THE DEVELOPMENT OF STATISTICAL METHODS IN COSMIC RAY PHYSICS

Chilingarian A.A.
Yerevan Physics Institute, Markarian St.2,
375036, Yerevan, Armenia, U.S.S.R.

1. Introduction

The treatment of the experimental data in cosmic ray physics is performed by simulating the experimental situation. The simulation realizations (the so-called training sample) are applied to find the statistical procedure parameters. We shall perform our analysis with two situations often encountered, when we deal with data handling. The first one is the identification - a finite action problem, when the space of possible states of nature consists of finite number of categories, the second, an infinite action problem, is the estimation of a parameter, when the space of possible states covers a continuous interval of values. There is the n-dimensional variable $\vec{x}$ (experimental observation), which, as we hope, is related to the state of nature, and a decision is required on the type or energy of the particle traversing the installation. The observation vector $\vec{x}$ is a set of energy release commited by π -meson or proton traversing the transition radiation detector (Chilingarian, 1982) or a set of secondary electrons accompanying μ -meson of unknown energy traversing the spark calorimeter (Chilingarian and Ter-Antonian, 1982).

The peculiarities of the training sample application in cosmic ray physics are characterized by reducing the initial information: the replacement of the energy release by the geometrical average, etc. and by the Likelihood function form specifying with the conditional parametric independence approximation. Then the Likelihood function is determined iteratively by fitting the training sample with a function from the chosen class by some goodness of fit procedure.

The demerit of this treatment is that the assumption of feature independence is often erroneous, it is never known if the reduction is carried out in the best possible way, and it is unknown whether the chosen functional class fits the Likelihood function well.

Point one of this paper is to remind that the training sample is the only information one should proceed from in estimation and decision making, and one must select the statistical procedures to use it effectively. Such procedures must be Bayesian for one to take the most complete account of the a priori information and provide the minimum misclassifications; nonparametric not to impose any external structure on the data, and adaptive - to allow the data to speak for themselves as fully as possible.

2. The Bayesian Decision Rules and Probability Density Local Estimation

Suppose we want to determine the particle type according to the detector response. The Bayesian identification rule has a form

$$d(\vec{x}) = \begin{cases} p, & \text{if } P(p)\cdot p(\vec{x}/p) > P(\pi)\cdot p(\vec{x}/\pi) \\ \pi & \text{otherwise} \end{cases} \qquad (1)$$

where $P(p)$ and $P(\pi)$ are a priori probabilities (the relative portion of protons and π -mesons determined in the previous most precise experiment), $p(\vec{x}/p)$, $p(\vec{x}/\pi)$ are the conditional densities, i.e. the probabilities that the energy is released by proton or π -meson, respectively.

The key procedure of the Bayesian decision rule (1) is the estimation of the densities $p(\vec{x}/p)$ and $p(\vec{x}/\pi)$ by the training sample. Among numerous density estimation methods (Tapia and Thompson, 1978), one can outline the Nearest Neighbour (NN) rules. They are nonparametric in spirit, do not presuppose any structure and are economic in their greed for computer memory and time. Despite their simplicity, the NN rules are the surprisingly powerful discrimination technique. To reduce the variance of the NN estimator due to small sample size the alternate NN rule was proposed (Rabiner et al., 1979), first, to estimate the density for different number of NN

$$\hat{P}_{i,M} = \frac{i}{M\cdot V_{i,M}}, \qquad V_{i,M} = \frac{2\pi^{d/2} r_i^{d}}{d\,\Gamma(d/2)}, \quad i=1,K \qquad (2)$$

where $V_{i,M}$ is the volume of the region, containing i NN of the $\vec{x}$ vector, M is the total number of vectors in the training sample, d is the feature space dimension, r_i is the distance to the i -th NN. Second, to average the obtained estimates

$$\hat{P}_{K,M}(\vec{x}) = \sum_{i=1}^{K} \hat{P}_{i,M}(\vec{x})/K \qquad (3)$$

Point two of the paper consists in the proposal to use instead of arithmetic means the well known in robust statistic theory (ed. by Launer and Wilkinson, 1979) linear combinations of order statistics

$$\hat{P}_{[K],M}(\vec{x}) = \sum_{i=1}^{K} \alpha_i \hat{P}_{[i],M}(\vec{x}), \quad \sum_{i=1}^{K}\alpha_i = 1 \qquad (4)$$

where $\hat{P}_{[i],M}$ is the ranged set of density estimates. By the appropriate selection of the α_i -coefficients one may obtain more precise density estimates. The median estimate is the simplest version of such estimates

$$\hat{P}^{MED}_{[K],M}(\vec{x}) = \hat{P}_{[M/2+1],M} \quad . \quad M \text{ is an odd number} \qquad (5)$$

3. The Estimation by KNN Rule

Earlier (Chilingarian and Ter-Antonian, 1982) we considered the interval estimation of μ -meson energy by multi-layered installations. The recognition (estimation) was performed by the KNN classifier (1) in the 13 selected classes (energy intervals), the distance to the NN was calculated in a metric, sensitive to the correlations between the features. This method can be applied for energy spectrum studying, but there is another aspect, too: the single μ -meson energy determination at the given energy spectrum. In this case the training sample should be generated according to the spectrum and not be divided into classes; the estimation is again carried out by treating the ordered sets of the estimated vector NN (Cover, 1967)

$$\hat{E}(\vec{x}) = \sum_{i=1}^{K} \beta_i E_{[i]} \, , \quad \sum_{i=1}^{K} \beta_i = 1 \qquad (6)$$

where $E_{[i]}$ is an ordered set of energy values corresponding to K NN of $\vec{x}$ vector. In this case one can also apply the median estimates (5), but the more flexible approach is not connected with the a priori selection of β_i coefficients. It implies an optimization procedure, proceeding from the concrete viewpoint on the data structure. The optimization of β_i coefficients will lead to the minimization of the mean square errors of energy estimation integrated over the whole energy spectrum or over certain part of it (e.g. for high energies).

The last point of the paper appeals not to regard the statistics as a collection of dogmatic procedures, but to construct the new more precise and powerful methods proceeding from experimental purposes and available data.

References

Chilingarian, A.A., scientific report (1982).

Chilingarian, A.A. and Ter-Antonian, S.V., scientific report (1982).

Tapia, R.A. and Thompson, J.R., (1978), Nonparametric Probability Density Estimation, The Johns Hopkins University Press, Baltimore and London.

Rabiner, L.R. et al., (1979), IEEE Trans. on Acoust., Speech and Signal Proc. ASSP-27, p336.

Robustness in Statistics, ed. by Launer, L. and Wilkinson, G.N., Academic Press, New York, San Francisco, London, (1979).

Cover, T.M., (1967), IEEE Trans. on Inform. IT-14, p50.