American Journal of Mathematics and Statistics 2013, 3(6): 346-348
DOI: 10.5923/j.ajms.20130306.07

The Problem of Statistical Learning Decision-Making for the Small Sample Size in Geoinformation Monitoring

F. A. Mkrtchyan

Department of Informatics, Institute of Radioengineering and Electronics, Russian Academy of Science, Fryazino, 141190, Russia

Abstract The application of geoinformation monitoring means often involves statistical decision making about the presence of one or other phenomenon on a surveyed part of the Earth surface. One of the features of information acquisition conditions for such a decision is that it is impossible to obtain large statistical samples. Therefore, the development and research of optimal algorithms for distinguishing between random signals with samples of limited size under conditions of parametric a priori uncertainty is a topical problem. In the present work, a generalized adaptive learning algorithm is developed for statistical decision making concerning exponential families of distributions under conditions of a priori parametric uncertainty for small sample sizes. Numerical examples are presented. The efficiency of the optimal procedure developed is demonstrated in the case of small samples. "The reported study was partially supported by RFBR, research project No.13-07-00146".

Keywords Statistical Decision, Small Samples, Exponential Classes, Geoinformation Monitoring, Spottiness

1. Introduction

The development of geoinformation monitoring systems requires a solution of some problems concerning the organization of flows of measurement data. One of important problems here is making a statistical decision about the presence of one or other phenomenon on a surveyed part of the Earth surface. One of the features of information acquisition conditions for such a decision is that it is impossible to obtain large statistical samples. Therefore, there is a need in the development and research of optimal algorithms for making statistical decisions for small-size samples under information constraints.

When the number of observations is large enough, the problem is solved by the method of estimating the parameters of probability distributions, which is effective when the size of samples used for estimating the parameters indefinitely increases. Under a restricted size of samples, the decision making rule obtained by the method of estimating parameters does not satisfy the necessary optimality conditions, i.e., the constancy of the mean probability of error of the first kind and unbiasedness.

In the present work, we develop an adaptive learning algorithm for statistical decision making for exponential families of distributions under conditions of a priori parametric uncertainty for small sample sizes[1, 2].

* Corresponding author:
ferd47@mail.ru (F. A. Mkrtchyan)
Published online at http://journal.sapub.org/ajms

2. Statement of the Problem

Very often, one encounters the following problem: under what class a measured random variable should be classified when a full probability description of these classes is unknown. The latter fact does not allow one to use classical results of the theory of statistical decision making. A decision can be obtained only by means of learning samples.

Let ξ, η, and ζ be independent random variables and $f_\xi(x/\omega_0)$, $f_\eta(y/\omega_1)$, and $f_\zeta(z/\omega)$ be the distributions of probabilities $\omega_0, \omega_1, \omega \in \Omega$. There are two alternatives for the parameter ω: $H_0: \omega = \omega_0$ and $H_1: \omega = \omega_1$.

Errors of the first kind are given by

$$\alpha(\varphi, \omega_0, x^*, y^*) = \int \varphi(x^*, y^*, z^*) f\omega_0(z^*) dz^*$$

and

$$\beta(\varphi, \omega_0, x^*, y^*) = \int [1 - \varphi(x^*, y^*, z^*)] f\omega_1(z^*) dz^*.$$

The optimality conditions are

1. $\acute{\alpha} = \alpha_0$ (the constancy of errors of the first kind) and
2. $1 - \beta > \acute{\alpha}$ (unbiasedness).

In this work, we develop a mathematical apparatus and propose a generalized adaptive procedure for solving the problem of learning to distinguish between random variables from exponential families of distributions with unknown parameters for small-size samples under information constraints.

We show that available discrimination learning procedures in which first one estimates the parameters and then makes a choice between hypotheses do not satisfy the above requirements on optimal procedures.

When the number of observations is large enough, the problem is solved by the method for estimating the parameters of probability distributions, which is effective when the size of samples used for estimating the parameter indefinitely increases. When the sample size is limited, the decision making rule obtained by the method of estimating parameters does not satisfy the necessary optimality conditions: the constancy of the mean probability of error of the first kind and unbiasedness.

3. Statement of the Problem

The classical method of solving the problem is based on the sufficiently well developed theory of point estimates for unknown parameters of probability distributions. In the present problem, one obtains estimates for the parameters ω_0, ω_1, θ_1, and θ_0 from n_0 observations $x^* = (x_1, x_2, \ldots\ldots\ldots x_{n_0})$ and n_1 observations $y^* = (y_1, y_2, \ldots\ldots\ldots y_{n1})$ of the random variables ξ and η, respectively.

Next, the method for constructing a decision rule is based on the Neyman--Pearson fundamental lemma: construct a likelihood relation

$$L(z^*/\theta_1,\theta_0) = [f(z_1,z_2,\ldots z_n/\ \theta_1)/\ f(z_1,z_2,\ldots z_n/\ \theta_0)] > C(\theta_1,\theta_0)$$

and choose a threshold $C(\theta_1,\theta_0)$.

1. $\theta_1 > \theta_0$ $\{s < t(n_0/n_1), s < n_0/G^{-1}{}_n(1 - \alpha_0)$,
2. $\theta_1 < \theta_0$ $\{ s > t(n_0/n_1)$, $s > n_0/G^{-1}{}_n(1- \alpha_0)$,

$s = x/z$, $t = y/z$,

$G^{-1}{}_n = [1/(n-1)!] \int \exp(-z)\ z^{n-1}\ dz$,

θ_1 and θ_0 are point estimates for ω_1 and ω_0.

The domains of making a hypothesis H_1 are shown in Fig. 1.

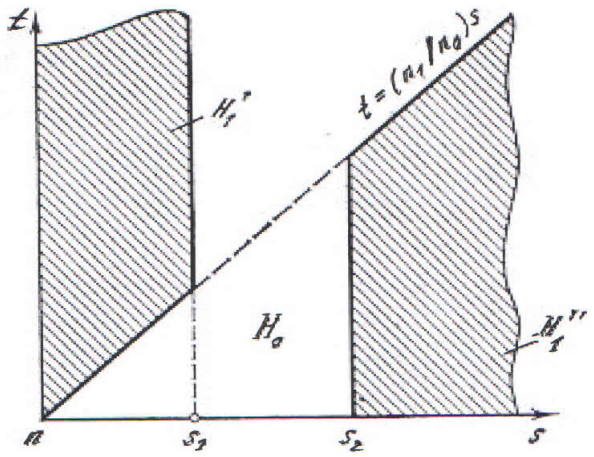

Figure 1. Domains of making hypotheses for the classical decision rule

The graphs of probabilities for making a correct decision and errors of the first kind are shown in Fig. 2.

Figure 2 shows that the probability of error of the first kind is greater than the admissible value α_0 by a factor of two and a half.

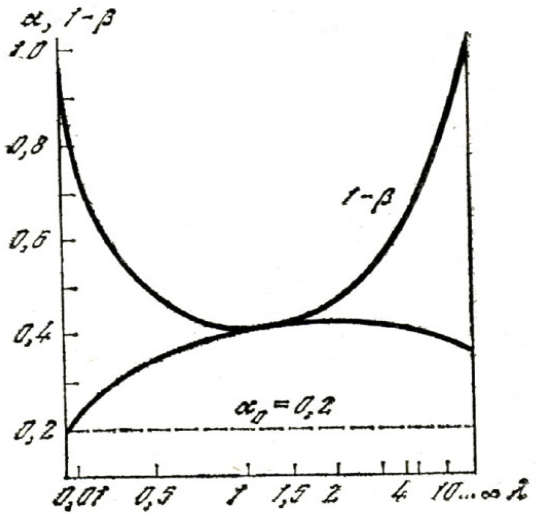

Figure 2. Probabilities of making a correct decision and errors of the first kind

4. Decision Rule Satisfying Necessary Optimality Conditions

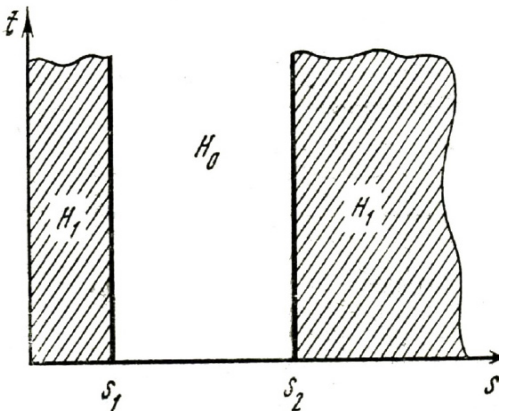

Figure 3. Domains of making hypotheses for an optimal decision rule

An optimal decision rule in the class of decision rules shown in Fig. 3 is given by

$$\{\max_{0<\alpha<\infty} \alpha\,(s_1, s_2 / \lambda) = \alpha_0\ ,\ \max_{0<\lambda<\infty}\ \min_{(s_1,s_2)}\ D(s_1, s_2 / \lambda)$$

$$D(s_1, s_2 / \lambda) = 1- \beta(s_1, s_2 / \lambda)$$

$$\alpha(s_1, s_2 / \lambda) = P_0(s_1) + [P_0(\infty) - P_0(s_1)]$$

$$D(s_1, s_2 / \lambda) = P_1(s_1) + [P_1(\infty) - P_1(s_2)]$$

$$P_i(k) = \iint f_i(s,t)\,ds\,dt\ ,\ \text{where } i = 0, 1.$$

Applying Lagrange's method of undetermined multipliers, we obtain the following set of equations for the optimal thresholds s_1 and s_2:

$$\{ (s_1/s_2)n_0 = [(s_1 + 1)/(s_2 + 1)](n+n_0)$$

$$\{ \{1-\prod [(n_0 + j)/n!][[(n_0+n-1)(x_1+1)(n_0-1) - n]/(x_1+1)(n_0+n-1)] -$$

$$\{ -[[(n_0+n-1)(x_2+1)(n_0-1) - n]/(x_2+1)(n_0+n-1)]] - \alpha_0 = 0.$$

The probabilities of making a correct solution and errors of the first kind are shown in Fig. 4.

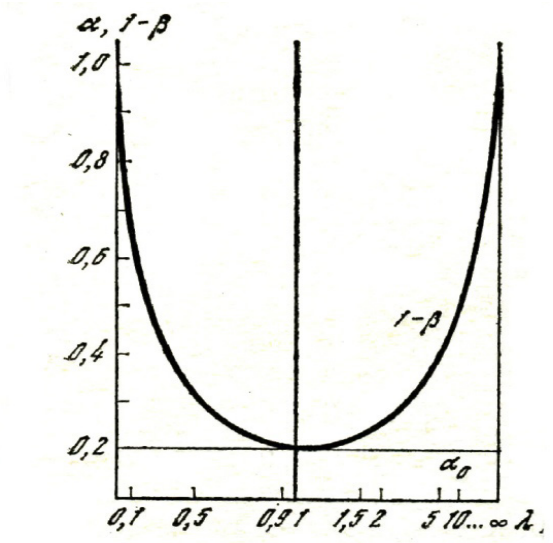

Figure 4. Probabilities of making a correct solution and errors of the first kind

The procedure proposed satisfies the following necessary conditions: (1) the constancy of the mean probability of errors of the first kind α and (2) the unbiasedness condition $1-\beta < \alpha$.

The method described allows one to obtain decision rules for particular distributions from the exponential family of distributions.

5. Application

The analysis of statistical characteristics of "spottiness" was carried out for three types of areas of the Pacific ocean. These statistical characteristics were obtained for the most informative thresholds. The statistical characteristics of spottiness for the same areas were chosen on the basis of the criteria of minimal value of the correlation coefficient for a joint sample of positive and negative spots. The analysis of these characteristics shows that the statistical characteristics of spottiness for regions of moderate sea roughness and storm regions coincide. The correlation coefficient ρ_{min} attains its minimum in the case of most informative thresholds. However, for a quiet region, the situation is different[2].

Figure 5 illustrates the operation of the automated system in the mode of monitoring the surface temperature of Northern Atlantic based on data from «Kosmos-1151» (April 8-14, 1980) satellite. The system allows one to obtain temperature maps on a rather rarefied grid of satellite trajectories. The dots on the map denote areas where ship measurements were carried out. The analysis of satellite and contact measurements shows that the satellite data on the sea surface temperature are systematically understated compared with the ship measurement data, the difference being about 1.6 K on average.

The root-mean-square deviation of satellite data from the ship over the entire sample data is 3.3 K. The dotted lines on the map indicate areas where the difference between the ship and satellite measurements exceeds 4 K. It is remarkable that all these areas were characterized by high cloudiness (by weather forecast data). If we disregard these areas, then the root-mean-square deviation of the satellite measurements of temperature from the ship measurements amounts to 1.4.

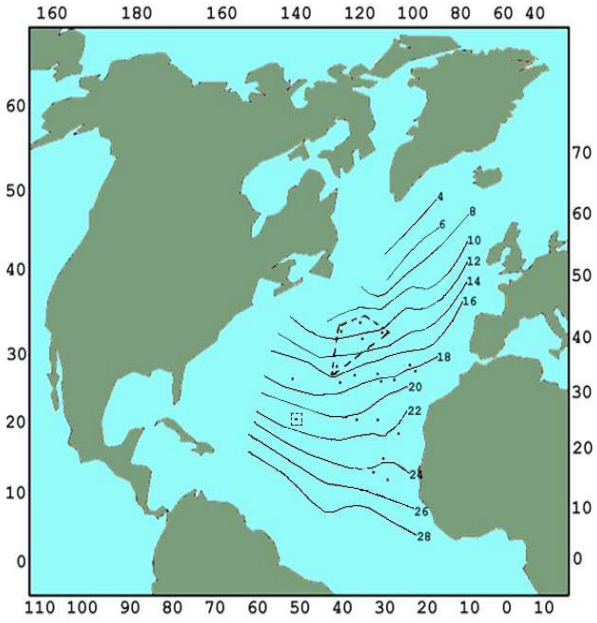

Figure 5. Temperature map of Northern Atlantic

It follows from the aforesaid that the statistical characteristics of "spottiness" of microwave brightness temperature can be used to detect and classify various phenomena on the surface of the ocean that differ by the degree of sea roughness.

6. Conclusions

The analysis of empirical distributions of the spottiness of microwave brightness temperature has shown that, in most cases, the (l^+, l^-) characteristics are consistent with the exponential distribution, while the amplitude characteristics are consistent with the normal distribution. Therefore, to detect and classify phenomena on the surface of the ocean, one should apply optimal algorithms for the computer training to making statistical decisions for the aforesaid distributions.

REFERENCES

[1] Mkrtchyan, F.A., *Optimal Distinguish of Signals and Monitoring Problems,* Moscow: Nauka, 1982 (in Russian).

[2] Armand, N.A., Krapivin, V.F., and Mkrtchyan, F.A., *Methods of Data Processing of Radiophysical Research of an Environment,* Moscow: Nauka, 1987 (in Russian).