A bootstrap method of distribution mixture proportion determination

A.A. CHILINGARIAN and G.Z. ZAZIAN
Yerevan Physics Institute, Alikhanian Brothers 2, SU-375036, Yerevan 36, Republica of Armenia, USSR

Received 23 April 1990
Revised 7 September 1990

Abstract: A simulation-based methodology for distribution mixture proportion determination is presented. A Monte Carlo study using Gaussian population data and cosmic-ray-experiments data proved the reliability of the methods proposed.

Key words: Bootstrap, classification, multivariate analysis, probability density estimation, distribution mixture, error rate estimation.

1. Introduction

The basic goal of this letter is to report development of a unified theory of statistical inference, based on nonparametric models, in which various nonparametric approaches (density estimation, Bayesian decision making, error rate estimation, feature extraction, sample control during handling, etc.) are used.

For a nonparametric model, that is a model in the form of a stochastic mechanism, whereby the data are generated, the underlying log-likelihood function cannot be given explicitly [1]. These, in a convenient way intractable statistical models often arise in modern high-energy physics experiments, where very sophisticated Monte Carlo models are being used.

A cosmic-ray-physics experiment, used as an example for the technique proposed, concerns the Primary Cosmic Rays (PCR) composition determination by the Extensive Air Showers (EAS) data [2].

A Monte Carlo study of the bootstrap method is also presented.

2. Classification of the distribution mixture

Let us consider the stochastic mechanism $(A, \mathscr{P})$ which generates the observation $\boldsymbol{v}$ in a multivariate feature space V ($\boldsymbol{v}$ is a d-dimensional vector of values measured experimentally, d is the dimensionality of the feature space, $\boldsymbol{v} \in V$). The basic states space A consists of alternative hypotheses or classes (for example, different primary nuclei). We know no law of nature like $(A, \mathscr{P})$, that is why, to determine the mutual probability measure on the direct product of A and V spaces, the total Monte Carlo simulation of the phenomenon under investigation is performed, including experimental data registration and handling.

The set of d-dimensional $\boldsymbol{u}$ vectors obtained in simulations is the analog of the experimentally measured values of $\boldsymbol{v}$. But, as opposed to experimental data, it is known to which of the alternative classes each of the events belongs. These 'labeled' events include a priori information about dynamics of the process under investigation, which is given in a nonparametric form, as finite samples. The sequence $\{u_i, t_j\}$, where $i = 1, \ldots, M_{\mathrm{TS}}$, $j = 1, \ldots, L$, t_j is a class index, we usually call a training set or sample (TS) which is also denoted by $(A, \tilde{\mathscr{P}})$.

0167-8655/90/$03.50 © 1990 — Elsevier Science Publishers B.V. (North-Holland)

Since both physical processes of particle production and those of registration are stochastic, only by careful measurement of probabilities can we gain an understanding of the phenomena under investigation. The high energy physics data analysis is uncertain in the sense that the probability distributions concerning the alternative hypotheses usually overlap significantly.

The only thing we can require when classifying a distribution mixture is to minimize the losses due to incorrect classification to some degree and to ensure use of a priori information completely. Such a procedure is the Bayes decision rule with nonparametric estimation of the multivariate probability density function, which, when using a simple loss function (the loss is zero in case of correct classification and it is the same for any error), takes the form:

$$\hat{A} = \eta(v, A, \tilde{\mathscr{P}}) = \operatorname{argmax}\{\hat{P}(A_i/v)\}, \quad i = 1, \dots, L \tag{2.1}$$

where $\hat{P}(A_i/v) = P_i \hat{P}(v/A_i)$ are a posterior densities, $\hat{P}(v/A_i)$ are conditional densities which are estimated by TS $(A, \tilde{\mathscr{P}})$ using one of many nonparametric methods available [3], L is the number of classes.

To estimate conditional densities, we used Parzen's method with automatic kernel width adaptation. In this method some probability density values are calculated which correspond to different values of method parameters. Then the sequence obtained is ordered and the median of this sequence is chosen as final estimate. Depending on the value of the probability density in the vicinity of $\boldsymbol{v}$, due to stabilizing properties of the median, each time we will choose an estimate with a width most fitting for that region [4]. The probability density is estimated by:

$$P(v/A_i) = 1/(2\pi^{d/2} h^d) \sum_{j=1}^{M_i} e^{-r_j^2/h^d} W_j, \quad i = 1, \dots, L \tag{2.2}$$

where d is the feature space dimensionality, M_i is the number of vectors in the ith TS class, r_j is the distance to the jth neighbor in the Mahalanobis metric:

$$r_j = (v - u_j)^{\mathrm{T}} R^{-1} (v - u_j), \tag{2.3}$$

where R is a sampling covariance matrix of the class to which u_j belongs, W_j are the event weights, h is the kernel width.

The classification methods, like all the statistical ones, include a procedure quality test as a necessary element. This stage beside all the others is also necessary for the determination of the mixture proportion. The most natural procedure quality estimate is the error probability which depends on both the degree of overlapping of alternative multivariate distributions and the decision rule being used (Bayes decision rules provide minimum error probability as compared to any other):

$$R_M^{\mathrm{B}} = E\{\theta[\eta(v, A, \mathscr{P})]\}, \tag{2.4}$$

where

$$\theta[\eta(v, A, \mathscr{P})] = \begin{cases} 0, & \text{for correct classification,} \\ 1, & \text{otherwise} \end{cases}$$

and E stands for mathematical expectation. The expectation is taken over all possible samples of volume M and over the whole d-dimensional space of measured values.

Since we do not know to which class the experimental vectors belong, we obtain an estimate of R_M^{B} via TS:

$$\hat{R}_M = \frac{1}{M_{\mathrm{TS}}} \sum_{i=1}^{M_{\mathrm{TS}}} \theta\{t_j, \eta(u_i, A, \tilde{\mathscr{P}})\}, \tag{2.5}$$

i.e., we classify the $\{u_i\}$ TS and check correctness of classification over the index of the class t_j, $j = 1, \dots, L$. However, as numerous investigations have shown (e.g., [5]), this estimate is systematically biased and hence, a cross-validation estimation is preferable:

$$R_M^{\mathrm{c}} = \frac{1}{M_{\mathrm{TS}}} \sum_{i=1}^{M_{\mathrm{TS}}} \theta\{t_j, \eta(u_i, A, \tilde{\mathscr{P}}_{(i)})\} \tag{2.6}$$

where $A, \tilde{\mathscr{P}}_{(i)}$ is a TS with a removed ith element, which is classified. This estimate is unbiased and has an essentially smaller r.m.s. deviation. The advantage of R_M^{c} is especially notable when the feature space has a high dimensionality [6].

Note, that we have the possibility to estimate the error probability of various types by classifying various TS classes $\{u_i, t_j\}$, $j = 1, \dots, L$.

By R_{ij} we denote the probability of classifying the ith class events as belonging to the jth class (misclassification).

Now let us estimate the a posterior fraction of various classes in the distribution mixture.

It is known [7] that the best estimate of the a posterior fraction (in case of uniform a priori information and absence of systematic errors) is the empirical fraction

$$P_i^e = M_i/M_{\text{tot}} \tag{2.7}$$

where M_i is the number of events classified as belonging to the class A_i, M_{tot} is the total number of events. It can be shown that with account of classification errors the fraction (proportion) can be obtained as the solution of the following set of linear equations:

$$\left(1-\sum_{j\neq i}^{L} R_{ij}\right)\hat{P}_i+\sum_{k\neq i}^{L}\hat{P}_k R_{ki}=P_i^e, \quad i=1,\dots,L. \tag{2.8}$$

In the first sum summation goes over j, in the second sum over k. All estimates of R_{ij} and P_i^e are obtained over one and the same TS using the same decision rules.

The accuracy of the estimates is defined by the TS size and the number of control data as well as by the value of the classification errors, which present the 'quality' of discrimination in the chosen feature subset. Note that the set (2.8) is a poorly defined system and at large values of classification errors the solutions of the set are unpredictable and hence, the choice of a feature combination providing a high percentage ($\geqslant$60%) of correct classification is a necessary preliminary stage.

3. The bootstrap procedure of fraction estimation

As we have shown in the previous section, to estimate the proportion of various classes in a distribution mixture, beside classification of a control sample by a TS, it is also necessary to calculate the misclassification coefficients, R_{ij}. The error in determination of the fraction is a function of the errors both from classification and in determination of R_{ij}.

The possibility to decrease the bias and variance of misclassification rates estimates was discussed in [8], where it was mentioned that it is possible to improve the accuracy of the R_{ij} estimates, if the TS size is large enough to separate the TS into independent subsamples.

Unfortunately, time consumption per model event generation increases abruptly with primary particle energy and the TS size is always limited for high energy events.

Thus, the problem of an efficient use of the information contained in training samples is very important for cosmic-ray and accelerator physics, since the classical sampling models do not allow to extract the whole information carried by a sample.

The methods of sample control during handling are widely used in the last few years. One of these is the leave-one-out-for-a-time test considered in the previous section, which allows to decrease the sample bias. A more efficient procedure actively developing in both applied and theoretical respects in the last decade is the bootstrap which lies in replication of the initial sample very many times by means of random sampling with replacement. The thus obtained conditionally independent bootstrap-replicas in many respects stand for independent samples from the general population (under the condition of sufficiently large size of the initial sample). In fact, the bootstrap substitutes the unknown general population by a single sample.

The theoretical basis of the bootstrap method is the analog of the central limit theorem (CLT) proved in [9]:

$$P\{\sqrt{B}(\mu_B-\mu_M)<tS_M\,|\,x_1,\dots,x_M\}\to\Phi(t), \tag{3.1}$$

when $M,B\to\infty$, $x_1,\dots,x_M$ are independent, identically distributed (iid) random quantities, $\Phi(t)$ is a normal (Gaussian) distribution, μ_M and S_M are sample estimates of the first and the second moments,

$$\mu_B=\sum_{j=1}^{B}\mu_j^*/B,$$

and

$$\mu_j^*=\sum_{i=1}^{M}x_i^{(j)}/M$$

is the jth bootstrap replica's mean. Moreover, analogies between sampling and the bootstrap are valid also for many other statistics. Referring to [10], we shortly summarize the main idea of the new procedure: the bootstrap-moments (denoted by E_*, σ_*) are introduced, which in many cases substitute the statistical moments calculated according to a distribution function (in most cases of interest it is unknown).

Owing to the fact that the bootstrap is very important for high-energy physics, and to investigate its possibilities for finite samples and a limited number of bootstrap replicas, we have carried out an investigation with the purpose to calculate the bootstrap expectation $(\mu_j^*-\mu_M)$, a 'butstrap' CLT test, and calculation of bootstrap expectations of the standard deviation of the mean iid random variables, $\delta_*^2=\sigma_*^2/M$. To do this we used samples from the standard, normal distribution $N(0,1)$; the sample size varied between 25 and 1000, the number of bootstrap replicas in a series was from 10 to 2000. The mean was calculated for each bootstrap replica, and for each bootstrap series the bootstrap-estimate of the mean standard deviation, δ_*, was calculated.

A round of recalculations including 100 series of the same size has been carried out using different initial samples; the obtained data were averaged and the mean-square deviations were calculated. The results of investigations, which are present in Table 1, illustrate the validity of 'butstrap' CLT and consistency of using the bootstrap expectation. Although the mathematical theorems were proved for the asymptotic cases $M, B \to \infty$, even with small sample sizes and small numbers of bootstrap replicas ($M, B = 50$), the obtained estimates fit to the expected theoretical ones.

There are two ways of distribution mixture coefficient estimation: (i) to obtain the bootstrap estimate of the misclassification coefficients R_{ij}^c, then classify and estimate the fraction, or (ii) carry out fraction estimation over each bootstrap

Table 1
Bootstrap expectations and bootstrap standard deviations of sampling statistics

	B	10	50	100	200
	$E_*\{\mu_B-\mu_M\}$	−0.0152	0.0031	−0.0048	−0.0003
$M=25$	$\sigma_*\{\mu_B\ \ \mu_M\}$	0.0639	0.0251	0.0174	0.0160
$\delta_{25}\ \ 0.2$	$E\{\delta_*\}$	0.1891	0.1974	0.1929	0.1977
	$\sigma\{\delta_*\}$	0.0560	0.0300	0.0031	0.0028
	$E_*\{\mu_B-\mu_M\}$	−0.0024	−0.0023	0.0003	−0.0001
$M=50$	$\sigma_*\{\mu_B-\mu_M\}$	0.0402	0.0227	0.0149	0.0097
$\delta_{50}\ \ 0.1414$	$E\{\delta_*\}$	0.1481	0.1398	0.1396	0.1395
	$\sigma\{\delta_*\}$	0.0286	0.0182	0.0167	0.0154
	$E_*\{\mu_B-\mu_M\}$	−0.0171	−0.0010	0.0004	−0.0008
$M=100$	$\sigma_*\{\mu_B-\mu_M\}$	0.0323	0.0152	0.0101	0.0066
$\delta_{100}=0.1$	$E\{\delta_*\}$	0.0897	0.0959	0.1000	0.0988
	$\sigma\{\delta_*\}$	0.0212	0.0107	0.0097	0.0086
	$E_*\{\mu_B\ \ \mu_M\}$	0.0038	0.0017	0.0001	0.0000
$M\ \ 200$	$\sigma_*\{\mu_B-\mu_M\}$	0.0231	0.0107	0.0082	0.0048
$\delta_{200}=0.0707$	$E\{\delta_*\}$	0.0593	0.0692	0.0694	0.0700
	$\sigma\{\delta_*\}$	0.0154	0.0078	0.0063	0.0049
	$E_*\{\mu_B-\mu_M\}$	−0.0018	0.0007	0.0004	0.0003
$M-500$	$\sigma_*\{\mu_B-\mu_M\}$	0.0115	0.0072	0.0040	0.0032
$\delta_{500}=0.0447$	$E\{\delta_*\}$	0.0430	0.0452	0.0442	0.0446
	$\sigma\{\delta_*\}$	0.0095	0.0043	0.0033	0.0024
	$E_*\{\mu_B-\mu_M\}$	0.0038	0.0001	0.0002	0.0003
$M=1000$	$\sigma_*\{\mu_B-\mu_M\}$	0.0079	0.0050	0.0030	0.0022
$\delta_{1000}=0.032$	$E\{\delta_*\}$	0.0322	0.0317	0.0316	0.0315
	$\sigma\{\delta_*\}$	0.0073	0.0033	0.0022	0.0017

replica, then obtain the fraction and the standard deviation bootstrap expectation. The second way is preferable, because obtaining of the standard deviation in the first case is time-consuming: the error propagation formulae obtained by the REDUCE symbolic manipulation program occupy several standard sheets in case of classification into four classes.

Finally, let us formalize the bootstrap method of the distribution mixture coefficient estimation. Let us define the solution of the set (2.8) as:

$$\boldsymbol{P} \equiv \boldsymbol{P}\{\hat{P}_1, \ldots, \hat{P}_l\} = f\{\upsilon, \tilde{\mathscr{P}}, \eta(\upsilon, \Lambda, \tilde{\mathscr{P}})\}. \quad (3.2)$$

This solution is a complicated function of experimental data and the TS as well as the decision rule η being used. By several TS bootstrap replicas we calculate the bootstrap expectation and the bootstrap standard deviation of the mixture coefficients $\hat{P}_i$, which are used as estimates of the fraction of different nuclei groups in the primary flux.

4. Results of calculations

To test the method, the generated events were grouped in two. The first were used to create a TS and the second as control events. The EAS characteristics—number of electrons, number of muons, age parameter (N_e, N_μ, S)—were used in the events classification. The TS consisted of four classes in accordance with the primary nuclei type (p: protons, α-particles, CNO: nuclei with atomic number $A = 7$–16, H: nuclei with atomic number $A = 24$–27, and VH: nuclei with atomic number $A = 50$–56).

Table 2 presents the Bayes error matrix obtained as a result of a leave-one-out test over TS. The diagonal elements of this matrix show the probability of correct classification and the nondiagonal elements represent the probability for misclassifications. It is seen from Table 2 that the percentage of correct classifications amounts to 70–80%. Classification of 'boundary' groups (protons and iron group nuclei) is essentially better than that of the intermediate groups.

Table 2
The Bayes error matrix obtained by the leave-one-out method

	p	CNO	H	VH
p	0.798	0.102	0.067	0.033
CNO	0.127	0.688	0.105	0.080
H	0.072	0.113	0.691	0.124
VH	0.034	0.090	0.150	0.726

Table 3
Recovered factions of four groups of nuclei (W_{in} is a 'true' fraction, W_{out} is a recovered one)

	N_{TS}	W_{in}	$E_*\{W_{out}\}$	$\sigma_*\{W_{out}\}$
p	200	0.370	0.345	0.038
CNO	188	0.272	0.299	0.067
H	194	0.168	0.232	0.057
VH	163	0.189	0.194	0.019

Table 3 shows the recovered nuclei fractions obtained by classification of control events for one interval over N_e. The errors presented are obtained by the bootstrap procedure. As is seen from this table, the proposed method allows to determine the fraction of protons and iron nuclei in the incident flux with quite a good accuracy.

References

[1] Diggle, P.J. and R. Gratton (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. B* 46, 193.

[2] Rich, J., D.L. Owen and M. Spiro (1987). Experimental particle physics without accelerators. *Phys. Reports* 151 (5,6).

[3] Devroy, L. and L. Gyorfi (1985). *Nonparametric Density Estimation. The L1 View.* Wiley, New York.

[4] Chilingarian, A.A. (1989). Statistical decisions under nonparametric a priori information. *Comp. Phys. Comm.* 54, 381–390.

[5] Toussaint, G.T. (1974). Bibliography of misclassification. *IEEE Trans. Information* 20, 472–478.

[6] Chilingarian, A.A. and S.Ch. Galfayan (1984). Calculation of Bayes risk by *K*NN method. *Stat. Problem of Control.* Vilnius, 66–76.

[7] Hey, J.D. (1983). *An introduction to Bayesian Statistical Inference.* Martin Robertson.

[8] Jain, A.K., R.C. Dubes and C.-C. Chen (1987). Bootstrap techniques for error estimation. *IEEE Trans. Pattern Anal. Machine Intell.* 9, 628.

[9] Bickel, P.J. and D.A. Freedman (1981). Some asymptotic theory for the bootstrap. *Ann. Stat.* 9, 1198.

[10] Efron, B. (1982). The Jackkife bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.