INFORMATICA, 1993, Vol.4, No.3–4, 360–383

ON SHAPE OF PATTERN ERROR FUNCTION, INITIALIZATIONS AND INTRINSIC DIMENSIONALITY IN ANN CLASSIFIER DESIGN

Šarūnas RAUDYS

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St.4, Lithuania

Abstract. An analytical equation for a generalization error of minimum empirical error classifier is derived for a case when true classes are spherically Gaussian. It is compared with the generalization error of a mean squared error classifier – a standard Fisher linear discriminant function. In a case of spherically distributed classes the generalization error depends on a distance between the classes and a number of training samples. It depends on an intrinsic dimensionality of a data only via initialization of a weight vector. If initialization is successful the dimensionality does not effect the generalization error. It is concluded advantageous conditions to use artificial neural nets are to classify patterns in a changing environment, when intrinsic dimensionality of the data is low or when the number of training sample vectors is really large.

Key words: feed forward neural nets, training sample size, generalization, intrinsic dimensionality, initialization, insufficient learning.

1. Introduction. The application of multivariate statistical analysis methods to investigate artificial neural-network (ANN) models has provided tools to characterize a variety of aspects of training abilities of multilayer networks (see, for example, Raudys and Jain, 1991; Jain and Raudys, 1992). Here we use these methods to analyze an influence of target values and a shape of an activation function on training of a feedforward ANN pattern classifier.

The goal is to investigate the influence of target values on a pattern error function used to train neural networks and then to compare training speeds of two networks utilizing two limiting

values of targets.

Consider a simple feedforward network in which the output is a following function of inputs $x_1, x_2, \ldots, x_p$

$$o = f(\text{net}), \tag{1}$$

where

$\text{net} = \sum_{i=1}^{p} a_i x_i + a_0$,

and

$f(\text{net})$ is a nonlinear differentiable monotonically increasing activation function.

To find unknown coefficients (weights) $a_0, a_1, a_2, \ldots, a_p$ of the network usually a following pattern error (cost) function

$$mse = \frac{1}{n}\sum_{j=1}^{n}\left(t_j - f\Big(\sum_{i+1}^{p} a_i x_{ij} + a_0\Big)\right)^2 \tag{2}$$

is minimized with respect to unknown parameters $a_0, \ldots, a_p$.

In equation (2) $t_1, t_2, \ldots, t_n$ are target (desired output) values of the network, determined for each training vector

$$\mathbf{X}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}, \qquad j = 1, 2, \ldots, n,$$

where n is a number of training vectors in a training set. In classification problem the target values are usually fixed to be constant for all training vectors of one class. Rumelhart, Hinton and Williams (1986), for example, used the activation function

$$f(\text{net}) = \frac{1}{1 + e^{-\text{net}}}, \tag{3}$$

and targets: $t^{(1)} = 0.1$ (for one class) and $t^{(2)} = 0.9$ (for another one). Another choice commonly used in ANN design is the hyperbolic tangent.

A specification of target values influences the pattern error function. In the literature only vague considerations are presented concerning best values of the targets. In following section we shall review main results concerning the generalization error of a mean squared error classifier which is obtained when the targets for both classes are close. In Section 3 we shall investigate analytically the minimum empirical error ANN classifier which is obtained when targets significantly differs and approaches their limit values, i.e., **0** or **1**. In Section 4 we describe simulation studies and Section 5 contains a discussion.

2. Target values, a shape of error function and the generalization error. In this paper we shall analyze only one type of the activation function, i.e., the sigmoid function (3). Analysis of the other functions is analogous. Let targets $t^{(1)}$, $t^{(2)}$ are very close to 0.5 (0.5 is a value of the activation function $f(\text{net})$ for $\text{net} = 0$): $t^{(1)} = 0.5-\varepsilon$ and $t^{(2)} = 0.5+\varepsilon$ (ε is a small positive constant). Then in a close neighborhood of the point $\text{net} = \mathbf{0}$ a Taylor series expansion with three terms results that the function $f(\text{net})$ is nearly linear:

$$f(\text{net}) = f(0) + f'(0)\,\text{net} = 0.5 + 0.25\,\text{net}.$$

Then the pattern error function (2) can be rewritten in a form

$$\begin{aligned} mse &= \frac{1}{n}\sum_{j=1}^{n}\left(t_j - 0.5 - 0.25\Big(\sum_{i=1}^{p} a_i x_{ij} + a_0\Big)\right)^2 \\ &= \frac{1}{16n}\sum_{j=1}^{n}\Big(4\varepsilon(-1)^{s+1} - \sum_{j=1}^{p} a_i x_{ij} - a_0\Big)^2, \end{aligned} \tag{4}$$

where s in an index of the class for training pattern vector $(x_{1j}, \ldots, x_{pj})'$.

In the case when numbers of training vectors from each class are equal among themselves, i.e., $N_1 = N_2 = N = n/2$, a minimization of the cost function (4) results a standard linear discriminant function (Koford and Groner, 1966)

$$g(x) = \mathbf{A}'\mathbf{X} + w_0 = \sum_{i=1}^{p} a_i x_i + a_0, \tag{5}$$

with

$$\mathbf{A} = \mathbf{S}^{-1}(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}) = (a_1, a_2, \ldots, a_p)',$$
$$a_0 = -\frac{1}{2}(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)})'\mathbf{A},$$

$\overline{\mathbf{X}}^{(1)}$, $\overline{\mathbf{X}}^{(2)}$ are sample estimates of the mean vectors of the classes and $\mathbf{S}$ is a sample estimate of the covariance matrix.

Small sample properties of this linear DF are well analyzed (see, e.g., reviews Raudys and Jain, 1991; Raudys and Pikelis, 1980; Raudys and Jain, 1991a). A variance of the conditional probability of misclassification (PMC) – (the generalization error in ANN terminology) of this classifier asymptotically as the number of training vectors and dimensionality increase tends to zero and its expectation PMC tends to a constant (Deev, 1970; Raudys, 1972).

$$P_N = q_1\Phi\left\{-\frac{\delta^2 + \frac{p}{N_1} - \frac{p}{N_2}}{2\sqrt{(\delta^2 + \frac{2p}{N_1+N_2})\frac{N_1+N_2}{N_1+N_2-p}}}\right\} + q_2\Phi\left\{-\frac{\delta^2 - \frac{p}{N_1} + \frac{p}{N_2}}{2\sqrt{(\delta^2 + \frac{2p}{N_1+N_2})\frac{N_1+N_2}{N_1+N_2-p}}}\right\}, \tag{6}$$

where

$$\Phi(c) = \int_{-\infty}^{c} \varphi(t)dt \quad \text{and} \quad \varphi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

Equation (6) is quite accurate (see, e.g., Pikelis, 1976; Wyman, Young and Turner, 1990) and indicates:

- when the number of training vectors tends to infinity $P_N \to P_\infty = \Phi(-\delta/2)$ – asymptotic PMC;
- when $N_1 + N_2 \to p$ – PMC $P_N \to 0.5$;
- smaller PMC are obtained if $N_2 = N_1$. If $N_2 \neq N_1$ an additional nonzero threshold should be added in order to reduce an effect of unequal number of training vectors (Raudys and Jain, 1991; Deev, 1970).

Suppose now the targets $t^{(1)}$ and $t^{(2)}$ are very close to **0** or **1** (limiting values of the activation function f(net) when net $\to -\infty$ or net $\to +\infty$): $t^{(1)} = \varepsilon$ and $t^{(2)} = 1 - \varepsilon$. Then the minimization of

sum (2) will force to make all scalar products $\mathbf{A}'\mathbf{X}_j^{(1)} + a_o$ as large as possible and all scalar products $\mathbf{A}'\mathbf{X}_j^{(2)} + a_o$ as small as possible. If the classes are well separated and one can obtain a zero empirical classification error then the minimization of the sum (2) for $\varepsilon = 0$ will inlimitedely increase absolute values of the components of the weight vector $\mathbf{A}$. The activation function then will act as a threshold function and the minimization of the sum squared error (2) actually will become the minimization of an empirical probability of misclassification P_{emp}.

Very little can be said about small sample properties of the minimum empirical error classifier. Due to the nonlinear character of the activation function an analytical investigation of the problem becomes difficult. Only upper bounds for the classification error exist (Vapnik, 1979)

$$P_N \leqslant P_{\text{emp}} + \frac{p(\ln \frac{n}{p} + 1) - \ln \eta}{2n} \left(1 + \sqrt{1 + \frac{4nP_{\text{emp}}}{p(\ln \frac{n}{p} + 1) - \ln \eta}}\right), \quad (7)$$

where $1 - \eta$ is a probability of inequality (7).

Bound (7) is obtained for a least favorable distribution of the training vectors and results estimates which hardly can be used in practice. Therefore "an important problem is to extend the learning theory to the regime of simple distributions" (Baum, 1990). For one of such models this will be done in the next section.

3. Generalization error of minimum empirical error classifier.

3.1. A model. We shall analyze a following hypothetical training algorithm.

According to some chosen prior density $f(a, \mathbf{A})$ of a vector $(a, \mathbf{A})$ we generate a set of random weights $a_0, a_1, a_2, \ldots a_p$ and test a condition $\mathcal{S}$

$$\mathcal{S}: \begin{array}{l} \text{for all training pattern vectors from } \pi_1 \; g(\mathbf{X} \mid a, \mathbf{A}) > 0 \\ \text{for all training pattern vectors from } \pi_2 \; g(\mathbf{X} \mid a, \mathbf{A}) \leqslant 0 \end{array} \Bigg\}. \quad (8)$$

If the condition $\mathcal{S}$ is satisfied the training will be called successful.

We shall calculate an expected PMC EP_N of successfully trained linear discriminant function. The expectation will be calculated both over all possible random training sets of the size n and over all possible sets of random weights generated accordingly the prior density $f_{\text{prior}}(a, \mathbf{A})$:

$$EP_N = \text{Prob}(MC \mid \mathcal{S}) = \int\int \text{Prob}(MC \mid a, \mathbf{A}) f_{\text{apost}}(a, \mathbf{A} \mid \mathcal{S}) da \; d\mathbf{A}, \tag{9}$$

where $P(MC \mid a, \mathbf{A})$ is a conditional probability of misclassification given the set of weights $(a, \mathbf{A})$ and $f_{\text{apost}}(a, \mathbf{A}) \mid \mathcal{S} = \text{true})$ is aposteriori density function of the weights if the training was successful, i.e., the conditions $\mathcal{S}$ were satisfied:

$$f_{\text{apost}}(a, \mathbf{A} \mid \mathcal{S}) = \frac{P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) f_{\text{prior}}(a, \mathbf{A})}{P(\mathcal{S} = \text{true})} = \frac{P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) f_{\text{prior}}(a, \mathbf{A})}{\int\int P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) f_{\text{prior}}(a, \mathbf{A}) da \, d\mathbf{A}}. \tag{10}$$

In order to obtain an analytical expression for EP_N suitable for numerical evaluation of the error rate we need to specify the prior density $f_{\text{prior}}(a, \mathbf{A})$ and true probability density functions of the pattern classes $f(\mathbf{X} \mid \pi_1)$, $f(\mathbf{X} \mid \pi_2)$.

To obtain easy to calculate equation for the expected PMC we shall analyze a case of very simple distributions:

– two multivariate spherically Gaussian classes with densities $N(\mathbf{X}, \mathbf{C}_1, \mathbf{I})$, $N(\mathbf{X}, \mathbf{C}_2, \mathbf{I})$, equal prior probabilities $q_1 = q_2 = 1/2$; equal number of training vectors from each class: $N_1 = N_2 = N = n/2$.

Also we shall assume the training vectors $\mathbf{X}_1^1, \mathbf{X}_2^1, \ldots, \mathbf{X}_N^1, \mathbf{X}_1^2, \ldots, \mathbf{X}_N^2$ are statistically independent and identically distributed in their own classes.

We shall analyze a limit case when $N \to \infty$. Then sample discriminant function is close to optimal one and all third order effects $1/n^3$ and higher will be neglected.

3.2. Conditional PMC. For the chosen model of the true distribution of the classes the linear discriminant function (5) will have Gaussian distribution

$$g(x) = \mathbf{A}'\mathbf{X} + a \sim N(\mathbf{A}'\mathbf{C}_i,\ \mathbf{A}'\,\mathbf{A})$$

and the conditional probability of misclassification

$$\begin{aligned}\mathrm{Prob}(MC \mid a, \mathbf{A}) &= \frac{1}{2}\mathrm{Prob}(\mathbf{A}'\mathbf{X} + a < 0 \mid \mathbf{X} \in \pi_1) + \frac{1}{2}\mathrm{Prob}(\mathbf{A}'\mathbf{X} + a \geqslant 0 \mid \mathbf{X} \in \pi_2) \\ &= \frac{1}{2}\Phi\left(-\frac{\mathbf{A}'\mathbf{C}_1 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\right) + \frac{1}{2}\Phi\left(\frac{\mathbf{A}'\mathbf{C}_2 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\right), \qquad (11)\end{aligned}$$

where

$$\Phi(c) = \int_{-\infty}^{c} \varphi(t)dt \quad \text{and} \quad \varphi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

The conditional PMC (11) depends on $(p+1)$-variate vector $(a\mathbf{A}')'$. For spherical case we can show this PMC depends only on two independent scalar variables. Let us perform a transformation

$$\mathbf{V} = \mathbf{TA} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_p \end{bmatrix},$$

$$\mathbf{L} = \mathbf{T}(\mathbf{C}_1 - \mathbf{C}_2) = \begin{bmatrix} \delta \\ l_2 \\ l_3 \\ \vdots \\ l_p \end{bmatrix},$$

where $\mathbf{T}$ is $p \times p$ orthogonal matrix with a first row vector $\mathbf{t}_1 = (\mathbf{C}_1 - \mathbf{C}_2)/(\mathbf{C}_1 - \mathbf{C}_2)'(\mathbf{C}_1 - \mathbf{C}_2)^{1/2}$ and $\delta^2 = (\mathbf{C}_1 - \mathbf{C}_2)'(\mathbf{C}_1 - \mathbf{C}_2)$ is a squared Mahalanobis distance. Then

$$\begin{aligned}\frac{\mathbf{A}'\mathbf{C}_1 + a}{\sqrt{\mathbf{A}'\mathbf{A}}} &= \frac{(\mathbf{TA})'\Big(\mathbf{T}(\mathbf{C}_1 - \mathbf{C}_2) + \mathbf{T}(\mathbf{C}_1 + \mathbf{C}_2)\Big) + 2a}{2\sqrt{(\mathbf{TA})'(\mathbf{TA})}} \\ &= \frac{v_1\delta + w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}} = u\delta/2 + w, \qquad (12A)\end{aligned}$$

where

$$w_0 = (\mathbf{TA})'\Big(\mathbf{T}(\mathbf{C}_1 + \mathbf{C}_2)\Big) + 2a = \mathbf{A}'(\mathbf{C}_1 + \mathbf{C}_2) + 2a,$$

$$u = \frac{v_1}{\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}},$$

$$w = \frac{w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}}.$$

Analogously

$$\frac{\mathbf{A}'\mathbf{C}_2 + a}{\sqrt{\mathbf{A}'\mathbf{A}}} = -u \cdot \delta/2 + w. \tag{12B}$$

Therefore

$$\begin{aligned} \text{Prob}(MC \mid a, \mathbf{A}) &= \text{Prob}(MC \mid u, w) \\ &= 1/2\Phi\{-u\delta/2 - w\} + 1/2\Phi\{-u\delta/2 + w\}. \end{aligned} \tag{13}$$

3.3. A probability of the successive training. For independent identically distributed training pattern vectors the conditional probability

$$\begin{aligned} &P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) \\ &= \prod_{j=1}^{N} \text{Prob}\{\mathbf{A}'\mathbf{X}_j^{(1)} + a > 0\} \prod_{j=1}^{N} \text{Prob}\{\mathbf{A}'\mathbf{X}_j^{(2)} + a \leqslant 0\} \\ &= \Big[\text{Prob}\{\mathbf{A}'\mathbf{X} + a > 0 \mid \mathbf{X} \in \pi_1\}\Big]^N \Big[\text{Prob}\{\mathbf{A}'\mathbf{X} + a \leqslant 0 \mid \mathbf{X} \in \pi_2\}\Big]^N \\ &= \Big[1 - \text{Prob}\{\mathbf{A}'\mathbf{X} + a \leqslant 0 \mid \mathbf{X} \in \pi_1\}\Big]^N \\ &\quad \times \Big[1 - \text{Prob}\{\mathbf{A}'\mathbf{X} + a > 0 \mid \mathbf{X} \in \pi_2\}\Big]^N \\ &= \left[1 - \Phi\Big(-\frac{\mathbf{A}'\mathbf{C}_1 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\Big)\right]^N \left[1 - \Phi\Big(-\frac{\mathbf{A}'\mathbf{C}_2 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\Big)\right]^N. \end{aligned}$$

Taking into account (12A) and (12B) the above equation we can rewrite in a form

$$\begin{aligned} P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) &= P(\mathcal{S} = \text{true} \mid u, w) \\ &= [1 - \Phi(-u\delta/2 - w)]^N [1 - \Phi(-u\delta/2 + w)]^N. \end{aligned} \tag{14}$$

In further analysis we shall use an expansion

$$(1-\Phi)^N = \exp\left(-N\Phi - N\Phi^2/2 - N\Phi^4/3 - \ldots\right),$$

where $\Phi \ll 1$.
Then

$$\begin{aligned} &P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) \\ &\quad = \exp\Big\{-N\big[\Phi(-u\delta/2 - w) + \frac{1}{2}\Phi^2(u\delta/2 - w) + \ldots \\ &\qquad + \Phi(-u\delta/2 + w) + \frac{1}{2}\Phi^2(-u\delta/2 + w) + \ldots\big]\Big\}. \end{aligned} \tag{15}$$

In the limit case for $N \to \infty$ the parameters of the discriminant function will approach to their ideal (true) values. Therefore $u \to 1$. Let us denote

$$\begin{aligned} &v = 1 - u \qquad \text{with} \qquad v > 0, \qquad \text{and} \\ &\Phi(-\delta/2) = \Phi, \\ &\varphi(-\delta/2) = \varphi, \\ &\varphi'(-\delta/2) = -(-\delta/2)\varphi(-\delta/2) = \delta\varphi/2. \end{aligned}$$

Then we can write

$$\begin{aligned} \Phi(-u\delta/2 \pm w) &= \Phi(-\delta/2(1-v) \pm w) \\ &= \Phi(-\delta/2 + \delta/2v \pm w) \\ &= \Phi(-\delta/2) + (\delta/2v \pm w)\varphi(-\delta/2) + \frac{1}{2}(\delta/2v \pm w)^2\varphi'(-\delta/2) + \ldots \\ &= \Phi + (\delta/2v \pm w)\varphi + (\delta/2v \pm w)^2\delta\varphi/4 + \ldots. \end{aligned} \tag{16}$$

Consequently taking into account (15) and (16)

$$\begin{aligned} P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) &= P(\mathcal{S} = \text{true} \mid u, w) \\ &= \exp\Big\{-N\Big(\Phi + \varphi\Big(\frac{\delta}{2}v - w\Big) + \frac{\varphi\delta}{4}\Big(\frac{\delta}{2}v - w\Big)^2 \\ &\qquad + \frac{1}{2}\Big[\Phi + \varphi\Big(\frac{\delta}{2}v - w\Big) + \frac{\varphi\delta}{4}\Big(\frac{\delta v}{2} - w\Big)\Big]^2 \\ &\qquad + \Phi + \varphi\Big(\frac{\delta}{2}v + w\Big) + \frac{\varphi\delta}{4}\Big(\frac{\delta v}{2} + w\Big)^2 \\ &\qquad + \frac{1}{2}\Big[\Phi + \varphi\Big(\frac{\delta}{2}v + w\Big) + \frac{\varphi\delta}{4}\Big(\frac{\delta v}{2} + w\Big)\Big]^2\Big)\Big\}, \end{aligned}$$

and after some simple algebra we have

$$P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) = P(\mathcal{S} = \text{true} \mid u, w)$$
$$= k \cdot \exp\left\{ - N\Big(\delta\varphi(1+\Phi)v + \frac{\delta^2}{4} \cdot \frac{\delta\varphi}{2}\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)v^2 + \frac{\delta\varphi}{2}\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)w^2\Big)\right\}, \tag{17}$$

where here and below in this paper coefficients k denote terms which do not depend either on v nor on w.

3.4. The prior distribution of v, w. We shall use a following prior density

$$f_{\text{prior}}(v, w) = \begin{cases} N(v, m_v, b_v^2)N(w, 0, b_w^2), & \text{when} \quad v \geqslant 0, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

A particular case. Let components of the vector $\mathbf{A}$ be chosen random from Gaussian distribution with zero mean and variance σ^2: $a_i \sim N(0, \sigma^2)$. Then the components of the vector $\mathbf{V}$ also will have Gaussian distribution $v_i \sim N(0, \sigma^2)$. It is easy to show that asymptotically when $p \to \infty$, $u = v_1/\sqrt{\mathbf{V}'\mathbf{V}} \sim N(0, 1/p)$. Consequently $f_{\text{prior}}(v) \quad N(v, 1-\varepsilon, 1/p)$ with $\varepsilon \to 0$ as $p \to \infty$.

3.5. Aposteriori p.d.f. f (a, B | S= true). A joint probability

$$P(\mathcal{S} = \text{true} \mid a, \mathbf{A}) f_{\text{prior}}(a, \mathbf{A}) = P(\mathcal{S} = \text{true} \mid u, w) f_{\text{prior}}(u, w)$$
$$= k \cdot \exp\left\{ - N\Big[\delta\varphi(1+\Phi)v + \frac{\delta^3\varphi}{8}\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)v^2 + \frac{\delta\varphi}{2}\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)w^2\Big]\right\} \cdot \exp\left\{ \frac{m_v}{\delta_v^2}v - \frac{1}{2b_v^2}v^2 - \frac{1}{2b_w^2}w^2 \right\}$$
$$= k \cdot \exp\left\{ - \frac{1}{2}\left(\Big(\frac{1}{b_v^2} + \frac{N\delta^3\varphi}{4}\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)\Big)v^2 + 2\Big[N\delta\varphi(1+\Phi) - \frac{m_v}{b_v^2}\Big]v\right)\right\}$$
$$\times \exp\left\{ - \frac{1}{2}\Big(\frac{1}{b_w^2} + N\delta\varphi\Big(1 + \frac{2\varphi}{\delta} + \Phi\Big)\Big)w^2\right\}$$

$$= k \cdot N\Big(v, -\frac{N\delta\varphi(1+\Phi) - m_v/b_v^2}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4}, \frac{1}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4}\Big)$$
$$\times N\Big(w, 0, \frac{1}{1/b_v^2 + N\delta\varphi(1+2\varphi/\delta+\Phi)}\Big), \qquad \text{when} \qquad v \geqslant 0.$$

Then aposteriori p.d.f.

$$f_{\text{apost}}(v, w \mid S) =$$
$$= \frac{P(S = \text{true} \mid v, w) f_{\text{prior}}(u, w)}{-\infty \int^{\infty} {}_0 \int^{\infty} P(S = \text{true} \mid v, w) f_{\text{prior}}(u, w) dv dw}$$
$$= k \cdot N\Big(v, \frac{m_v/b_v^2 - N\delta\varphi(1+\Phi)}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4},$$
$$\frac{1}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4}\Big)$$
$$\times N\Big(w, 0, \frac{1}{1/b_v^2 + N\delta\varphi(1+2\varphi/\delta+\Phi)}\Big), \quad \text{when} \quad v \geqslant 0. \quad (19)$$

3.6. Expected PMC. A use of Taylor series expansion for (13) results

$$\text{Prob}(MC \mid v, w) = \Phi + v\varphi\delta/2 + v^2\varphi\delta^3/16 + w^2\varphi\delta/4, \qquad (20)$$

Inserting (19), (20) into (9) we obtain

$$\text{Prob}(MC \mid S = \text{true}) = \Phi + \overline{v}\varphi\delta/2 + \overline{v^2}\varphi\delta^3/16 + \overline{w^2}\varphi\delta/4 \qquad (21)$$

where

$$\overline{w}^2 = -\infty \int^{\infty} w^2 N\Big(w, 0, \frac{1}{1/b_w^2 + N\delta\varphi(1+2\varphi/\delta+\Phi)}\Big) dw$$
$$= \frac{1}{1/b_w^2 + N\delta\varphi(1+2\varphi/\delta+\Phi)}$$
$$= \frac{1}{N\delta\varphi(1+2\varphi/\delta+\Phi)} - \frac{1}{N^2\delta^2\varphi^2(1+2\varphi/\delta+\Phi)^2 b_w^2}, \qquad (22)$$
$$\overline{v}^i = \int_0^{\infty} v^i N\Big(v, \frac{m_v/b_v^2 - N\delta\varphi(1+\Phi)}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4},$$
$$\frac{1}{1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4}\Big) dv \qquad i = 1, 2 \quad (23)$$

are moments of a truncated Gaussian distribution.

For a truncated Gaussian $N(x, m_x, \sigma_x^2)$ it is known (G.Korn and T.Korn, 1961)

$$Ex = \mu_x + \frac{\sigma_x \varphi\left(\frac{m_x}{\sigma_x}\right)}{\Phi\left(\frac{m_x}{\sigma_x}\right)}$$

and

$$Ex^2 = \mu_x^2 + \frac{\sigma_x \varphi\left(\frac{m_x}{\sigma_x}\right)}{\Phi\left(\frac{m_x}{\sigma_x}\right)}\mu_x + \sigma_x^2.$$

For the density determined in (19)

$$\frac{m_v}{\sigma_v} = [m_v/b_v^2 - N\delta\varphi(1+\Phi)] \times \left[1/b_v^2 + N\delta^3\varphi(1+2\varphi/\delta+\Phi)/4\right]^{-1/2} \to -\infty \quad \text{as} \quad N \to \infty.$$

For large negative $c = m_v/\sigma_v$ we shall use an expansion (Zoritch, 1984)

$$\Phi(c) = \frac{\varphi(c)}{-c}\left(1 - \frac{1}{c^2} + \frac{3}{c^4} - \frac{15}{c^6} + \cdots\right).$$

Thus

$$\left.\begin{aligned} Ex &= \frac{\sigma_x^2}{\mu_x}\left(1 - 2\frac{\sigma_x^2}{\mu_x}\cdot\frac{1}{\mu}\right) \\ Ex^2 &= 2\left(\frac{\sigma_x^2}{\mu_x}\right)^2 \end{aligned}\right\}. \tag{24}$$

Use of (24) in (23) results

$$\begin{aligned} \overline{v} = Ev &= \frac{1}{N(\varphi\delta(1+\Phi) - m_v/(Nb_v^2))} - \frac{\delta^3\varphi(1+2\varphi/\delta+\Phi)}{N^2 2\left[\varphi\delta(1+\Phi) - m_v/(Nb_v^2)\right]^3} \\ &= \frac{1}{N\varphi\delta(1+\Phi)} + \frac{m_v}{N^2\varphi^2\delta^2(1+\Phi)^2 b^2 v} - \frac{1+2\varphi/\delta+\Phi}{2N^2\varphi^2(1+\Phi)^3}, \\ \overline{v}^2 = Ev^2 &= 2\frac{1}{N^2\left[\varphi\delta(1+\Phi) - m_v/(Nb_v^2)\right]^2} = \frac{2}{N^2\varphi^2\delta^2(1+\Phi)^2}. \end{aligned} \tag{25}$$

Inserting (25), (22) into (21) we obtain a final result – the expected probability of misclassification EP_N

$$\text{Prob}(MC \mid \mathcal{S} = \text{true}) = \Phi + \frac{\varphi\delta}{2}\left(\frac{1}{N\varphi\delta(1+\Phi)} + \frac{m_v}{N^2\varphi^2\delta^2(1+\Phi)^2 b_v^2}\right.$$

$$
\begin{aligned}
&- \frac{1+2\varphi/\delta+\Phi}{N^2 2\varphi^2(1+\Phi)^3}\Big) + \frac{\varphi\delta^3}{16N^2}\cdot\frac{2}{\varphi^2\delta^2(1+\Phi)^2} \\
&+ \frac{\varphi\delta}{4}\Big(\frac{1}{N\delta\varphi(1+2\varphi/\delta+\Phi)} - \frac{1}{N^2\delta^2\varphi^2(1+2\varphi/\delta+\Phi)^2 b_w^2}\Big) \\
= \Phi &+ \frac{1}{N}\Big(\frac{1}{2(1+\Phi)} + \frac{1}{4(1+2\varphi/\delta+\Phi)}\Big) \\
&+ \frac{1}{N^2}\frac{\delta}{\varphi}\Big(\frac{1}{8(1+\Phi)^2} - \frac{1+2\varphi/\delta+\Phi}{4(1+\Phi)^3}\Big) \\
&+ \frac{1}{N^2}\frac{1}{2\delta\varphi(1+\Phi)^2}\cdot\frac{m_v}{b_v^2} - \frac{1}{N^2}\frac{1}{4\delta\varphi(1+2\varphi/\delta+\Phi)^2}\cdot\frac{1}{b_w^2}. \quad (26)
\end{aligned}
$$

When classes are well separated $\Phi \ll 1$, $\varphi \ll 1$ and we can assume $1+\Phi \approx 1$ and $1+2\varphi/\delta+\Phi \approx 1$. Then we obtain a following asymptotic equation

$$
EP_N = P_\infty + \frac{3}{4N} - \frac{1}{8N^2}\frac{\delta}{\varphi} + \frac{1}{2N^2\delta\varphi}\frac{m_v}{b_v^2} - \frac{1}{4N^2\delta\varphi}\frac{1}{b_w^2}, \quad (27)
$$

where we inserted $P_\infty = \Phi$ as the asymptotic PMC.

We see in Eq. (27) a constant positive contribution term $\frac{3}{4N}$, which means the minimum empirical error classifier depends only on the number of training samples when N is very large. In Eq. (27) we do not see an explicit influence of dimensionality. It influences the expected PMC only via a determination the prior distribution of $a, \mathbf{A}$ (or v, w). E.g., insertion the density $f_{\text{prior}}(v) = N(v, 1-\varepsilon, 1/p)$ with $\varepsilon \to 0$ discussed in subsection 3.4 results the following contribution term in Eq. (27)

$$
\frac{1}{N^2}\frac{1}{2\delta\varphi}\frac{m_v}{b_v^2} = \frac{1}{2N^2\delta\varphi}\frac{1}{1/p} = \frac{p}{N^2}\cdot\frac{1}{2\delta\varphi}. \quad (28)
$$

The very last term in Eq. (27) explains an influence of a correct apriori selection of the threshold w. The optimal threshold for two spherical distribution discussed in this paper is $w = 0$. If we would set $E_{\text{prior}} w = 0$ as was done in (18) and have small apriori variance b_w^2, then the large last negative term in Eq. (27) will indicate that an use of correct prior information concerning the value of the threshold w can reduce the expected PMC. If, however, we do not have information concerning the value of the threshold w and b_w^2 is

large then with increase in the number of training samples N the influence of prior setting of w reduces and disappears at last.

3.7. Intrinsic dimensionality and generalization error. Small sample properties of parametric local statistical pattern classifiers mainly depend on an intrinsic dimensionality of the data in local areas of the multivariate feature space (Raudys, 1991). The ANN minimum empirical error classifier is in fact also nonparametric local classifier designed without any assumptions on a general structure of the data. Therefore one may hope small sample properties of this classifier could be determined by the intrinsic dimensionality of the data (Duin, 1993). This statement can be easily confirmed for the spherical r-dimensional data lying in r-dimensional subspace of p-variate feature space.

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{c}_{i1} \\ \mathbf{c}_{i2}, \end{pmatrix} \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0}, & \mathbf{I}\cdot\lambda^2 \end{pmatrix} \right), \quad \text{if} \quad \mathbf{X} \in \pi_i, \tag{29}$$

and we'll use a discriminant function

$$g(x) = \mathbf{A}_1'\mathbf{X} + \mathbf{A}_2'\mathbf{X}_2 + a,$$

where $\mathbf{c}_{i1}$, $\mathbf{A}_1$ and $\mathbf{X}_1$ are r-variate vectors, $\mathbf{I}_r$ is $r \times r$ identity matrix and $\lambda^2 \ll 1$.

The conditional classification error $\text{Prob}(MC|a, \mathbf{A})$ and the conditional probability $P(S = \text{true}|a, \mathbf{A})$ for this model will be

$$\text{Prob}(MC|a, \mathbf{A}) = \frac{1}{2}\Phi\left(-\frac{\mathbf{A}_1'\mathbf{c}_{11} + \mathbf{A}_2'\mathbf{c}_{12} + a}{\sqrt{\mathbf{A}_1'\mathbf{A}_1 + \mathbf{A}_2'\mathbf{A}_2\lambda^2}} \right) + \frac{1}{2}\Phi\left(\frac{\mathbf{A}_1'\mathbf{c}_{21} + \mathbf{A}_2'\mathbf{c}_{22} + a}{\sqrt{\mathbf{A}_1'\mathbf{A}_1 + \mathbf{A}_2'\mathbf{A}_2\lambda^2}} \right),$$

$$P(S = \text{true}|a, \mathbf{A}) = \left[1 - \Phi\left(-\frac{\mathbf{A}_1'\mathbf{c}_{11} + \mathbf{A}_2'\mathbf{c}_{12} + a}{\sqrt{\mathbf{A}_1'\mathbf{A}_1 + \mathbf{A}_2'\mathbf{A}_2\lambda^2}} \right) \right]^N \times \left[1 - \Phi\left(\frac{\mathbf{A}_1'\mathbf{c}_{21} + \mathbf{A}_2'\mathbf{c}_{22} + a}{\sqrt{\mathbf{A}_1'\mathbf{A}_1 + \mathbf{A}_2'\mathbf{A}_2\lambda^2}} \right) \right]^N.$$

When $\lambda^2 \to 0$, then all the data will lie in a linear r-dimensional subspace, all terms $\mathbf{A}_2', \mathbf{C}_{12}$, $\mathbf{A}_2'\mathbf{C}_{22}$, $\mathbf{A}_2'\mathbf{A}_2\lambda^2$ will tend to zero and the analysis of the small sample behavior of the minimum error classifier can be carried out only in r-dimensional space. Then the generalization error will be determined by Eq. (27) with a new "reduced" contribution term

$$\frac{r}{N^2}\,\frac{1}{2\delta\varphi}. \tag{30}$$

4. Simulation studies. Eq. (6) and (27) are derived for the certain statistical classifiers, which are similar to but at the same time slightly different from adaptive linear classifiers obtained by minimizing squared error function (2). Thus the statistical classifier (6) requires a matrix inversion, and the zero empirical error classifier discussed in Section 3 requires a multiple generation of random weighs and subsequent selection of the proper classifier.

Therefore in order to make sure that above theoretical results are valid for an analysis of small sample properties of the adaptive ANN classifiers some simulation studies were performed.

Two p-variate Gaussian populations were generated by means of pseudo random numbers generator. The populations differ in mean of the first variable $\delta = 3.76$. The one layer linear ANN classifier was trained by means of a standard Back propagation training algorithm until zero empirical error was obtained or until a number of training sweeps exceeded m. Learning speed constant $\eta = 0.1$, momentum term $\alpha = 0.3$. A maximal number of training sweeps $m = 100$. A number of different training sets ($N_1 = N_2 = 100$) used to obtain estimates of the generalization error for different conditions $t = 50$. Number of vectors used to estimate generalization error P_{gen}, $N_t = 500 + 500$. Two series of simulation studies with spherical Gaussian data were performed. In first of them weights of the ANN were initialized randomly in an interval (-1, +1). In second series of experiments in subsequent 49 training experiments with new training data sets the old weights of the trained ANN classifier were retained and used for initialization. Since the training data was new and different from previous one after a new initialization

the empirical error usually increased until 0.05÷0.15 and then while the training progressed it diminished again. This type of experiments corresponds to the case of a good initialization with very small ε and b_v^2 in (18).

Each type of the experiments was performed with different number of features: $p = 2, 4, 6, 8, 10, 20, 40, 60$ and three different target values: $t = 0.0001$, 0.1 and 0.495. The results are presented in Fig. 1 and Fig. 2. Simulation results with the spherical data confirm theoretical conclusion obtained from Eq. (27):

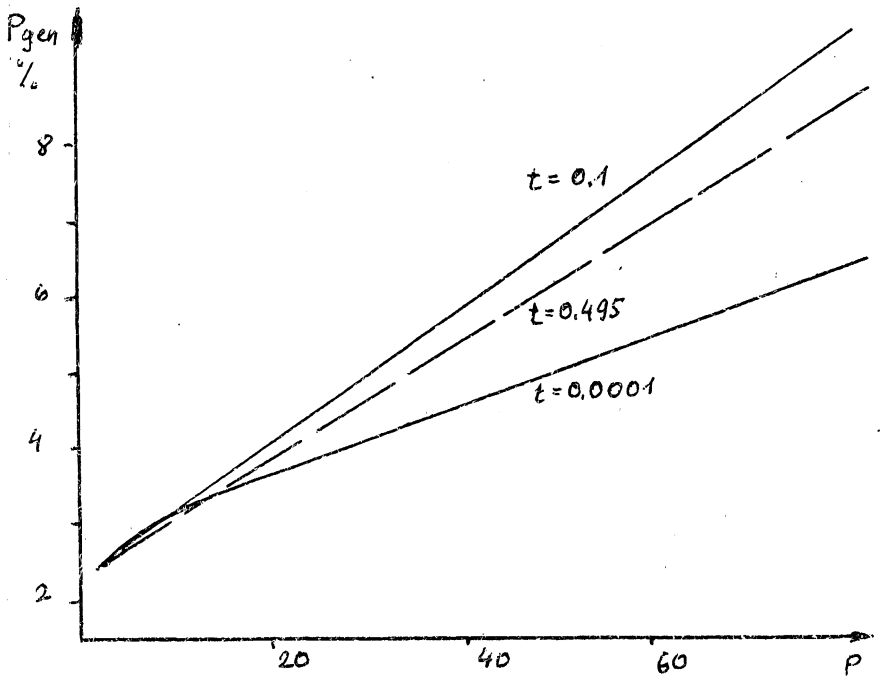

Fig. 1. Generalization error EP_N versus dimensionality p (random initialization).

1) for random initialization generalization error of minimum error classifier increases linearly with the increase in dimensionality;

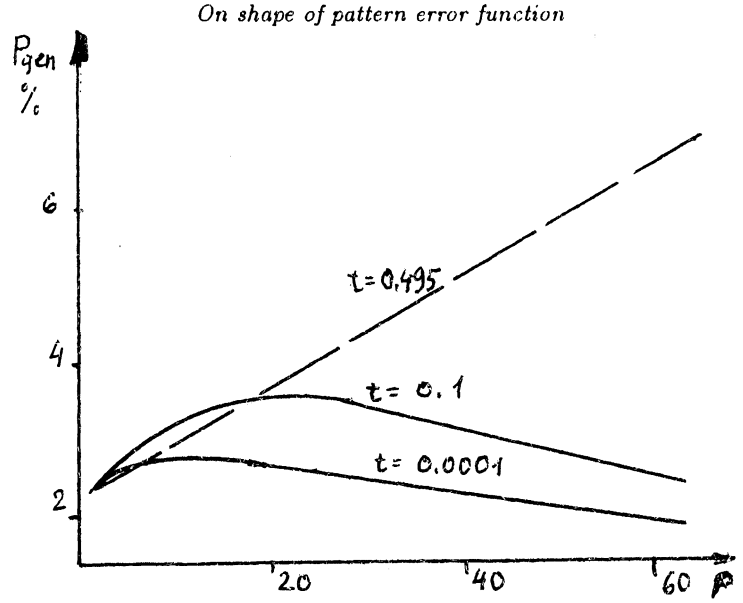

Fig. 2. Generalization error EP_N versus dimensionality p (non-random initialization).

2) a good initialization of minimum error ANN classifier (old weights in our simulation experiment and small target values) depreciate the effect of dimensionality;

3) small sample properties of the ANN classifier essentially depend on the type of pattern error function used to train classifier (in our case on the value of the target).

We pay readers attention that "initialization term" $\frac{p}{N^2}\frac{1}{2\delta\varphi}$ in Eq. (27) for essentially large N can be very small.

The last two conclusions offer an important perspective to use the ANN classifier to classify patterns in changing environment. In slightly changed conditions the weights of the ANN classifier could be accurate enough to serve as the good initialization. However it is not obvious does this conclusion is valid for multilayer ANN.

To check this guess we performed an experiment with non-spherical multivariate Gaussian data and the multilayer ANN classifier with 4 neurons in the hidden layer. Two 8-variate Gaussian

populations were generated with standard deviations:

1 2 1 2 1 2 1 2 (first population),
2 1 1 2 2 1 1 2 (second population).

In the first population means of all features were zero, in second one – 1.5. Asymptotical PMC of 4 hidden neurons ANN classifier was less than 2% . The number of training vectors $N_1 = N_2 = 20$, the number of test set vectors 500+500. Ten different randomly selected training sets were used to obtain average estimates of generalization error. The standard Back propagation training algorithm ($\eta = 0.1$, $\alpha = 0.1$, $m = 800$) was used to train the classifier. For random initialization of the weights of the ANN classifier in the interval (-0.1, +0.1) in all 10 experiments during the training the empirical classification error diminished from 50% until zero. The average generalization error $\overline{P_{\text{gen}}} = 16.1\%$ of misclassifications.

In following simulation studies with nonrandom initializations we selected 4 different weights vectors. Two of them corresponded to a successful initializations (with the generalization errors 10.4% and 11.7%) and two of them – to fairly successfull initializations (with the generalization errors – 29.1% and 19.6%). In each of 4 subsequent series of the experiments the ANN classifiers were trained 10 times with 10 different training sets. The results are presented in Table 1.

Table 1. Dependence of the generalization error on the type of weights initialization

Initialization	Average $\overline{P_{\text{emp}}}$		Average $\overline{P_{\text{gen}}}$
	begin of training	end of training	
Random	0.50	0.0	0.161
nonrandom with $P_{gen} = 0.291$	0.230	0.0	0.209
nonrandom with $P_{gen} = 0.196$	0.128	0.01	0.174
nonrandom with $P_{gen} = 0.117$	0.088	0.002	0.145
nonrandom with $P_{gen} = 0.104$	0.088	0.005	0.123

Results obtained for the multilayer ANN classifier do not contradict to previous conclusion: the results of the training highly depend on the initialization – for successive initializations we obtain smaller generalization error. In next experiments the multilayer ANN classifiers were initialized by weights obtained in previous experiment with the previous training set. The generalization error averaged over 40 training experiments was 14.0%.

Thus one may hope that the multilayer ANN classifier could be sucessfully used to classify patterns in slowly changing environment, when the set of previous weights is good enough for subsequent initialization but bad enough for classification of changed patterns. Term (28) is proportional to $\frac{p}{N^2}$. Thus Eq. (27) indicates conditions when the good initialization could be useful: for larger N the "initialization term"

$$\frac{p}{N^2} \cdot \frac{1}{2\delta\varphi} \tag{28}$$

can become too small to feel the influence of initialization. Really in our experiments with $N_1 = N_2 = 40$ we did not feel the positive effect of sucessive initialization.

In order to study the effect of *intrinsic dimensionality* several ANN training experiments were performed with "singular" data determined by the model (29) with $r = 2$. In simulation experiments addition of $p - 2$ new "singular" features with the standard deviation $\lambda = 0.001$ (or even 0.1) practically did not effect the empirical and the generalization errors if the targets were small ($t = 0.00001$) – the generalization errors were the same for 2 and for 60 features (Graph 1 in Fig. 3). However for $t = 0.495$ and $r = 2$ it was difficult to train the network. Too small variances ($\sigma^2 = 10^{-6}$) of $p - r$ features practically prevented the training of $p - r$ weights of the network. Then the training process depends on initializations essentially. Graph 2A for the generalization error and graph 2B for the empirical error were obtained for a case when weights were initialized from interval $(-1, +1)$.

For comparision in Fig. 3. we present graph 3 – the generalization error for target $t = 0.495$ and a "full" dimensionality (when $r = p$). Theoretically the graphs 2A and 3 must coincide.

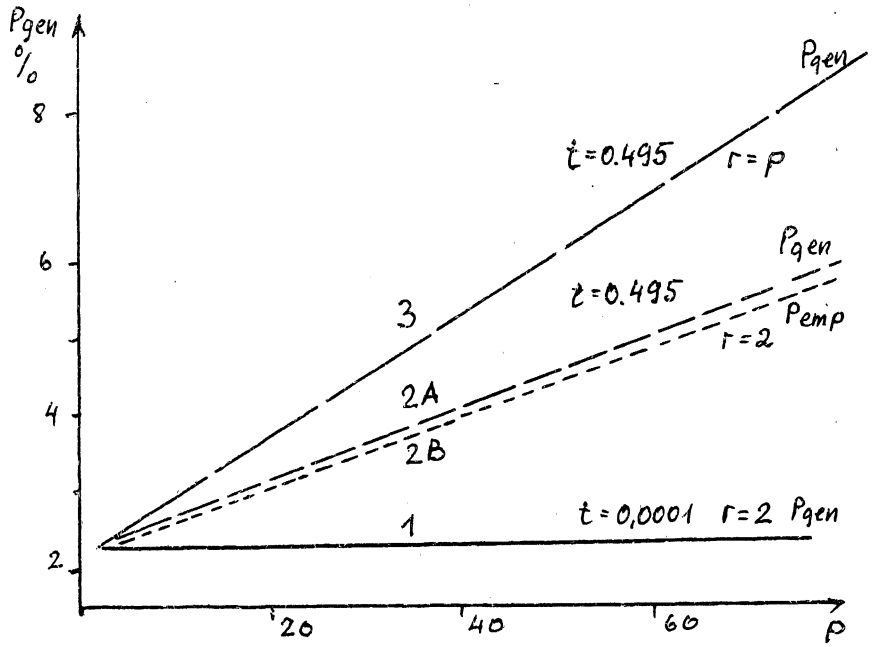

Fig. 3. Generalization and empirical errors versus dimensionality p (random initialization, intrinsic dimensionality $r = 2$ or $r = p$ (for curves 3 only)).

An "early stopping" of the training algorithm limits the classifiers capacity (Kraaijveld, 1993; Schmidt, 1993): the empirical error increases. We present the graphs 2A and 2B of 'unsucessfull' training in order to illustrate that in small training sample size the limitation of the classifiers capacity caused by insufficient training can reduce the generalization error.

The simulation results with singular low intinsic dimensionality data confirm theoretical conclusions and indicates that low intrinsic dimensionality has double effect on sample size considerations of the ANN classifier. For small target values the minimum empirical error classifier is a local classification rule and its generalization abilities are determined by the intrinsic dimensionality. From another hand low intrinsic dimensionality in case of nonideal initialization can cause an early stopping of the training algorithm, reduce the actual capacity of the ANN and as a consequence improve small sample properties. More theoretical and simulation studies are needful.

5. Discussion. Two types of pattern error functions frequentely used in ANN pattern classifier design were discussed. Both of them can be derived from the standard mean squared error function (2) by using different values of the targets.

If the target values of both classes are close to each other we work in a linear part of the activation function $f(\text{net})$ and the classifier obtained coincides with the standard Fisher linear discriminant function. If the targets are different and close to limiting values of the activation function $f(\text{net})$ then essentially we minimize the empirical classification error. In first case all training set pattern vectors have their contribution to patern error function. Morever an assumption that pattern vectors are Gaussian with equal covariance matrices for both classes are used in an implisit way. If this additional information is correct (the classes are Gaussian and the covariance matrices are equal) then it can improve small sample properties of the classification rule obtained. In highdimensional case however the assumption mentioned can cause problems. One of them is that estimation of the covariance matrix requires a large number of training vectors. Another one is that in highdimensional case even if the number of training vectors is large some eigenvalues of the sample covariance matrix are extremely small and numerical difficulties with matrix inversion and/or a stability of the ANN training algorithm arise.

In spite of the fact that the minimum empirical error classifier

utilizes less information concerning a shape of the distribution densities of the pattern classes under a certain conditions this classifier can have good small sample properties. The analytical investigation (Eq. (27)) as well as simulation experiments with the linear and nonlinear multilayer ANN classifier indicate that such conditions arise when the weights of the ANN are correctly initialized. Then the training procedure only shifts a discriminating boundary to its right direction and the increase in the generalization error due the finite number of training samples can be small. The initialization term

$$\frac{p}{N^2} \cdot \frac{1}{2\delta\varphi}$$

in Eq. (27) indicates also that if the number of training samples is sufficiantely large the influence of dimensionality can be also small.

The application of the adaptive ANN classifier may be advantagues when we can get a "good" initialization" or when the data lie in a subspace of low dimensionality. The conditions for the good initialization arise when one applies the ANN to classify patterns in changing environment when the old ANN weights can be used as the starting weight vectors in new changed environment. The conditions for low intrinsic dimensionalities arise when features are highly correlated.

Other advantegeous conditions to use the ANN classifier are when the problem is complex and we have to use the complex multilayer ANN with the great number of inputs and the number of training vectors is large.

Above analysis indicates new partialy unexpected and counterintuitive properties of the ANN classifiers. In future research it would be interesting to obtain analytical formulae for more general distribution of the pattern vectors and nonzero empirical error. In simulation studies it is worth to investigate cases when the intrinsic dimensionality of the data is low and when the ANN is really complex and the number of training vectors is large.

REFERENCES

Baum, E.B. (1990). When k-nearest neighbor and back propagation accurate for feasible sized sets of examples. In L.B. Almeida and C.J. Wellekens (Eds.), *Proceedings of the EURASP Workshop on Neural Networks,* Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, Delft.

Deev, A.D. (1970). Representation of statistics of discriminant analysis and asymptotic expansions in dimensionalities comparable with sample size. *Reports of Academy of Sciences of the USSR,* **195**(4), 756–762 (in Russian).

Duin, R.P.W. (1993). Superlearning capabilities of neural networks. In *Proc. of the 8th Scandinavian Conference on Image Analysis,* NOVIM, Norwegian Society for Image Processing and Pattern Recognition, Tromso, Norway. pp. 547–554.

Jain, A., and Š. Raudys (1992). On training sample size and complexity of artificial neural net classifier. *Informatica,* **3**(2), 301–337.

Koford, J.S., and G.F. Groner (1966). The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE Trans. Inf. Theory,* **IT-12**, 42–50.

Korn, G., and T. Korn (1961). *Mathematical Handbook for Scientists and Engineers.* MGraw-Hill, NY.

Kraaijveld, M.A. (1993). Small sample behavior of multi-lager feedforward network classifiers: theoretical and practical aspects. *Ph. D. Thesis,* Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, Delft.

Pikelis, V. (1976). Comparison of methods of computing the expected classification errors. *Automation and Remote Control,* **5**, 59–63 (in Russian).

Raudys, Š., and A.K. Jain (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **13**, 252–264.

Raudys, Š., and A. Jain (1991a). Small sample problems in designing artificial neural networks. In I.K. Sethi and A.K. Jain (Eds.), Elsevier Sci. Publ. NY, pp. 33–50.

Raudys, Š., and V. Pikelis (1980). On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition. *IEEE Trans. on PAMI,* **PAMI-2**(3), 242–252.

Raudys, Š. (1991). On the effectiveness of Parzen window classifier. *Informatica,* **2**(3), 434–454.

Raudys, Š. (1992). On amount at a prior information in designing the classification algorithm. *Proc. Acad. of Sciences of the USSR. Tech. Cyber.,* **4**, 168–174.

Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. KcClelland (Eds.), *Parallel Distributed Processing,* MA:MIT Pres, Cambridge.

Schmidt, W.F. (1993). Neural pattern classifying systems, theory and experiments with trainable pattern classifiers. *Ph. D. Thesis,* Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, Delft.

Vapnik, V.N. (1979). *Estimation of Dependences Based on Empirical Data.* Springer, NY.

Wyman, F., D. Young and D. Turner (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition,* **23**, 775–783.

Zoritch, V. (1984). *Mathematical Analysis.* Part 2. Nauka, Moscow.

Received November 1993

Š. Raudys received the M.S. degree in electrical and computer engineering from Kaunas University of Technology in 1963, and the Candidate of Sciences and Doctor of Sciences degree from the Institute of Mathematics and Cybernetics, Academy of Sciences, Lithuania, in 1969 and 1978, respectively.

He is currently Head of the Department of Data Analysis in the Institute of Mathematics and Cybernetics and Professor in the Vytautas Magnus University, Kaunas. His current research interests include statistical pattern recognition, artificial neural nets, expert systems, machine learning, and data analysis methods.

He is an Associate Editor of International Journals: Pattern Recognition, Pattern Recognition and Image Analysis, Informatica. He has been a member of the Program Committee of many international conferences.