PARAMETRIZATION OF PROBABILISTIC DISTRIBUTIONS FOR IMAGE RECOGNITION BASED ON NORMALIZING TRANSFORMATIONS

K.T. Protasov

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences,
Received March 16, 1994

Within the framework of the Bayes approach to constructing decision rules for image recognition in the space of definitions of high dimensionality, we solve the problem of retrieving multidimensional conditional functions of the density of classes, based on normalizing transformations. These transformations satisfy two conditions. The first of them demands that estimated distributions of a separate feature agree with actual single–dimensional distributions at a practically acceptable reliability. The second condition states that the approximating distribution should describe statistical relations between the components of the vector of observations.

At a subject oriented deciphering of aerospace video information one of the central problems is that of segmenting spectrozonal images of the underlying surface and cloud cover. In the course of such a deciphering the digitized initial image is decomposed into separate fragments, which are then treated as observed signals. The algorithm of image recognition, expressed as a decision rule, starts from such an observed signal and identifies the situation which had initially produced the observed class of input data.

Among the numerous approaches to constructing decision rules of image recognition, one of the most formalized is the Bayes approach, which in fact makes testing of statistical hypotheses. It is within the framework of this approach, methodologically based on the theory of testing statistical hypotheses and on the theory of synthesis of adaptive information systems, that optimal decision rules for making decisions were obtained for situations when full *a priori* information is available under the condition of minimum of the function of losses. However, direct application of such an approach faces serious difficulties, mostly because in the majority of practical tasks of image recognition *a priori* information is incomplete. Indeed, we know neither the *a priori* distribution of images (classes), nor the conditional functions of density, which describe situations to be recognized (the latter is more important).

The quality criterion of the algorithm of recognition is determined by averaging the payment matrix. Elements of this matrix are set quite arbitrarily, providing only for the function of losses to be convex. Therefore, it is admissible to set the probabilities starting with the condition of a maximum *a priori* indefinitness according to Laplace–– Bayes postulate. As to the conditional functions of probability density distribution, these should be reconstructed only from teaching sets of classes, which are of a limited bulk. Besides, in the problems of image analysis the observed signals have high dimensionality, exceeding the bulk of sampled data.

In this connection it seems to be worthwhile to consider the problem on reconstructing multidimensional conditional functions of density from samples of teaching sequences, while describing observed values or their frequency distributions by a flexible system of mathematical formulas.

We start from a standard formulation of the problem of image recognition in its statistical interpretation.[1]

Probabilistic measures with their *a priori* distributions of classes $P(\lambda)$ and conditional functions of probability density $f(\mathbf{x}/\lambda)$, $\lambda \in \Lambda$, $\mathbf{x} \in E^n$ are defined in a Euclidean n–dimentional space of observations E^n, with the elements being the fragments of digitized images, presented in the form of a vector $\mathbf{X} \in E^n$ and in the space of hypotheses $\Lambda = \{1, \ldots, L\}$, where L is the number of hypotheses (images). A simple matrix of losses due to decisions taken $(1 - \delta_{\lambda\mu})$ is also defined there, where $\delta_{\lambda\mu}$ is the Kronecker symbol. The Bayes decision rule for selecting the hypothesis $\lambda \in \Lambda$ from a series of mutually exclusive hypotheses, optimal in the sense of minimum of average losses, has the form

$$u(\mathbf{x}) = \arg \max_{\lambda \in \Lambda} P(\lambda)\, f(\mathbf{x}/\lambda)\,, \qquad (1)$$

where the solution u belongs to the space Λ as well.

This decision rule may also be presented in an equivalent form, as a likelihood ratio compared to a threshold.

If we have a sample of the observed vectors $\mathbf{X}_1^{\lambda}, \ldots, \mathbf{X}_{N_\lambda}^{\lambda}$, classified by a "teacher", where N_k is the size of the sample, $\lambda \in \Lambda$, the average risk may be estimated by the empirical risk as follows:

$$R = \sum_{\lambda \in \Lambda} \frac{1}{N_\lambda} \sum_{k=1}^{N_\lambda} P(\lambda)\, I\, \{\lambda = \arg \max_{\mu \in \Lambda} P(\mu)\, f(\mathbf{X}_k^{\lambda}/\mu)\}\,, \qquad (2)$$

where $I\{\text{true}\} = 0$, $I\{\text{false}\} = 1$ is the characteristic function; N_λ is the sample size of class $\lambda \in \Lambda$.

It is natural to demand that the parametric functions of density $f(\mathbf{x}/\lambda)$, $\lambda \in \Lambda$ to be reconstructed for decision rule (1) must satisfy two conditions. First, the estimates of distributions of separate features must agree with the true one–dimensional distributions sufficiently reliable for practical purposes. Second, the approximating distribution must describe, at least to some extent, the statistical relations between the components of the random vector. In this case, a wide class of multidimensional parametric

distributions may be obtained, following the idea of transforming the distribution of the observed variable into a Gaussian distribution.

Generally the idea of using transformations of random variables to construct multidimensional parametric functions of density, which have an increased approximating capability to describe probabilistic properties of the selected data sets, may be illustrated as follows.

Let it be necessary to construct, to an accuracy of some parameters, a transformation

$$\mathbf{y} = \mathbf{y}(\mathbf{x}), \tag{3}$$

which transforms some random variable $\mathbf{X} \in E^n$ with its distribution $F(\mathbf{x})$ and the function of probability density $f(\mathbf{x})$ into a random vector value $\mathbf{Y} \in E^n$ with the distribution $G(\mathbf{y})$ and a function of probability density $g(\mathbf{y})$. After that the function of density of the random vector $\mathbf{X}$ is found using the following standard operation:

$$f(\mathbf{x}) = g(\mathbf{y}(\mathbf{x})) \left| \frac{D\,\mathbf{y}(\mathbf{x})}{D\,\mathbf{x}} \right| , \tag{4}$$

where $\left| \frac{D\,\mathbf{y}(\mathbf{x})}{D\,\mathbf{x}} \right|$ is the Jacobian of transformation (3).

The approximating capability of the function $f(\mathbf{x})$ thus obtained increases, because, in addition to the parameters of a "simpler" family $g(\mathbf{y})$ it also depends on the parameters of transformation (3).

To implement the representation (3) let us make use of transformations brought by continuous integral distribution functions (see Ref. 2, p. 187), which transform the random vector $\mathbf{X}$ to $\mathbf{Y} \in E^n$ with all its necessary properties. Consider the simplified version of that transformation, presenting it as a following two stage procedure.

At its first stage we define the transformation, performed with a one–dimensional distributions, which transform the components of random vector $\mathbf{X}$ into the components of an auxiliary vector $\mathbf{Z}$, $\mathbf{Z} \in E^n$ so that each component of the vector $\mathbf{Z}$ is uniformly distributed over the interval [0, 1]

$$Z^i = F_i(X^i) , \qquad i = 1, \ldots, n , \tag{5}$$

where $F_i(X^i)$ is the marginal distribution of the component X^i of vector $\mathbf{X} \in E^n$.

At the second stage we set strictly increasing functions of distributions $G_i(y^i)$, such that

$$G_i(Y^i) = Z^i , \qquad i = 1, \ldots, n , \tag{6}$$

where Z^i is the auxiliary random value from expression (5).

Transformation (3), which provides for the component–by–component transition from vector $\mathbf{X}$ to vector $\mathbf{Y}$, will, with the account of Eqs. (5) and (6), acquire the form

$$y^i = G_i^{-1}(F_i(x^i)) , \qquad i = 1, \ldots, n , \tag{7}$$

where $G_i(y^i) = \int_{-\infty}^{y^i} g_i(t)\,\mathrm{d}t$, $i = 1, \ldots, n$, and $g_i(t)$ is the one–dimensional function of probability density of ith component of vector $\boldsymbol{Y}$. Below we assume the functions $G_i(y^i)$, $i = 1, \ldots, n$ in Eqs. (6) and (7) to be Gaussian distributions, for clarity reasons.

In order to compensate, at least to a certain extent, for imperfection of transformations (7) and to account for the dependence of the components of vector $\mathbf{Y}$, we assume the common distribution of the components $Y^1, \ldots, Y^n$ to be a multidimensional Gaussian distribution with the correlation matrix Σ and the vector of mean values $\boldsymbol{\mu}$. Then, with the account of Eqs. (4) and (7), the sought function of density of the initial vector $\mathbf{X}$ is written in the following form:

$$f(\mathbf{x}) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^n} \exp\{-\frac{1}{2}(\boldsymbol{G}^{-1}(\boldsymbol{F}(\mathbf{x})) - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} \times$$

$$\times (\boldsymbol{G}^{-1}(\boldsymbol{F}(\mathbf{x})) - \boldsymbol{\mu})\} \prod_{i=1}^{n} \left\{ \frac{d\,G_i^{-1}(F_i(x^i))}{d\,F_i} f_i(x^i) \right\}, \tag{8}$$

where $\boldsymbol{G}^{-1}(\boldsymbol{F}(\mathbf{x})) - \boldsymbol{\mu} = G_i^{-1}(F_i(x^i)) - \mu^i)$, $f_i(x^i) = \frac{d\,F_i(x^i)}{d\,x^i}$, $i = 1, \ldots, n$, and symbol T denotes the operation of transposition.

Thus, performing the component–by–component normalization and accounting for the intercomponent relations, we obtain a parametric presentation of the function of probability density to describe the images to be recognized by the decision rule (1).

When using expression (8) one should bear in mind that $G_i^{-1}(\cdot)$ is a function inverse to the Gaussian distribution, which does not have a simple analytical form. However, accurate enough approximations are known both for the integral of probabilities, and for the function inverse to it. For example, the λ–distribution, introduced by J. Tyuki[3] approximates the inverse function of continuous distributions to a high degree of accuracy

$$y = G^{-1}(z) = \lambda_1 + (z^{\lambda_3} - (1 - z)^{\lambda_3}) \,/\, \lambda_2 , \tag{9}$$

where $z \in [0, 1]$, the average and the median of Y coincide and are equal to λ_1 , $\lambda_1 = M[Y]$ (M is mathematical expectation), λ_1 is the parameter of locality, λ_2 is the parameter of scale, and λ_3 is the parameter of shape.

In particular, to determine the inverse function of a normal distribution with an average μ and variance σ^2, we have

$$\lambda_1 = \mu, \quad \lambda_2 = 0.1975/\sigma , \quad \lambda_3 = 0.1349 .$$

Besides, to determine the multidimensional function of density $f(\mathbf{x})$ by formula (8) one needs to reconstruct one–dimensional probabilistic characteristics $F_i(x^i)$, $f_i(x^i)$, $i - 1, \ldots, n$.

To solve that problem from selected data it is natural to use the system of Pearson curves[4] or to approximate the selected distributions by splines. The described procedure of constructing the functions of density for a decision rule for image recognition is significantly simplified if one uses already known normalizing transformations.[4,5]

Among such normalizing transformations one should note those proposed by Johnson to approximate two wide classes of unimodular and bimodular distributions.[5,6] The procedure of reconstructing parametric function may then be presented as a two–step operation.

At the first step the Johnson transformation is selected for each feature, which is in agreement with the true unknown distribution to a sufficient degree of reliability, while at the second one estimates of the coefficients of

correlation between the features are made to describe statistical relations between the transformed components. After that the joint distribution of the components of the random vector is presented in standard form.

Consider the first stage in more detail. Let $X \in E^1$ be a random variable, for which we try to choose the Johnson distribution. It may be expressed in a general form

$$\xi = \gamma + \delta\, \tau(x;\, \varepsilon,\, \lambda)\,, \qquad (10)$$

where parameters γ, δ, ε, λ, and the function $\tau(\cdot)$ are chosen so that ξ has a normal distribution $N(0,\, 1)$ with zero average and a unit variance. Johnson suggested the following three families of functions $\tau(\cdot)$:

$$S_L : \tau_L(x;\, \varepsilon,\, \lambda) = \ln\left(\frac{x-\varepsilon}{\lambda}\right), \quad x \geq \varepsilon\,,$$

$$S_B : \tau_B(x;\, \varepsilon,\, \lambda) = \ln\left(\frac{x-\varepsilon}{\varepsilon+\lambda-x}\right), \quad \varepsilon \leq x \leq \varepsilon+\lambda\,, \qquad (11)$$

$$S_U : \tau_U(x;\, \varepsilon,\, \lambda) = \ln\left(\frac{x-\varepsilon}{\lambda} + \sqrt{\left(\frac{x-\varepsilon}{\lambda}\right)^2 + 1}\right),$$

$$-\infty < x < \infty.$$

Knowing the empirical estimates $\hat{\mu}_i$ of the true central moments μ_i ($i = 2,\, 3,\, 4$) of the initial random variable X, one may decide what family of functions $\tau(\cdot)$ is preferable to describe the distribution X, provided the sample $X_1,\, \ldots\,,\, X_n$ of size N is preset.

The technique of choosing the functions $\tau(\cdot)$ is based on estimating the index of asymmetry $\beta_1 = \mu_3^2/\mu_2^3$ and the relative index of excess $\beta_2 = \mu_4/\mu_2^2$. The technique consists in the following. If a curve, given by the parametric equation

$$\begin{cases} \beta_1 = (\omega - 1)\,(\omega + 2)^2\,, \\ \beta_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3\,, \end{cases} \qquad (12)$$

is plotted in the plane $(\beta_1,\, \beta_2)$, one should choose the S_L approximation (the lognormal distribution) for those distributions, for which β_1 and β_2 lie either on this curve or close to it; if β_1 and β_2 lie above the S_L line, then the S_B approximation should be used, and in case of β_1 and β_2 lying below this line one should use S_U Johnson approximation.

To use Eq. (12) in practice, it is convenient to express ω using the first equation as follows:

$$\omega^3 + 3\omega^2 - (4 + \beta_1) = 0\,,$$

and to find the real root

$$\omega_1 = \sqrt[3]{\frac{2+\beta_1}{2} + \sqrt{\left(\frac{2+\beta_1}{2}\right)^2 - 1}} + \sqrt[3]{\frac{2+\beta_1}{2} - \sqrt{\left(\frac{2+\beta_1}{2}\right)^2 - 1}} - 1.$$

By substituting ω_1 into the second Eq. (12) we determine the sign and the value of the discrepancy with β_2

$$\varepsilon_0 = \beta_2 - (\omega_1^4 + 2\omega_1^3 + 3\omega_1 - 3)\,. \qquad (13)$$

If we now chose ε_1 for the admissible discrepancy, then the algorithm for choosing the family of distributions may be described as follows:

if Eq. (13) yields $|\varepsilon_0| \leq \varepsilon_1$, we choose the family of distributions, related to the equation of curve (12), i.e., S_L;

if $|\varepsilon_0| > \varepsilon_1$, and ε_0 is negative, then the S_U family is to be chosen;

if $\varepsilon_0 > 0$, then the S_B family is to be chosen.

After specifying the form of the function $\tau(\cdot)$ in the expression (10) the Johnson distribution describing the density of probability of random variable X have the form

$$f(x) = \frac{|\mathrm{d}\, \tau'_x(x;\, \varepsilon,\, \lambda)|}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(\gamma + \delta\, \tau(x;\, \varepsilon,\, \lambda))^2\right\}, \qquad (14)$$

where

$$\tau'_x(x;\, \varepsilon,\, \lambda) = d\, \tau(x;\, \varepsilon,\, \lambda)/d\, x\,.$$

Assume that the parameters of Johnson distributions are somehow set. At the second step of the approximating procedure $\mathbf{X}$ is a vector, $\mathbf{X} \in E^n$, $\mathbf{X} = (X^1,\, \ldots\,,\, X^n)^{\mathrm{T}}$ each component of which is transformed into the component of vector $\xi \in E^n$ by formula (10). To reconstruct the intercomponent relations between the random values ξ^i and ξ^j of the vector ξ, we use the estimates of coefficients of correlation, from the teaching sample $\mathbf{X}_1,\, \ldots\,,\, \mathbf{X}_N$; here N is the size of the sample, and $\mathbf{X} \in E^n$

$$r_{ij} = \frac{1}{N} \sum_{k=1}^{N} \xi_k^i\, \xi_k^j\,, \qquad (15)$$

where

$$\xi_k^s = \gamma^s + \delta^s\, \tau(X_k^s;\, \varepsilon^s,\, \lambda^s)\,; \quad s = i,\, j\,; \quad i,\, j = 1,\, \ldots,\, n\,.$$

Then, following the assumed normal character of distribution, the joint distribution of the components of vector ξ will be written as follows:

$$f(\xi) = \frac{1}{(2\pi)^{n/2}\, |R|^{1/2}} \exp\left\{-\frac{1}{2}\, \xi^T\, R^{-1}\, \xi\right\}, \qquad (16)$$

where $R = (r_{ij})$ is the matrix of correlation composed of coefficients (15). Hence, the estimate of distribution of the initial vector $\mathbf{X}$ has the form

$$f(\mathbf{x}) = \frac{\prod_{i=1}^{n} \delta^i\, \tau'_x(x^i;\, \mathrm{e}^i;\, 1^i)}{(2\pi)^{n/2}\, |R|^{1/2}} \exp\left\{-\frac{1}{2}(\gamma + \delta\, \tau(\mathbf{x};\, \varepsilon,\, \lambda))^T \times \right.$$

$$\left. \times\, R^{-1}\, (\gamma + \delta\, \tau(\mathbf{x};\, \varepsilon,\, \lambda))\right\} \qquad (17)$$

with the vector

$$(\gamma+\delta\tau(x;\ \varepsilon,\ \lambda))=\begin{pmatrix}\gamma^1+\delta^1\tau(x^1;\varepsilon^1,\lambda^1)\\ \cdot\\ \cdot\\ \cdot\\ \gamma^n+\delta^n\tau(x^n;\varepsilon^n,\lambda^n)\end{pmatrix}.$$

Let us now consider the problem on estimating parameters of the families S_L and S_B of Johnson distributions. First, note that ε and λ have simple meaning: parameter ε gives the value of the lower boundary, and parameter $\varepsilon + \lambda$ – that of the upper boundary of the random variable X. In many cases these parameters may be estimated from the physical meaning of the measured values, and also directly from the teaching sample. If parameters ε and λ are known, parameters γ and δ may be obtained using the technique of maximum likelihood.

Let $y = \{X_1, \ldots, X_N\}$ be the independent sampling values of the random variable X (N is the sample size). With the account of Eq. (14) the function of likelihood have the form

$$L(\gamma,\ \delta\ /\ y) = \left(\frac{\delta}{\sqrt{2\pi}}\right)^N \prod_{j=1}^{N} \tau'_x(X_j;\ \varepsilon,\ \lambda) \times$$

$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\gamma + \delta\ \tau\ (X_i;\ \varepsilon,\ \lambda))^2\right\}. \qquad (18)$$

By differentiating $L(\gamma,\ \delta/y)$ by γ and δ, and equating these derivatives to zero, we obtain a system of equations

$$\gamma + \delta\ \frac{1}{N}\sum_{j=1}^{N}\tau\ (X_j;\ \varepsilon,\ \lambda) = 0\ , \qquad (19)$$

$$\delta^2\ \frac{1}{N}\sum_{j=1}^{N}\tau^2(X_j;\ \varepsilon,\ \lambda) + \gamma\ \delta\ \frac{1}{N}\sum_{j=1}^{N}\tau\ (X_j;\ \varepsilon,\ \lambda) = 1\ .$$

Whence it follows that

$$\delta = \left\{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left[\tau\ (X_j;\ \varepsilon,\ \lambda) - \frac{1}{N}\sum_{i=1}^{N}\tau\ (X_i;\ \varepsilon,\ \lambda)\right]^2}\right\}^{-1}, \qquad (20)$$

$$\gamma = \left\{-\frac{1}{N}\sum_{j=1}^{N}\tau\ (X_j;\ \varepsilon,\ \lambda)\right\}\delta\ .$$

Thus, defining the parameters of the families of Johnson curves from sampling data at the teaching stage, one may use the approximations of the unknown conditional functions of density (17) which are used for the Bayes decision rules of form (1) (see Ref. 6).

Note, in conclusion, that, besides the quality of the multidimensional conditional functions of density (8) and (17), the quality of the obtained decision rule (1), determined by the empirical risk (2), depends on the mutual position of points of the teaching sequences.

REFERENCES

1. V.G. Repin and G.P. Tartakovskii, *Statistical Synthesis in Case of A Priori Indefiniteness and Adaptation of Information Systems* (Sov. Radio, Moscow, 1977), 432 pp.
2. V.S.Pugachev, *Probability Theory and Mathematical Statistics* (Nauka, Moscow, 1979), 496 pp.
3. J.S. Ramberg and B.W. Schmeiser, Communication of the Acm. **15**, No. 11, 987–990 (1972).
4. M. Kendall and A. Stuart, *Theory of Distributions* [Russian translation] (Nauka, Moscow, 1966), 588 pp.
5. G. Khan and S. Shapiro, *Statistical Models in Engineering Problems* (Mir, Moscow, 1968), 395 pp.
6. A.P. Serykh, K.T. Protasov, and F.Ya. Borkun, Izv. Vyssh. Ucheb. Zaved. SSSR, Ser. Neft' i Gaz, No. 1, 3–9 (1972).