportional to the multiplicity of secondary particles produced by this particle) or by a single parameter $S = \sum_i c_i I_i$ (where the coefficients c_i depend on the distance to the axis of the beam of secondary particles [9]), which is used in the common one-dimensional method of restoring the primary energy in the NUCLEON project [9].

Figure 2 shows the calibration curves of the classification method as functions of the value of the threshold ε.

Figure 3 is a graphic interpretation of the inner mechanism used in the proposed primary-particle classification method. Depicted are, first, the expectation vectors M_1 and M_2 of the measured variables, i. e., the mean values of the signals from the matrix of the silicon stripped detector for each strip; and, second, the values (on a reduced scale) of the coefficients into which the signal from the respective strip is multiplied in the classification process [8]. Clearly, the spatial distribution of the signal from the secondary cascades produced by the interaction of primary particles in the target is broader and has not-so-sharp a peak than the distribution caused by the interaction in the converter.

It is these two features of difference in the two classes of the distributions that are effectively accounted for by formula (1) for the Bayes classifier, thus enabling a separation of primary particles by the position of the first-interaction point; the coefficient of the Bayes classifier, as a function of the strip number, has a local maximum near the central strip in the direction of which the primary-particle beam is oriented. As we move away from the central strip, the coefficient changes its sign, since on the average the cascade distribution referring to the interaction in the target begins to prevail over the cascade distribution in the

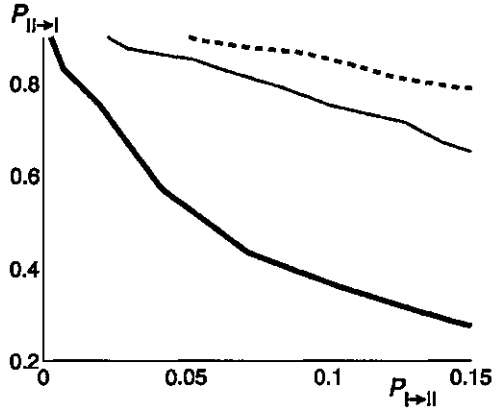

Fig. 1

Dependence of $P_{I\to II}$ on $P_{II\to I}$ for different primary-particle classification algorithms. The primary particles are helium nuclei with an energy of 79 GeV per nucleon. The heavy solid curve represents the results provided by the multidimensional recognition method, the light solid curve represents the results of classification according to the values of a single parameter S [9], and the dotted curve represents the results of classification by the values of a single parameter N (multiplicity of secondary particles).

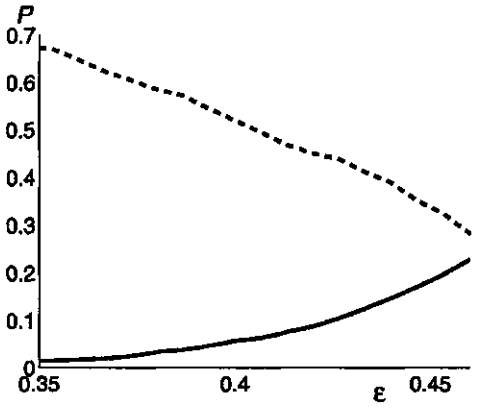

Fig. 2

Dependence of classification errors on the threshold ε for calcium nuclei with an energy of 158 GeV per nucleon. The solid curve represents the error of the $P_{II\to I}$ type, and the dotted curve represents the error of the $P_{I\to II}$ type. The value of ε is given in fractions of the interval of possible values of the Bayes classifier $t(\xi)$ (1).

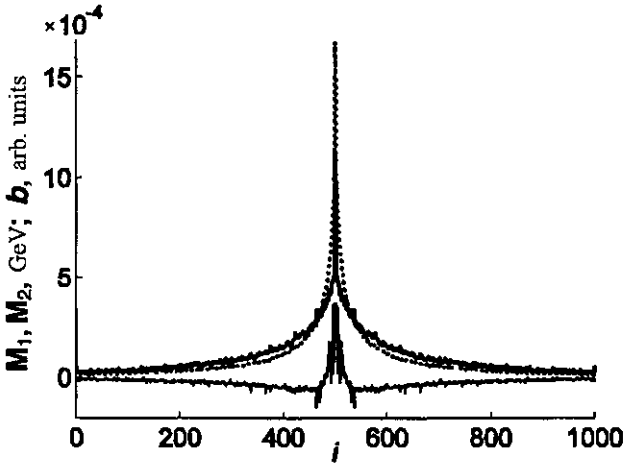

Fig. 3

Mean values of signals from the matrix of the stripped detector generated by the cascades produced by the interaction of primary particles (carbon nuclei, 315 GeV per nucleon) and the target's substance (the heavy solid curve represents the coordinates of M_1) and the converter's substance (the dotted curve represents the coordinates of M_2). Here i is the strip number; $i = 500$ corresponds to the center of the secondary-particle cascade. The light curve represents the values of the coefficient b in the classification method [8] on a reduced scale.

converter. As we move still farther from the beam center, the absolute value of the coefficient tends to zero,

i.e., the weight of the signals in the respective strips decreases due to the fact that the shapes of the two distributions become more and more alike.

CONCLUSION

The developed method of separating primary particles into two classes is not only effective and flexible but makes possible a graphic interpretation. It is simple to implement (say, with the MATLAB package) and, in view of the possibility of varying the value of the threshold in the computational formula, allows the researchers to select a variant of the method that suits his or her needs with respect to such parameters as the values of errors in referring the primary particles to each of the two classes and the ratio of these errors.

The computational experiments with model data done in order to solve a specific illustrative problem, the separation of primary particles according to the position of the point of first inelastic interaction, have shown that a sufficiently good classification quality (the total weighted classification error was no greater than 10% [8]) in the investigated energy range is observed both for light (helium) and heavy (calcium) nuclei. In contrast to the method of determining the primary energy, one-dimensional data-processing algorithms for solving the separation problem in the same NUCLEON project proved to be inadequate.

The conclusions drawn as a result of this research suggest the usefulness of the method in many applied problems of space physics where primary cosmic-ray particles need to be separated into two classes on the basis of measurements of a large number of physical variables.

REFERENCES

1. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edn., Boston, 1999.
2. F.A. Aharonian, A.A. Chilingaryan, A.K. Plyasheshnikov, and A.K. Konopelko, *Preprint Yerevan Phys. Inst.*, no. 1171(48), Yerevan, 1989.
3. T. Antoni, W.D. Apel, F. Badea, et al., *Astroparticle Phys.*, vol. 16, no. 3. p. 245, 2002.
4. D.M. Podorozhnyi, E.B. Postnikov, and L.G. Sveshnikova, *Yad. Fiz.*, vol. 68, no. 1, p. 51, 2005.
5. D.M. Podorozhnyi, E.B. Postnikov, L.G. Sveshnikova, and A.N. Turundaevskii, *Preprint MSU Research Institute of Nuclear Physics*, no. 2003-12/725, Moscow, 2003.
6. E.B. Postnikov, G.L. Bashindzhagyan, N.A. Korotkova, et al., *Izv. Ross. Akad. Nauk, Ser. Fiz.*, vol. 66, no. 11. p. 1634, 2002.
7. *GEANT User's Guide*, CERN DD/EE/83/1, Geneva, 1983.
8. E.B. Postnikov, *Preprint MSU Research Institute of Nuclear Physics*, no. 2004-23/762, Moscow, 2004.
9. N.A. Korotkova, D.M. Podorozhnyi, E.B. Postnikov, et al., *Yad. Fiz.*, vol. 65, no. 5. p. 884, 2002.

30 November 2004

Research Institute of Nuclear Physics