Adaptive Probabilistic Neural Networks for Pattern Classification in Time-Varying Environment

Leszek Rutkowski, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new class of probabilistic neural networks (PNNs) working in nonstationary environment. The novelty is summarized as follows: 1) We formulate the problem of pattern classification in nonstationary environment as the prediction problem and design a probabilistic neural network to classify patterns having time-varying probability distributions. We note that the problem of pattern classification in the nonstationary case is closely connected with the problem of prediction because on the basis of a learning sequence of the length n, a pattern in the moment $n + k$, $k \geq 1$ should be classified. 2) We present, for the first time in literature, definitions of optimality of PNNs in time-varying environment. Moreover, we prove that our PNNs asymptotically approach the Bayes-optimal (time-varying) decision surface. 3) We investigate the speed of convergence of constructed PNNs. 4) We design in detail PNNs based on Parzen kernels and multivariate Hermite series.**

***Index Terms*—Orthogonal series kernel, Parzen kernel, pattern classification, probabilistic neural networks (PNNs), time-varying environment.**

I. Introduction

PROBABILISTIC neural networks (PNNs), introduced by Specht [33]–[35], have their predecessors in the theory of statistical pattern classification. In the 1950s and 1960s, problems of statistical pattern classification in the stationary case were accomplished by means of parametric methods using the available apparatus of statistical mathematics [9], [45]. The knowledge of the probability density to an accuracy of unknown parameters was assumed and the parameters were estimated based on the learning sequence. Typical techniques included maximum likelihood and Bayesian approaches. Having observed tendencies present in literature within the last twenty years we should say that these methods have been almost completely replaced by the nonparametric approach [8], [10]. [11], [15], [21], [32], [38], [46]. In the nonparametric approach, it is assumed that a functional form of probability densities is unknown. The latter are estimated by nonparametric techniques like Parzen's approach, orthogonal series, or nearest neighbor algorithms. It is well known that these techniques are convergent in the probabilistic sense, e.g., in probability or with probability one. Moreover, pattern classification procedures derived from nonparametric estimates are convergent when the length of the learning sequence increases to Bayes' rules. Asymptotically-optimal pattern classification rules were examined by several authors [7], [12], [13], [26], [27], [29], [47]. The PNNs studied in literature implement in a parallel fashion nonparametric estimation techniques. They are characterized by fast training and convergence to the Bayes-optimal decision surface. For interesting applications of the PNNs, the reader is referred to [18], [20], [22], [25], and [37]. The crucial problem in these applications is the choice of the smoothing parameter. Most techniques are based on vector quantization [4], [48], cluster analysis [36], or the genetic algorithm [19]. A short survey of other available methods is given in [16].

The above review of literature concerned the stationary situation. However, a lot of phenomena have a nonstationary character, e.g., the conventer-oxygen process of steelmaking, the change of catalyst properties in an oil refinery or, in the process of carbon dioxide conversion. In particular, a large group of problems from the domain of technology, biology, and physics is described by nonstationary probability densities, e.g., of the type "movable argument,"$f_n(x) = f(x - c_n)$ [23]. The problem of pattern classification in a time-varying environment was analyzed in only a few works, de Figueiredo [6] and Tzypkin [40] used the parametric approach approximating probability densities (or discriminant functions) that change with time by means of linear combinations of a certain fixed set (base) of functions. Appropriate coefficients were estimated by making use of dynamic-stochastic approximation algorithms. Of course, such an approach does not ensure an asymptotic optimality of the classification rules. Rutkowski [28] considered the asymptotically-optimal classification rules in a quasi-stationary environment when class conditional densities are convergent to a finite limit. Some other results concerning learning in a time-varying environment are scattered in literature [14], [17], [24].

In this paper, we propose a new class of probabilistic neural networks working in a nonstationary environment. The novelty is summarized as follows.

1) We formulate the problem of pattern classification in a nonstationary environment as a prediction problem and design a probabilistic neural network to classify patterns having time-varying probability distributions. We note that the problem of pattern classification in the nonstationary case is closely connected with the problem of prediction because on the basis of a learning sequence of length n, a pattern in the moment $n + k$, $k \geq 1$, should be classified.

Manuscript received March 15, 2003; revised October 30, 2003.

The author is with the Department of Computer Engineering, Technical University in Czestochowa and the Department of Artificial Intelligence, WSHE University, Lodz, Poland (e-mail: lrutko@kik.pcz.czest.pl).

Digital Object Identifier 10.1109/TNN.2004.828757

2) We present, for the first time in literature, definitions of optimality of PNNs in time-varying environment. Moreover, we prove that our PNNs asymptotically approach the Bayes-optimal (time-varying) decision surface. Time-varying discriminant functions are estimated by means of a general learning procedure presented in Section III and convergence of algorithms is a consequence of theorems presented in that section.
3) We investigate the speed of convergence of the constructed PNNs.
4) We design, in detail, PNNs based on the Parzen kernels and multivariate Hermite series.

It should be emphasized that the design of the PNNs in a time-varying environment is much more difficult than in the stationary case. In order to design PNNs approaching the Bayes-optimal decision surfaces (time-varying), we should pick up not only a smoothing parameter (denoted in this paper by h_n for the Parzen kernel and by $q(n)$ for the orthogonal series kernel) but also a learning sequence a_n which should satisfy conditions typical for stochastic approximation procedures [1]. In this context, it is worth to quote Bendat and Piersol [2], who believed that the problem of estimation of nonstationary probability density requires the possession of many realizations of the stochastic process. We would like to emphasize that the PNNs constructed in this paper allow to track time-varying discriminant functions (in particular, tracking time-varying probability densities) with use of only one realization of the stochastic process—subsequent observations of a learning sequence. For illustration of the capability of our PNNs, we mention that having a sequence $\{X_n\}$ of independent random variables with probability densities $f_n(x) = f(x - n^t)$, we are able to estimate time-varying densities despite that both f is unknown and parameter t $(0 < t < 1)$ are unknown. Consequently, we are able to estimate time-varying discriminant functions and corresponding classification rules. The PNNs studied in this paper are adaptive in the sense that they adopt to changes of the time-varying environment.

This paper is organized into XI sections. In Section II, we present kernel functions on which the construction of PNNs will be based. Section III is a short introduction to PNNs in stationary environment. Moreover, in this section, we extend the idea of the classical PNNs to the recursive PNNs with a gain $1/n$. In the following sections, we replace the gain $1/n$ by a more general a_n (like in stochastic approximation methods) in order to enhance the recursive PNNs for tracking nonstationary signals. Since the existing theories do not allow to analyze "enhanced" recursive PNNs in a time-varying environment, in Section IV we present appropriate theorems which are very useful in the next sections. In Section V, we describe the problem of pattern classification in a time-varying environment. Estimates of time-varying discriminant functions are presented and classification rules are proposed. In Section VI, it is shown that our PNNs approach Bayes (time-varying) decision discriminant functions. Moreover, we investigate the speed of convergence. In Section VII and VIII, the PNNs based on the Parzen kernel and orthogonal series kernel are discussed in details. A specific case of the above mentioned nonstationarity of the type "movable argument" is elaborated in Section IX. In Section X, we present simulation results.

II. Kernel Functions for the PNNs Construction

All PNNs studied in this paper are based on a sequence $\{K_n\}$, $n = 1, 2, \ldots$, of bivariate Borel, measurable functions (so-called general kernel functions) defined on $A \times A$, $A \subset R^p$, $p \geq 1$. The concept of general kernel functions stems from the theory of nonparametric density estimation. We will use ideas of the two methods: Parzen's approach and orthogonal series.

A. Application of the Parzen Kernel

Sequence K_n based on the Parzen kernel in the multidimensional version takes the following form:

$$K_n(x,u) = h_n^{-p} K\left(\frac{x-u}{h_n}\right) \tag{1}$$

where h_n is a certain sequence of numbers and K is an appropriately selected function. Precise assumptions concerning sequence h_n and function K that ensure the convergence of PNNs will be given in the next sections. It is convenient to assume that function K can be presented in the form

$$K(x) = \prod_{i=1}^{p} H\left(x^{(i)}\right).$$

Then, sequence K_n is expressed by means of formula

$$K_n(x,u) = h_n^{-p} \prod_{i=1}^{p} H\left(\frac{x^{(i)} - u^{(i)}}{h_n}\right). \tag{2}$$

The most popular is the Gaussian kernel given by

$$H(v) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}} v^2 \tag{3}$$

and

$$K_n(x,u) = h_n^{-p} (2\pi)^{-\frac{1}{2}} \prod_{i=1}^{p} \exp\left(\frac{x^{(i)} - u^{(i)}}{h_n}\right)^2. \tag{4}$$

B. Application of Orthogonal Series

Let $g_j(.)$, $j = 0, 1, 2, \ldots$, be a complete orthonormal system in $L_2(\Delta)$, $\Delta \in R$, such that

$$\max_x |g_j(x)| \leq G_j \tag{5}$$

where G_j is a sequence of positive numbers. It is well known that the system composed of all possible products

$$\left\{ \begin{array}{l} \Psi_{j_1}, \ldots, _{j_p}\left(x^{(1)}, \ldots, x^{(p)}\right) \\ \quad = g_{j_1}\left(x^{(1)}\right) \ldots g_{j_p}\left(x^{(p)}\right) \\ j_k = 0, 1, 2, \ldots, \quad k = 1, \ldots, p \end{array} \right\}. \tag{6}$$

is a complete orthonormal system in $L_2(A)$, where

$$A = \underbrace{\Delta \times \ldots \times \Delta}_{p\text{–times}}$$

It constitutes the basis for construction of the following sequence K_n:

$$K_n(x,u) = \sum_{j_1=0}^{q} \cdots \sum_{j_p=0}^{q} g_{j_1}\left(x^{(1)}\right) g_{j_1}\left(u^{(1)}\right) \cdots g_{j_p}\left(x^{(p)}\right) g_{j_p}\left(u^{(p)}\right) \tag{7}$$

where q depends on the length of the learning sequence, i.e., $q = q(n)$. It can be given in a shortened form as

$$K_n(x, u) = \sum_{|\underline{j}| \leq q} \Psi_{\underline{j}}(x) \Psi_{\underline{j}}(u) \tag{8}$$

where

$$\underline{j} = (j_1, \ldots, j_p) \text{ and } |\underline{j}| = \max_{1 \leq k \leq p} (j_k).$$

If $\Delta = (-\infty, \infty)$, then we design the PNNs based on the Hermite series given by

$$g_j(x) = \left(2^j j! \pi^{\frac{1}{2}}\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2}} H_j(x)$$

where

$$H_{j+1}(x) = 2xH_j(x) - 2jH_{j-1}(x)$$

and $H_0(x) = 1$, $H_1(x) = 2x$. It is easly seen that the orthonormal functions of the Hermite series can be recursively generated by

$$\left.\begin{array}{c} g_0(x) = \pi^{-\frac{1}{4}} e^{-\frac{x}{2}} \\ g_1(x) = 2^{\frac{1}{2}} \pi^{-\frac{1}{4}} x e^{-\frac{x}{2}} = 2^{\frac{1}{2}} x g_0(x) \\ g_{j+1}(x) = \left(\frac{2}{(j+1)}\right)^{\frac{1}{2}} x g_j(x) \\ -\left(\frac{j}{(j+1)}\right)^{\frac{1}{2}} g_{j-1}(x) \\ \text{for } j = 1, 2, \ldots, \end{array}\right\}. \tag{9}$$

It is known [43] that for the Hermite series $G_j = \text{const} \cdot j^{-12}$.

If $\Delta = [0, \infty)$, then we design the PNNs based on the Laguerre series given by

$$g_j(x) = e^{-\frac{x}{2}} L_j(x)$$

where

$$(j+1)L_{j+1}(x) = (2j + 1 - x)L_j(x) - jL_{j-1}(x)$$

and $L_0(x) = 1$, $L_1(x) = 1 - x$. It is easly seen that the orthonormal functions of the Laguerre series can be recursively generated by

$$\left.\begin{array}{c} g_0(x) = e^{-\frac{x}{2}} \\ g_1(x) = e^{-\frac{x}{2}}(1 - x) = g_0(x)(1 - x) \\ (j+1)g_{j+1}(x) = (2j + 1 - x)g_j(x) \\ -jg_{j-1}(x) \\ \text{for } j = 1, 2, \ldots, \end{array}\right\}.$$

It is known [43] that for the Laguerre series $G_j = \text{const} \cdot j^{-1/4}$.

If $\Delta = [-1, 1]$, then we design the PNNs based on the Legendre series given by

$$g_j(x) = \sqrt{\frac{2j+1}{2}} P_j(x)$$

where

$$(j+1)P_{j+1}(x) = (2j+1)xP_j(x) - jP_{j-1}(x)$$

and $P_0(x) = 1, P_1(x) = x$. It is easly seen that the orthonormal functions of the Legendre series can be recursively generated by

$$\left.\begin{array}{c} g_0(x) = \sqrt{\frac{1}{2}} \\ g_1(x) = \sqrt{\frac{3}{2}} x \\ (j+1)g_{j+1}(x) = \sqrt{(2j+1)(2j+3)} x g_j(x) \\ -\sqrt{\frac{(2j+3)}{(2j-1)}} j g_{j-1}(x) \\ \text{for } j = 1, 2, \ldots, \end{array}\right\}.$$

It is known [43] that for the Legendre series $G_j = \text{const} \cdot j^{1/2}$.

III. PNNs for Pattern Classification in Stationary Environment

Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ be a sequence ofindependent identically distributed (i.i.d.) pairs of random variables, Y takes values in the set of classes $S = \{1, \ldots, M\}$, whereas X takes values in $A \subset R^p$. The problem is to estimate Y from X and W_n, where $W_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ is a learning sequence. Suppose that p_m and f_m, $m = 1, \ldots, M$ are the prior class probabilities and the class conditional densities, respectively. We define a discriminant function of class j

$$d_j(x) = p_j f_j(x). \tag{10}$$

Let $L(i, j)$ be the loss incurred in taking action $i \in S$ when the class is j. We assume 0–1 loss function. For a decision function $\varphi{:}A \to S$ the expected loss is

$$R(\varphi) = \sum_{j=1}^{M} p_j \int_A L(\varphi(x), j) f_j(x) dx. \tag{11}$$

A decision function φ^* which classifies every $x \in A$ as coming from any class m for which

$$p_m f_m(x) = \max_j p_j f_j(x) = \max_j d_j(x) \tag{12}$$

is a Bayes-decision function and

$$R^* = R(\varphi^*) = \sum_{j=1}^{M} p_j \int_A L(\varphi^*(x), j) f_j(x) dx \tag{13}$$

is the minimal Bayes risk. The function $d_m(x)$ is called the Bayes-discriminant function. Let n_j be the number of observations from class j, $j = 1, \ldots, M$. We partition observations $X_1, \ldots, X_n$ into M subsequences

$$\begin{array}{c} X_1^{(1)}, \ldots, X_{n_1}^{(1)} \\ X_1^{(2)}, \ldots, X_{n_2}^{(2)} \\ \ldots \\ X_1^{(M)}, \ldots, X_{n_M}^{(M)}. \end{array} \tag{14}$$

As estimates of conditional densities f_j we apply nonparametric estimator in the form

$$\widehat{f}_{n_j}(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} K_{n_j}\left(x, X_i^{(j)}\right). \tag{15}$$

The prior probabilities p_j are estimated by

$$\widehat{p}_j = \frac{n_j}{n}. \tag{16}$$

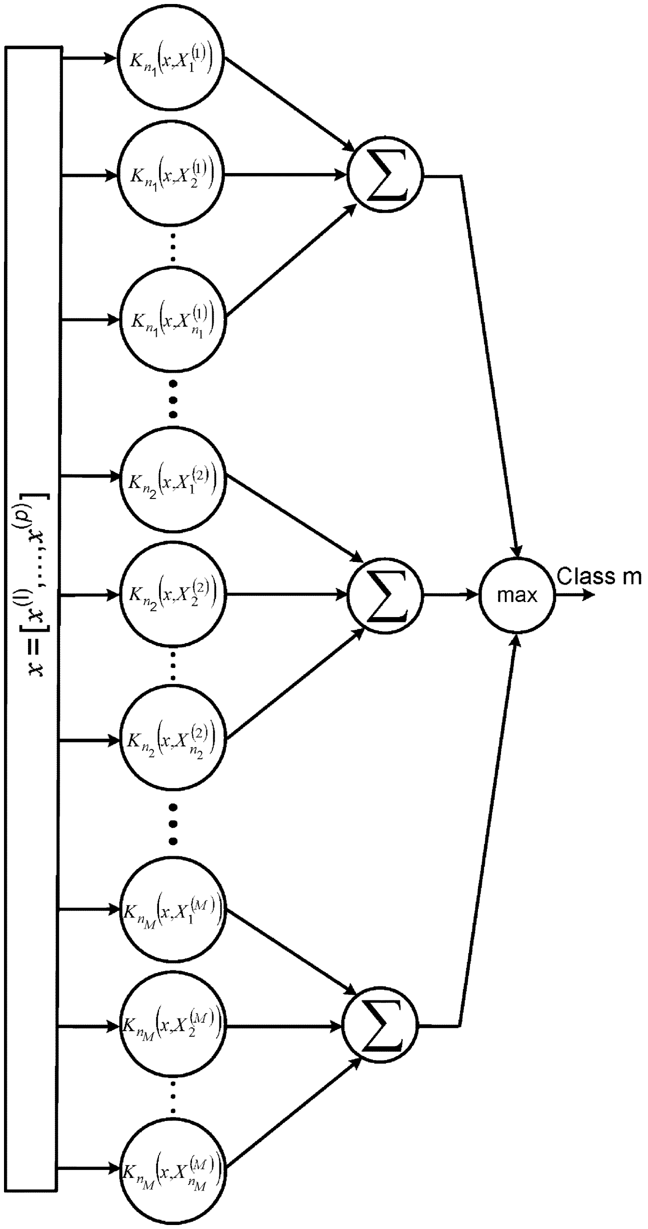

Fig. 1. Probabilistic neural network for pattern classification.

Combining (10), (15), and (16) we get the following discriminant function estimate:

$$\widehat{d}_{j,n}(x)=\frac{1}{n}\sum_{i=1}^{n_j}K_{n_j}\left(x,X_i^{(j)}\right) \tag{17}$$

and corresponding classification procedure

$$\begin{aligned}\widehat{\varphi}_n(x)=m\ \text{ if }\ &\sum_{i=1}^{n_m}K_{n_m}\left(x,X_i^{(m)}\right)\\ &\geq\sum_{i=1}^{n_j}K_{n_j}\left(x,X_i^{(j)}\right)\\ \text{for } i\neq m,\quad & i=1,\ldots,M.\end{aligned} \tag{18}$$

The probabilistic neural network realizing procedure (18) is shown in Fig. 1.

It was shown [7], [12], [47] that

$$R(\varphi_n)\xrightarrow{n}R^* \tag{19}$$

in probability (with pr. 1) if estimators (15) converge in probability (with pr. 1).

Example 1: For the Parzen kernel, procedure (18) classifies every $x\in A$ as coming from a class which maximizes

$$\frac{1}{h_{n_j}^p}\sum_{i=1}^{n_j}K\left(\frac{x-X_i^{(j)}}{h_{n_j}}\right)$$

for $j=1,\ldots,M$.

We will now derive classification procedures from the generalized-regression probabilistic neural network [35]. Instead of partition (14), we define

$$T_{ji}=\begin{cases}1 & \text{if } Y_i=j\\ 0 & \text{if } Y_i\neq j\end{cases} \tag{20}$$

for $i=1,\ldots,n$ and $j=1,\ldots,M$.

Observe that discriminant functions d_j, $j=1,\ldots,M$, can be presented in the form

$$d_j(x)=p_jf_j(x)=f(x)E[T_{ji}|X_i=x] \tag{21}$$

where

$$f(x)=\sum_{j=1}^{M}p_jf_j(x). \tag{22}$$

Therefore, the estimates of discriminant functions derived from regression model take the form

$$\widehat{d}_{j,n}(x)=\frac{1}{n}\sum_{i=1}^{n}T_{ji}K_n(x,X_i). \tag{23}$$

The classification procedure derived from estimator (23) takes the form

$$\begin{aligned}\widehat{\varphi}_n(x)=m\ \text{ if }\ &\sum_{i=1}^{n}T_{mi}K_n(x,X_i)\\ &\geq\sum_{i=1}^{n}T_{ji}K_n(x,X_i)\\ \text{for } i\neq m,\quad & i=1,\ldots,M.\end{aligned} \tag{24}$$

Generalized-regression neural network for pattern classification is presented in Fig. 2.

Example 2: For the Parzen kernel, procedure (24) classifies every $x\in A$ as coming from a class which maximizes

$$\frac{1}{h_n}\sum_{i=1}^{n}T_{ji}K\left(\frac{x-X_i}{h_n}\right)$$

for $j=1,\ldots,M$. The appropriate probabilistic neural network is shown in Fig. 3 assuming use of kernel (4) and the normalization of vectors x and X_i.

A recursive version of estimate (23) is given by

$$\widehat{d}_{j,n}(x)=\frac{1}{n}\sum_{i=1}^{n}T_{ji}K_i(x,X_i) \tag{25}$$

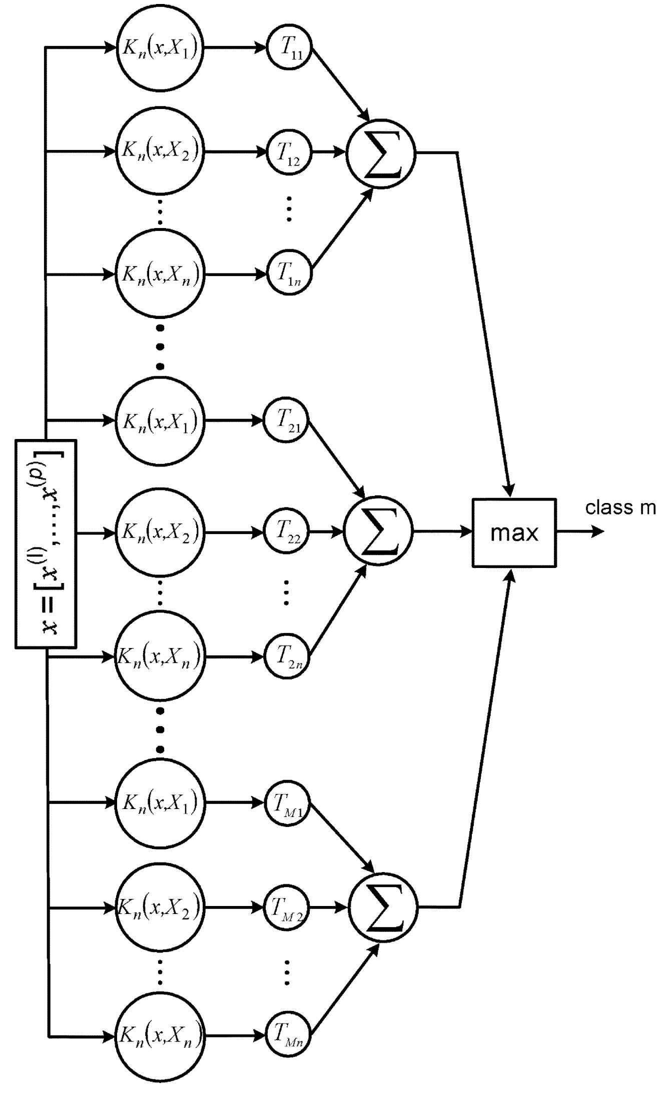

Fig. 2. Generalized-regression neural network for pattern classification.

or alternatively in the form

$$\widehat{d}_{j,n+1}(x) = \widehat{d}_{j,n}(x) + \frac{1}{n+1} \cdot \left[T_{j,n+1}K_{n+1}(x,X_{n+1}) - \widehat{d}_{j,n}(x)\right]. \quad (26)$$

The classification procedure becomes

$$\widehat{\varphi}_n(x) = m \text{ if } \sum_{i=1}^{n} T_{mi}K_i(x,X_i) \geq \sum_{i=1}^{n} T_{ji}K_i(x,X_i) \quad \text{for } i \neq m, \quad i = 1,\ldots,M. \quad (27)$$

The probabilistic neural network realizing procedure (27) is shown in Fig. 4. Whereas its Parzen-kernel version (for $M = 2$) is depicted in Fig. 5. The net in Fig. 5 consists of one neuron in the first layer having p inputs, coordinates of the vector X_n, $n = 1, 2, \ldots$. Let us notice that the role of weights is played by the coordinates of vector x. The second layer consists of two neurons with the feedback typical for recurrent neural networks.

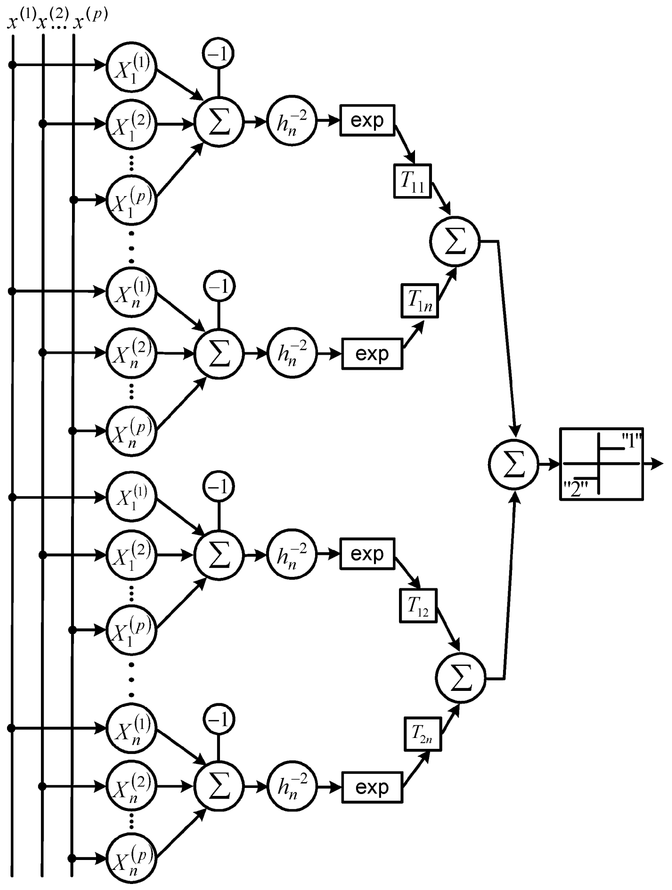

Fig. 3. Generalized-regression neural network based on the Parzen kernel for pattern classification $(M = 2)$.

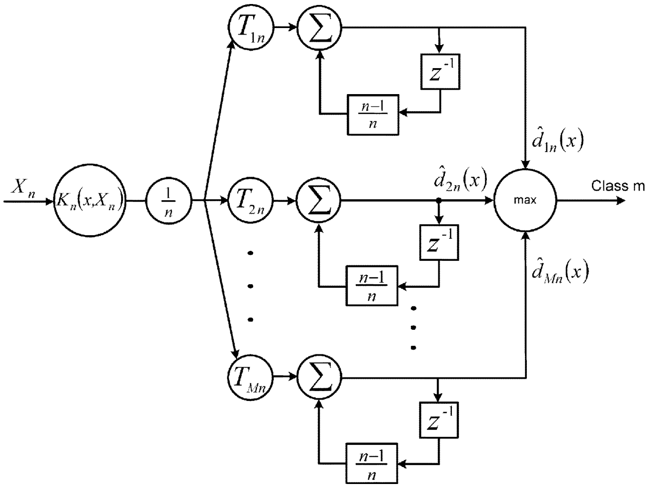

Fig. 4. Recursive generalized-regression neural network for pattern classification.

IV. Preliminaries to PNNs in Time-Varying Environment

In this section, we study a general problem of learning in the nonstationary environment. The results and theorems will be a starting point for construction of the PNNs in the next sections. Let us consider a sequence $\{(X_n, Y_n)\}$, $n = 1, 2, \ldots$, of independent pairs of random variables, where X_n-random variables having the probability density f_n taking values in the set $A \subset R^p Y_n$-random variables taking values in the set $B \subset R$.

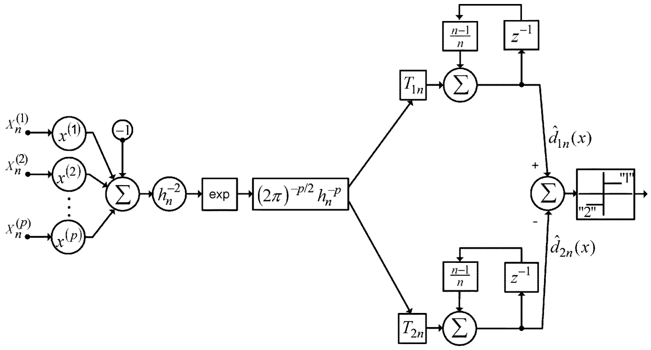

Fig. 5. Recursive generalized-regression neural network based on the Parzen kernel for pattern classification $(\mathrm{M} = 2)$.

We assume that probability distributions of the above random variables are completely unknown.

Let us define the following function:

$$R_n(x) \overset{df}{=} f_n(x)E[Y_n|X_n = x] \; n = 1, 2, \ldots. \tag{28}$$

From the assumption that the probability distributions are completely unknown, it follows that the sequence of functions (28) is also unknown. The goal of learning will be tracking the changing function R_n, $n = 1, 2, \ldots$.

Let $\{a_n\}$ be a sequence of numbers satisfying the following conditions:

$$a_n > 0, \quad a_n \xrightarrow{n} 0, \quad \sum_{n=1}^{\infty} a_n = \infty. \tag{29}$$

We will consider a nonparametric learning procedure of the following type:

$$\widehat{R}_{n+1}(x) = \widehat{R}_n(x) + a_{n+1}\left[Y_{n+1}K_{n+1}(x, X_{n+1}) - \widehat{R}_n(x)\right]$$
$$n = 0, 1, 2, \ldots \; \widehat{R}_0(x) = 0. \tag{30}$$

As the global measure of the learning process quality, we take

$$I_n = \int \left(\widehat{R}_n(x) - R_n(x)\right)^2 dx. \tag{31}$$

We will show that

$$EI_n \xrightarrow{n} 0 \quad \text{and} \quad I_n \xrightarrow{n} 0 \quad \text{with pr. 1.}$$

Define

$$r_n(x) = E\left[Y_n K_n(x, X_n)\right]. \tag{32}$$

Theorem 1: If the following conditions are satisfied:

$$a_n \int \operatorname{var}\left[Y_n K_n(x, X_n)\right] dx \xrightarrow{n} 0 \tag{33}$$
$$a_n^{-2} \int \left(r_n(x) - R_n(x)\right)^2 dx \xrightarrow{n} 0 \tag{34}$$
$$a_n^{-2} \int \left(R_{n+1}(x) - R_n(x)\right)^2 dx \xrightarrow{n} 0 \tag{35}$$

then

$$EI_n \xrightarrow{n} 0 \quad \text{with pr. 1.} \tag{36}$$

Theorem 2: If the following conditions are satisfied:

$$\sum_{n=1}^{\infty} a_n^2 \int \operatorname{var}\left[Y_n K_n(x, X_n)\right] dx < \infty \tag{37}$$
$$\sum_{n=1}^{\infty} a_n^{-1} \int \left(r_n(x) - R_n(x)\right)^2 dx < \infty \tag{38}$$
$$\sum_{n=1}^{\infty} a_n^{-1} \int \left(R_{n+1}(x) - R_n(x)\right)^2 dx < \infty \tag{39}$$

then

$$I_n \xrightarrow{n} 0 \quad \text{with pr. 1.} \tag{40}$$

Theorem 3: If the following conditions are satisfied:

$$\int \operatorname{var}\left[Y_n K_n(x, X_n)\right] dx = 0(n^A), \quad A > 0 \tag{41}$$
$$\int \left(R_{n+1}(x) - R_n(x)\right)^2 dx = 0(n^{-B}), \quad B > 0 \tag{42}$$
$$\int \left(R_n(x) - r_n(x)\right)^2 dx = 0(n^{-C}), \quad C > 0 \tag{43}$$

$$a_n = \frac{k}{n^a}, \quad k > 0, \quad 0 < a \leq 1 \tag{44}$$

then

$$EI_n \leq l_1 n^{-C} + l_2 n^{-r} \tag{45}$$

where l_1 and l_2 are positive constants and $r = \min[a - A, B - 2a, C - 2a]$ with $r > 0$ for $0 < a < 1$ and $0 < r < 1$ for $a = 1$.

It would be interesting to investigate if procedure (30) on the basis of learning set

$$(X_1, Y_1), \ldots, (X_n, Y_n) \tag{46}$$

allows to predict

$$R_{n+k}(x) = f_{n+k}(x)E[Y_{n+k}|X_{n+k} = x] \tag{47}$$

for $k \geq 1$. In the considered situation, performance measure (31) takes the form

$$I_{n,k} = \int \left(\widehat{R}_n(x) - R_{n+k}(x)\right)^2 dx. \tag{48}$$

The following result is a corollary from Theorems 1 and 2 and allows to predict R_{n+k}, $k \geq 1$, on the basis of a learning set of length n.

Corollary 1:

i) If the assumptions of Theorem 1 are satisfied, then

$$EI_{n,k} \xrightarrow{n} 0. \tag{49}$$

ii) If the assumptions of Theorem 2 are satisfied, then

$$I_{n,k} \xrightarrow{n} 0 \quad \text{with pr. 1.} \tag{50}$$

The next corollary follows immediately from Theorem 3.

Corollary 2: Under conditions of Theorem 3

$$EI_{n,k} \leq l_1 n^{-C} + l_2 n^{-r} + l(k) n^{-B}. \tag{51}$$

Above, symbol $l(k)$ denotes a positive constant which depends on k, the rest of the symbols are identical as in Theorem

3. Of course, the more steps k in prediction, the bigger value of the right side of expression (51) because $l(k)$ increases with bigger k (this is shown in the proof of the above corollary).

V. Problem Description and Presentation of Classification Rules

Let (X_n, V_n), $n = 1, 2, \ldots$ be a sequence of independent pairs of random variables. Random variable X_n has an interpretation of the pattern connected with a given class and takes values in space A, $A \subset R^p$. Random variable V_n takes values in set $\{1, \ldots, M\}$ called the set of classes, specifying the class number.

A priori probabilities of occurrence of class m in moment $n\ (m = 1, \ldots, M; n = 1, 2, \ldots)$ will be denoted by p_{mn}, i.e., $p_{mn} = P(V_n = m)$. It is assumed that there are conditional probability densities f_{mn} of random variable X_n on condition that $V_n = m$. These densities are called densities in classes.

The classification rule is a measurable mapping φ_n: $A \rightarrow \{1, \ldots, M\}$. The measure of quality of the rule φ_n is the probability of misclassification

$$P\left(\varphi_n(X_n) \neq V_n\right) \overset{df}{=} L_n(\varphi_n). \tag{52}$$

The rule that minimizes the above index is called the Bayes' rule. The Bayes' rule in moment n is denoted by φ_n^* and the value of $L_n(\varphi_n^*)$ is denoted by L_n^*, i.e.

$$L_n\left(\varphi_n^*\right) = L_n^*. \tag{53}$$

We will define the following function:

$$d_{mn}(x) = p_{mn} f_{mn}(x). \tag{54}$$

This function will be called the discriminant function of class m in moment n. Generalizing considerations for the stationary case [9], it is easily seen that the rule φ_n^* has the form

$$\begin{aligned} &\varphi_n^*(X_n) = m, \\ &\text{if } d_{mn}(X_n) > d_{in}(X_n) \\ &\qquad \text{for } i \neq m, \quad i = 1, \ldots, M, \quad n = 1, 2, \ldots. \end{aligned} \tag{55}$$

We assume that both *a priori* probabilities p_{mn} and densities in classes f_{mn}, $m = 1, \ldots, M$, $n = 1, 2, \ldots$, are completely unknown. For this reason, we use empirical classification rules based on estimators of discriminant functions.

The problem of nonparametric pattern classification in a nonstationary case boils down to constructing empirical classification rules that on the basis of the learning sequence

$$(X_1, V_1), \ldots, (X_n, V_n) \tag{56}$$

would classify pattern X_{n+k}, $k \geq 1$. It is, of course, an issue of prediction of patterns having nonstationary probability distributions. In the case of complete probabilistic information, i.e., the knowledge of discriminant functions

$$d_{m,n+k}(x) = p_{m,n+k}\, f_{m,n+k}(x) \tag{57}$$

pattern X_{n+k} could be classified by means of rule (55).

Let $\widehat{d}_{mn}$ be the estimator constructed on the basis of the learning sequence (14) of function $d_{m,n+k}$, $k \geq 1$. We will consider the empirical rules of the form

$$\begin{aligned} &\widehat{\varphi}_n(X_{n+k}) = m, \\ &\text{if } \widehat{d}_{mn}(X_{n+k}) > \widehat{d}_{in}(X_{n+k}) \\ &\qquad \text{for } i \neq m, \quad i = 1, \ldots, M. \end{aligned} \tag{58}$$

The sequence of empirical rules $\widehat{\varphi}_n$ is called the classification-learning algorithm. The rule $\widehat{\varphi}_n$ is a function of the learning sequence $(X_1, V_1), \ldots, (X_n, V_n)$ and classified pattern X_{n+k}, $k \geq 1$.

We will now construct an estimator of function (57). We will first show that the procedure (30) that was presented in Section IV can be used for estimation of time-varying discriminant functions (57). Let

$$T_{mn} = \begin{cases} 1 & \text{if } V_n = m \\ 0 & \text{if } V_n \neq m \end{cases}. \tag{59}$$

Discriminant function (54) can be presented as

$$d_{mn}(x) = f_n(x) E[T_{mn} | X_n = x] \tag{60}$$

where

$$f_n(x) = \sum_{m=1}^{M} p_{mn} f_{mn}(x). \tag{61}$$

Comparing (60) and (30), setting $Y_n = T_{mn}$ for the fixed m, we use procedure (30) for the estimation of discriminant functions (54)

$$\begin{aligned} &\widehat{d}_{m,n+1}(x) \\ &\quad = \widehat{d}_{m,n}(x) + a_{n+1}^{(m)} \left(T_{m,n+1} K_{n+1}^{(m)}(x, X_{n+1}) - \widehat{d}_{m,n}(x)\right) \\ &d_{m,0}(x) \\ &\quad = 0 \quad \text{for } m = 1, \ldots, M, \quad n = 0, 1, 2, \ldots. \end{aligned} \tag{62}$$

On the basis of considerations of Section IV (see Corollary 1), $\widehat{d}_{mn}$ can be used not only for estimation of d_{mn}, but also for estimation of $d_{m,n+k}$, $k \geq 1$.

The structural scheme of a system that realizes the classification algorithm with the use of procedure (58) is presented in Fig. 6. Sequences $\{K_n\}$ and $\{a_n\}$, on which this procedure is based, should generally satisfy different conditions depending on the class number m. That is why in Fig. 6., symbols $K_n^{(m)}$ and $a_n^{(m)}$, $m = 1, \ldots, M$ are used. For convenience, we drop a dependence on m in theorems presented in the next sections.

We should point out that in order to classify pattern X_{n+k}, $k \geq 1$, it is necessary to store the whole learning set of the length n. Next, when the pattern X_{n+k} to be classified appears, procedure (62) is activated starting from $n = 0$ and putting $x = X_{n+k}$.

VI. Asymptotic Optymality of Classification Rules

As was mentioned in the introduction, the concept of asymptotic optimality of classification rules in the nonstationary case has not been studied in literature yet.

In this section we will present appropriate definitions and show that when the length of learning set (14) increases, clas-

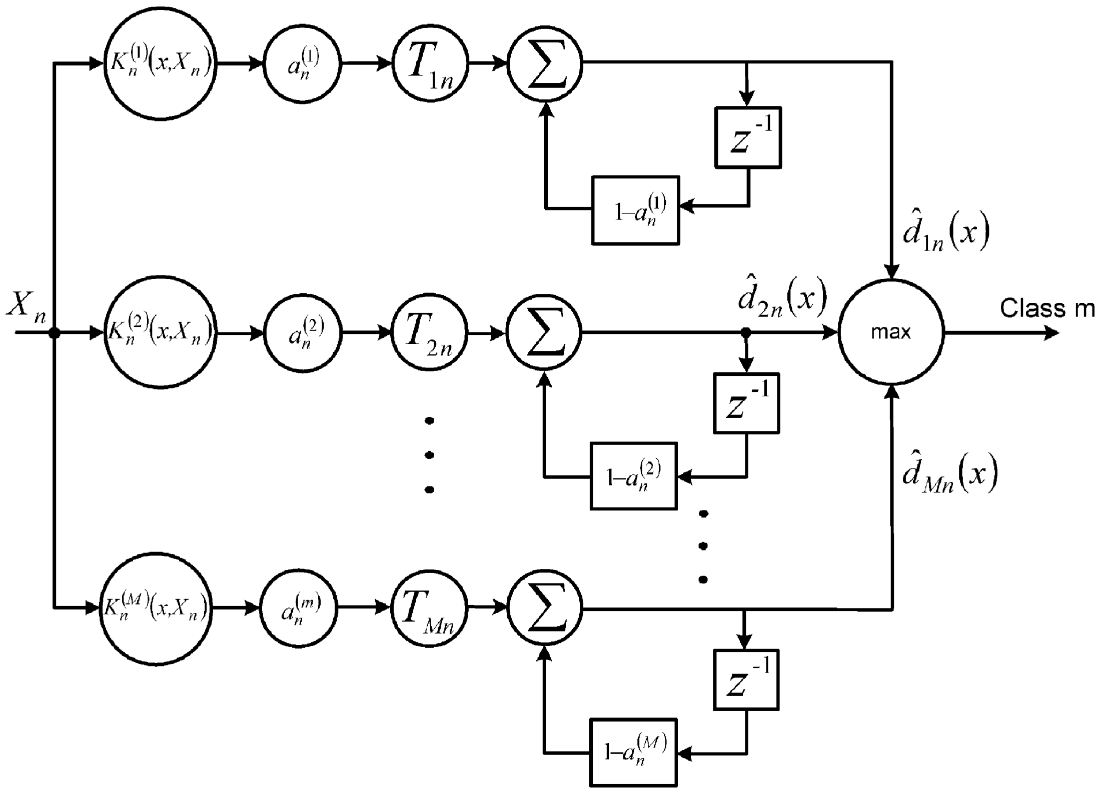

Fig. 6. Recursive generalized-regression neural network for pattern classification.

sification algorithms become increasingly similar to optimal algorithm (55) which could be determined when *a priori* probabilities p_{mn} and densities in classes f_{mn} are known, $m = 1, \ldots, M, n = 1, 2, \ldots$.

The global performance measure of classification rule (55), classifying X_{n+k}, is the probability of misclassification in moment $n + k$

$$L_{n+k}\left(\varphi_{n+k}^{*}\right) = P\left(\varphi_{n+k}^{*}(X_{n+k}) \neq V_{n+k}\right). \qquad (63)$$

As a performance measure of empirical rule (58) we take

$$L_{n+k}(\widehat{\varphi}_n) = (P\widehat{\varphi}_n(X_{n+k}) \neq V_{n+k}|X_1, V_1, \ldots, X_n, V_n) \qquad (64)$$

i.e., the probability of missclassification of pattern X_{n+k} determined on the basis of rule (58) and learning sequence (14).

Definition 1: Classification algorithm $\widehat{\varphi}_n$ defined by (58) is weakly asymptotically optimal when

$$EL_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi_{n+k}^{*}\right) \xrightarrow{n} 0. \qquad (65)$$

Definition 2: Classification algorithm $\widehat{\varphi}_n$ defined by (58) is strongly asymptotically optimal when

$$L_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi_{n+k}^{*}\right) \xrightarrow{n} 0 \qquad (66)$$

with probability 1.

The following theorem ensures asymptotic optimality of the rule (58) if estimator $\widehat{d}_{mn}$ (expressed by formula (62) "follows" the changes of the discriminant function $d_{m,n+k}$, when $n \to \infty$.

Theorem 4: Let χ_n, $n = 1, 2, \ldots$, be a sequence of sets in R^p such that for $\varepsilon > 0$, the following conditions are satisfied

$$\int_{\chi_n} f_{n+k}(x)dx \geq 1 - \frac{\varepsilon}{2}, \quad n = 1, 2, \ldots \qquad (67)$$

and

$$\mu(\chi_n) \leq \text{const.}, \quad n = 1, 2, \ldots \qquad (68)$$

where

$$f_n(x) = \sum_{m=1}^{M} p_{mn} f_{mn}(x).$$

A. *If*

$$E\int\left(\widehat{d}_{mn}(x) - d_{m,n+k}(x)\right)^2 dx \xrightarrow{n} 0 \qquad (69)$$

then the pattern classification rule (58) is weakly asymptotically optimal.

B. *If*

$$\int\left(\widehat{d}_{mn}(x) - d_{m,n+k}(x)\right)^2 dx \xrightarrow{n} 0 \qquad (70)$$

with pr. 1, then the pattern classification rule (58) is strongly asymptotically optimal.

Remark 1: It is always possible to select sequence χ_n in such a way so that condition (67) could be met. However, it does not mean that the condition (68) is automatically satisfied. For instance, if densities in classes are of the exponential type

$$f_{mn}(x) = \lambda_{mn} e^{-\lambda_{mn} x}, \quad x \geq 0 \qquad (71)$$

and

$$\lambda_{mn} \xrightarrow{n} 0, \quad m = 1, \ldots, M \tag{72}$$

then there does not exist sequence χ_n that satisfies conditions (67) and (68) at the same time. However, if the densities in the classes are of the type "movable argument"

$$f_{mn}(x) = f_m(x - c_{mn}) \tag{73}$$

then it is possible to take (in the scalar case)

$$\chi_n = [c_{mn} - \nu, \, c_{mn} + \nu] \tag{74}$$

for sufficiently large ν.

The speed of convergence of procedure (58) in the sense (69) can be evaluated by means of Theorem 3 taking into account Corollary 2 concerning prediction. For this purpose, it is necessary to specify constants A, B and C that are present in the assumptions of that theorem. We will do it in Section IX, considering a particular type of nonstationarity. Now, we will connect the speed of convergence of (65) with the speed of convergence of (69). As we know (Definition 1), convergence (65) ensures a weak asymptotic optimality of the rule (58).

Let us denote

$$t_{m,n+k} = \int \cdots \int \left|x^{(1)}\right|^s \ldots \left|x^{(p)}\right|^s f_{m,n+k}(x)dx, \quad s > 0 \tag{75}$$

and

$$t_{n+k} = \sum_{m=1}^{M} t_{m,n+k} \tag{76}$$

where $t_n < \infty, n = 1, 2, \ldots$.

Theorem 5: Let us assume that

$$E \int \left(\widehat{d}_{mn}(x) - d_{m,n+k}(x)\right)^2 dx = 0(u_n). \tag{77}$$

A. *If sequence t_n is bounded, i.e.,*

$$t_n \leq \text{const.}, \quad n = 1, 2, \ldots \tag{78}$$

then

$$EL_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi^*_{n+k}\right) = 0\left(u_n^{\frac{s}{2s+1}}\right). \tag{79}$$

B. *If sequence t_n is not bounded, i.e.,*

$$t_n \geq \text{const.} > 0, \quad n > n_0 \tag{80}$$

then

$$EL_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi^*_{n+k}\right) = 0\left(t_{n+k}u_n^{\frac{s}{2s+1}}\right). \tag{81}$$

From the proof of Theorem 5 it follows that if nonstationarity densities in classes are of the "movable argument" type (73), then conclusion a) is true when

$$\int \ldots \int \left|x^{(1)}\right|^s \ldots \left|x^{(p)}\right|^s f_m(x)dx < \infty \tag{82}$$

for a certain $s > 0$.

In the next sections, we will consider procedures of type (58) constructed on the basis of the Parzen kernel and the orthogonal series method. Using general Theorems 1 and 2 from Section IV and Theorem 4, we will present conditions ensuring the convergence of algorithm (58).

VII. Classification Procedures Based on Parzen Kernels

As we remember (Section II), kernel K can be expressed in the following way:

$$K(x) = \prod_{i=1}^{p} H\left(x^{(i)}\right). \tag{83}$$

Let us assume that

$$\sup_{v \in R} |H(v)| < \infty \tag{84}$$

$$\int_R H(v)dv = 1 \tag{85}$$

$$\int_R H(v)v^j dv = 0, \quad j = 1, \ldots, r-1 \tag{86}$$

$$\int_R \left|H(v)v^k\right| dv < \infty, \quad k = 1, \ldots, r. \tag{87}$$

For $r = 2$, the above conditions are satisfied by function (3). In Fig. 7 we show the PNNs with the Parzen kernel (4). For $r = 4$, conditions (83)–(87) are met by the function

$$H(v) = \frac{3}{2\sqrt{2\pi}}\left(1 - \frac{v^2}{3}\right)e^{-\frac{1}{2}v^2}.$$

The appropriate conditions for the convergence of the classification algorithm will depend on smooth properties of the density function f_{mn} $(m = 1, \ldots, M, n = 1, 2, \ldots)$. We define

$$\delta^{\underline{i}}_{mn} = \int \left[\frac{\partial^r}{\partial x^{(i_1)} \ldots \partial x^{(i_r)}} f_{mn}(x)\right]^2 dx \tag{88}$$

where $\underline{i} = (i_1, \ldots, i_r)$, $i_k = 1, \ldots, p$, $k = 1, \ldots, r$.

The following result is a corollary from Theorems 1 and 4.

Corollary 3: If function K satisfies conditions (83)–(87), assumptions (67) and (68) hold, $h_n \to 0$, and

$$a_n h_n^{-p} \xrightarrow{n} 0 \tag{89}$$

$$a_n^{-2} p_{mn}^2 h_n^{2r} \delta^{\underline{i}}_{mn} \xrightarrow{n} 0 \tag{90}$$

$$a_n^{-2} |p_{m,n+1} - p_{mn}| \int f_{mn}^2(x)dx \xrightarrow{n} 0 \tag{91}$$

$$a_n^{-2} p_{mn}^2 \int \left(f_{m,n+1}(x) - f_{mn}(x)\right)^2 dx \xrightarrow{n} 0 \tag{92}$$

then the pattern classification rule (58) is weakly asymptotically optimal.

Corollary 4 is a consequence of Theorems 2 and 4.

Corollary 4: If function K satisfies conditions (83)–(87), assumptions (67) and (68) hold, $h_n \to 0$, and

$$\sum_{n=1}^{\infty} a_n^2 h_n^{-p} < \infty \tag{93}$$

$$\sum_{n=1}^{\infty} a_n^{-1} h_n^{2r} \delta^{\underline{i}}_{mn} < \infty \tag{94}$$

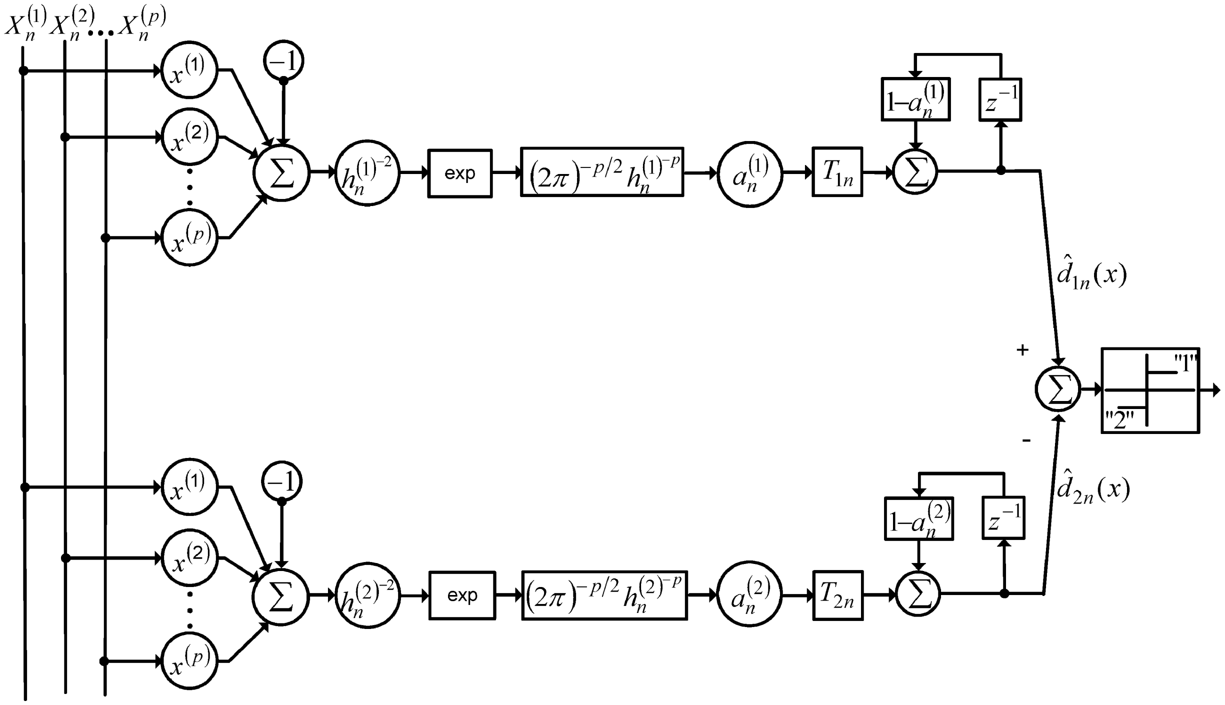

Fig. 7. Recursive generalized-regression neural network based on the Parzen kernel for pattern classification in time-varying environment $(\mathrm{M} = 2)$.

$$\sum_{n=1}^{\infty} a_n^{-1} (p_{m,n+1} - p_{mn})^2 \int f_{mn}^2(x)dx < \infty \qquad (95)$$

$$\sum_{n=1}^{\infty} a_n^{-1} p_{mn}^2 \int (f_{m,n+1}(x) - f_{mn}(x))^2 \, dx < \infty \qquad (96)$$

then the pattern classification rule (58) is strongly asymptotically optimal.

VIII. Classifications Procedures Based on Orthogonal Series

Let us denote

$$S_{mn} = \int \left(\sum_{|\underline{j}| \leq q} l_{\underline{j}n}^m \Psi_{\underline{j}}(x) - f_{mn}(x) \right)^2 dx \qquad (97)$$

where

$$l_{\underline{j}n}^m = \int f_{mn} \Psi_{\underline{j}}(x)dx. \qquad (98)$$

The following result is a corollary from Theorems 1 and 4.

Corollary 5: If conditions (91) and (92) are satisfied, $q(n) \to \infty$, and

$$a_n \left(\sum_{j=0}^{q(n)} G_j^2 \right)^p \xrightarrow{n} 0 \qquad (99)$$

$$a_n^{-2} p_{mn}^2 S_{mn} \xrightarrow{n} 0 \qquad (100)$$

then the pattern classification rule (58) is weakly asymptotically optimal.

Corollary 6 is a consequence of the general Theorems 2 and 4.

Corollary 6: If conditions (95) and (96) are satisfied, $q(n) \to \infty$, and

$$\sum_{n=1}^{\infty} a_n^2 \left(\sum_{j=0}^{q(n)} G_j^2 \right)^p < \infty \qquad (101)$$

$$\sum_{n=1}^{\infty} a_n^{-1} p_{mn}^2 S_{mn} < \infty \qquad (102)$$

then the pattern classification rule (58) is strongly asymptotically optimal.

Conditions (100) and (102) take a more concrete form depending on the smooth properties of function f_{mn} and the orthogonal series used. As an example, we will use the multidimensional Hermite series.

Let

$$D_{mn}^l(x; f_{mn}) \in L_2(R^p) \qquad (103)$$

where

$$D_{mn}^l(x; f_{mn}) = \prod_{j=1}^{p} \left(x^{(j)} - \frac{\partial}{\partial x^j} \right)^l f_{mn}(x), \quad l > 1.$$

Then

$$S_{mn} \leq \left\| D_{mn}^l(x; f_{mn}) \right\|_{L_2}^2 q^{-pl}(n). \qquad (104)$$

The above inequality is a generalization to the multidimensional and nonstationary case of Walter's result [43]. By means of this inequality, conditions (100) and (102) can be expressed as

$$a_n^{-2} p_{mn}^2 q^{-pl}(n) \left\| D_{mn}^l \right\|_{L_2}^2 \xrightarrow{n} 0 \tag{105}$$

and

$$\sum_{n=1}^{\infty} a_n^{-1} p_{mn}^2 q^{-pl}(n) \left\| D_{mn}^l \right\|_{L_2}^2 < \infty. \tag{106}$$

As we will see in the next section, these conditions take a simple form for a particular type of nonstationarity of function f_{mn}.

IX. Non-Stationarity of the Type "Movable Argument"

In order to simplify our considerations, let us examine the one-dimensional case and assume that *a priori* probabilities p_{mn} do not change with time. In regard to density in classes, we assume that they are of the form

$$f_{mn}(x) = f_m(x - c_{mn})$$

for $n = 1, 2, \ldots$, $m = 1, \ldots, M$, where $x \in R^1$. The above case occurs most often in practice [23]. In Tables I and II we present conditions that ensure the weak and strong asymptotic optimality of the PNNs [procedure (58)] constructed on the basis of the Parzen kernel. Tables III and IV show the analogous conditions for the Hermite orthonormal series.

In regard to smooth properties of function f_m, $r = 2$ (conditions (90) and (94)) is assumed in the case of the use of the Parzen kernel and $m = 2$ (conditions (100) and (102)) in the case of the use of the Hermite series. However, as follows from the last columns of Tables III and IV, the use of the orthogonal series method requires more assumptions imposed on function f_m and its derivatives. For instance, let us assume that sequence c_{mn} representing density function nonstationarity is of the type

$$c_{mn} = n^{t_m}, \quad t_m > 0, \quad m = 1, \ldots, M \;\; n = 1, 2, \ldots. \tag{107}$$

Analyzing all conditions specified in Tables I–IV, it is possible to establish within what limits parameters t_m, $m = 1, \ldots, M$ should be contained so that Corollaries 3–6 could be true. The results of such an analysis are presented in Table V.

It is easily seen that the use of the orthogonal series requires much more strict assumptions as regards the range within which parameters t_m can be contained.

Using Theorem 3 and Corollary 1 we will now evaluate the speed of convergence of algorithms (58) and (62). In procedure (62), let us select sequences h_n, $q(n)$ and a_n of the following type:

$$h_n = k_1 n^{-H}, \quad k_1 > 0, \quad H > 0 \tag{108}$$

$$q(n) = [k_2 n^Q], \quad k_2 > 0, \quad Q > 0 \tag{109}$$

$$a_n = \frac{k}{n^a}, \quad k > 0, \quad a > 0. \tag{110}$$

TABLE I
Conditions for Weak Convergence of PNNs Based on the Parzen Kernel

Condition	$L_{n+k}(\widehat{\varphi}) - L_{n+k}(\varphi_{n+k}^*) \xrightarrow{n} 0$ in probability	Assumptions
(89)	$a_n h_n^{-1} \xrightarrow{n} 0$	
(90)	$a_n^{-1} h_n^2 \xrightarrow{n} 0$	$f_m^{''} \in L_2, \;\; r = 2$
(92)	$a_n^{-1} \lvert c_{m,n+1} - c_{mn} \rvert \xrightarrow{n} 0$	$f_m, \; f_m^{'}, \; f_m^{''} \in L_2$

TABLE II
Conditions for Strong Convergence of PNNs Based on the Parzen Kernel

Condition	$L_{n+k}(\widehat{\varphi}) - L_{n+k}(\varphi_{n+k}^*) \xrightarrow{n} 0$ with pr. 1	Assumptions
(93)	$\sum_{n=1}^{\infty} a_n^2 h_n^{-1} < \infty$	
(94)	$\sum_{n=1}^{\infty} a_n^{-1} h_n^4 < \infty$	$f_m^{''} \in L_2, \; r = 2$
(96)	$\sum_{n=1}^{\infty} a_n^{-1} (c_{m,n+1} - c_{mn})^2 < \infty$	$f_m, \; f_m^{'}, \; f_m^{''} \in L_2$

TABLE III
Conditions for Weak Convergence of PNNs Based on the Orthogonal Series

Condition	$L_{n+k}(\widehat{\varphi}) - L_{n+k}(\varphi_{n+k}^*) \xrightarrow{n} 0$ in probability	Assumptions
(99)	$a_n q^{5/6}(n) \xrightarrow{n} 0$	$G_j = \text{const. } (j+1)^{-\frac{1}{12}}$ $l = 2$
(100)	$a_n^{-2} (c_{mn}^4 + 1) q^{-2}(n) \xrightarrow{n} 0$	$f_m, \; f_m^{'}, \; f_m^{''} \in L_2$ $\int x^4 f_m^2(x)\, dx < \infty$
(92)	$a_n^{-1} \lvert c_{m,n+1} - c_{mn} \rvert \xrightarrow{n} 0$	$\int x^2 f_m^2(x)\, dx < \infty$

TABLE IV
Conditions for Strong Convergence of PNNs Based on the Orthogonal Series

Condition	$L_{n+k}(\widehat{\varphi}) - L_{n+k}(\varphi_{n+k}^*) \xrightarrow{n} 0$ with. pr. 1	Assumptions
(101)	$\sum_{n=1}^{\infty} a_n^2 q^{5/6}(n) < \infty$	$G_j =$ $\text{const.}(j+1)^{-\frac{1}{12}}$
(106)	$\sum_{n=1}^{\infty} a_n^{-1} (c_{mn}^4 + 1) q^{-2}(n) < \infty$	$l = 2$ $f_m, \; f_m^{'}, \; f_m^{''} \in L_2$
(96)	$\sum_{n=1}^{\infty} a_n^{-1} (c_{m,n+1} - c_{mn})^2 < \infty$	$\int x^4 f_m^2(x)\, dx < \infty$ $\int x^2 f_m^2(x)\, dx < \infty$

TABLE V
Conditions Imposed on Parameter t_m—Nonstationarity of the Type "Movable Argument"

Method	Weak convergence	Strong convergence
Parzen	$0 < t_m < 1$	$0 < t_m < \frac{1}{7}$
Orthogonal series	$0 < t_m < \frac{1}{11}$	conditions of Corollary 6 are not satisfied

A. *Speed of Convergence of Algorithms Based on the Parzen Kernel*

With reference to the symbols from Theorem 3, we obtain

$$A = H, \quad B = 2(1 - t_m), \quad C = 4H \tag{111}$$

consequently

$$E \int \left(\widehat{d}_{mn}(x) - d_{m,n+k}(x) \right)^2 dx \\ \leq l_1 n^{-4H} + l_2 n^{-r} + l(k) n^{-2(1-t_m)} \tag{112}$$

where

$$r = \min \left[a - H, 2(1 - t_m - a), 2(2H - a) \right]. \tag{113}$$

If

$$\int |x|^s f_m(x) dx < \infty, \quad s > 0 \tag{114}$$

then, from Theorem 5, we obtain

$$EL_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi^*_{n+k}\right) = 0\left(n^{-\frac{As}{s+1}}\right) \tag{115}$$

where

$$A = \min \left[4H, r, 2(1 - t_m) \right]. \tag{116}$$

B. *Speed of Convergence of Algorithms Based on the Hermite Orthonormal Series*

In this case we have

$$A = \frac{5}{6}Q, \quad B = 2(1 - t_m), \quad C = 2(Q - 2t_m). \tag{117}$$

From Corollary 1 it follows that:

$$E \int \left(\widehat{d}_{mn}(x) - d_{m,n+k}(x) \right)^2 dx \\ \leq l_1 n^{2(2t_m - Q)} + l_2 n^{-r} + l(k) n^{-2(1-t_m)} \tag{118}$$

where

$$r = \min \left[a - \frac{5}{6}Q, 2(1 - t_m - a), 2(Q - 2t_m - a) \right. . \tag{119}$$

If condition (114) holds, then

$$EL_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi^*_{n+k}\right) = 0\left(n^{-\frac{Bs}{s+1}}\right) \tag{120}$$

where

$$B = \min \left[2(Q - t_m), r, 2(1 - t_m) \right]. \tag{121}$$

Analyzing formulas (115) and (120) we see that the influence of parameters t_m on the speed of convergence of both algorithms is much more significant in the case of use of the algorithm based on the Hermite series (it results in a decrease of this speed).

In all the above considerations, the same degree of smooth properties of function f_m was assumed: $r = 2$ for the algorithm based on the Parzen kernel and $l = 2$ for the algorithm based on the Hermite orthogonal series method. It is easy to prove that for $r > 2$, the range within which the parameters t_m, $m = 1, \ldots, M$ are contained and which ensures weak asymptotic optimality of the algorithm does not widen. For $l \geq 2$ (with additional assumptions as regards function f_m and its derivatives up to the lth order), the following inequality holds:

$$E \int \left(\widehat{d}_{mn}(x) - d_{m,n+k}(x) \right)^2 dx \\ \leq l_1 n^{l(2t_m - Q)} + l_2 n^{-r} + l(k) n^{-2(1-t_m)} \tag{122}$$

where

$$r = \min \left[a - \frac{5}{6}Q, \ 2(1 - t_m - a), \ Ql - 2lt_m - 2a \right].$$

From the last inequality it follows that:

$$0 < t_m < \frac{3l - 5}{8l - 5} \approx \frac{3}{8} \tag{123}$$

for a sufficiently large l. In other words, for the algorithm based on the Hermite series, a significant increase of smooth properties of function f_m allows the widening of the range within which the parameters parameters t_m are contained, but this "widened" range is nevertheless significantly narrower than in the case of the use of the algorithm based on the Parzen kernel ($0 < t_m < 1$).

X. Simulation Results

In this section, we apply the PNNs based on the Parzen kernel for estimation of time-varying probability densities and for classification of time-varying signals.

A. *Estimation of Time-Varying Probability Densities*

Let $\{X_n\}$ be a sequence of independent random variables with probability densities $f_n(x) = f(x - n^t)$. It is easily seen that if

$$p_{mn} = \begin{cases} 1 & \text{when } m = 1 \\ 0 & \text{when } m \neq 1 \end{cases}$$

then algorithm (62) can be used for nonparametric learning of time-varying probability densities f_n. In this case convergence of the PNNs follows from Corollaries 3–6. This problem was also investigated in works [28] and [41] but the authors assumed then that the sequence of nonstationary probability density functions is convergent in a specified sense to a finite limit.

Procedure (62) applied for nonparametric estimation of time-varying probability densities takes the form

$$\widehat{f}_{n+1}(x) = \widehat{f}_n(x) + a_{n+1}\left(K_{n+1}(x, X_{n+1}) - \widehat{f}_n(x) \right).$$

Let us choose

$$h_n = kn^{-H}, \quad k > 0, \quad H > 0 \\ a_n = n^{-a}.$$

Depending on the parameter t in model $f_n(x) = f(x - n^t)$ we pick up parameters a and H such that

$$a < 1 - t, \quad \frac{a}{2} < H < a$$

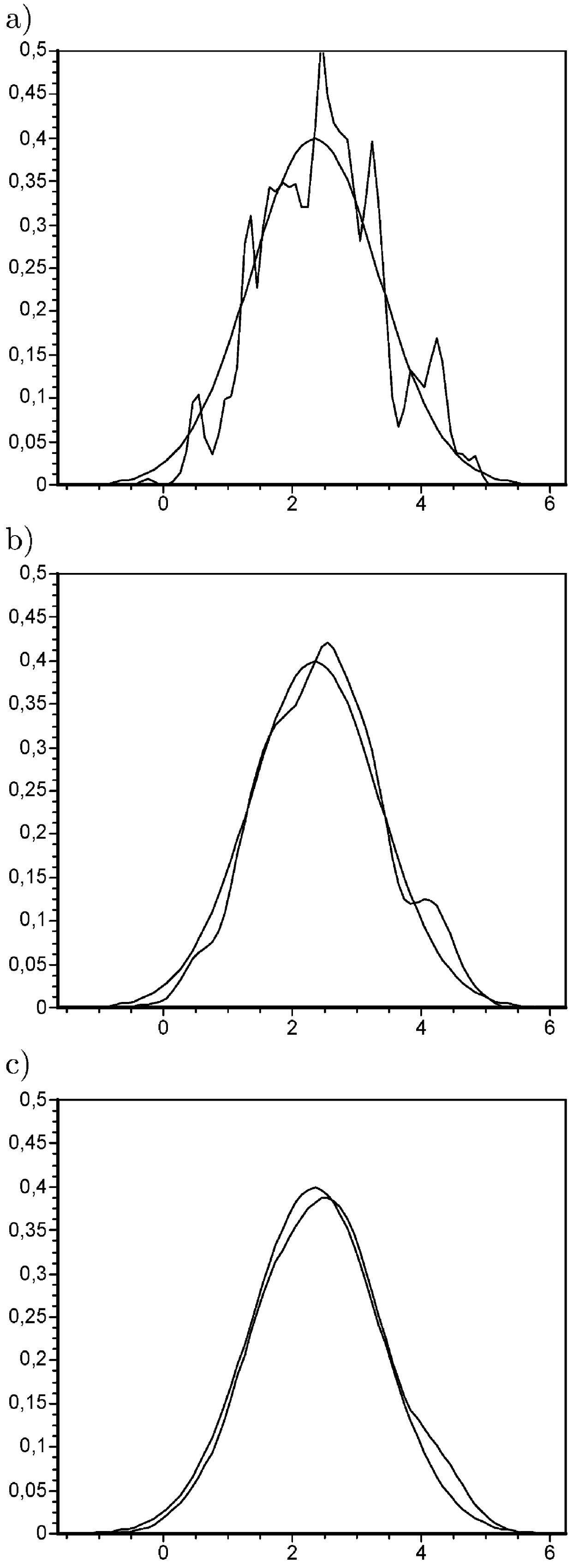

Fig. 8. Time-varying probability density estimation for $t = 0.1$, $a = 0.3$, $H = 0.5$, $k = 1, 3, 5$.

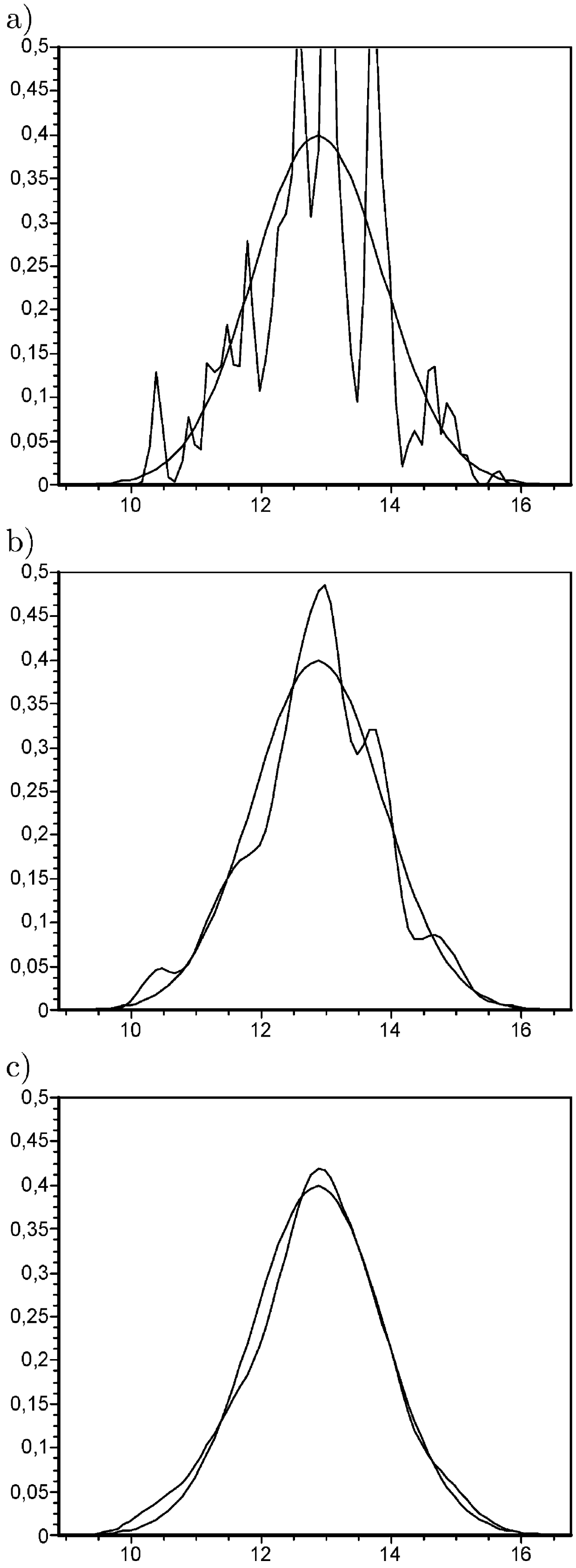

Fig. 9. Time-varying probability density estimation for $t = 0.3$, $a = 0.4$, $H = 0.3$, $k = 1, 3, 5$.

ensuring the convergence of the PNNs. In Figs. 8, 9, and 10 we show the results of simulations for $t = 0.1$, 0.3 and 0.5 respectively. In each case $n = 5000$, kernel (3) is used and the smoothing parameter $k = 1, 3, 5$. We observe that the best re-

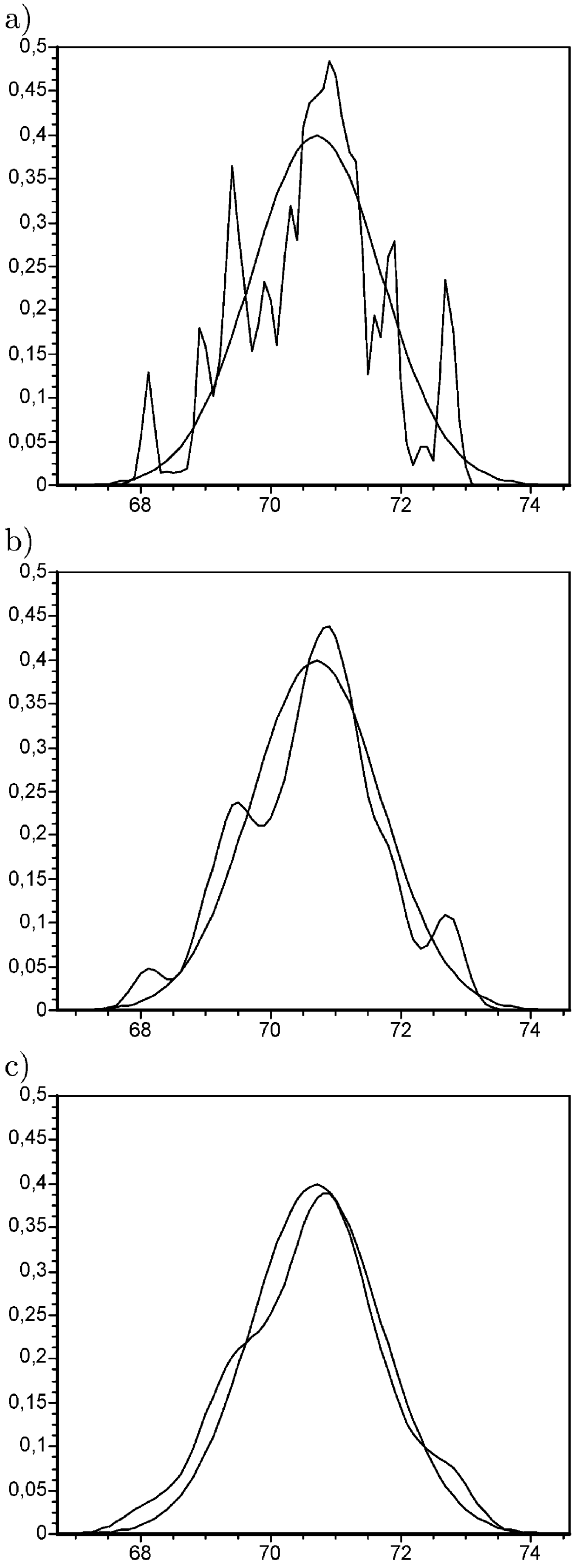

Fig. 10. Time-varying probability density estimation for $t = 0.5$, $H = 0.3$, $a = 0.4$, $k = 1, 3, 5$.

sults are obtained for $k = 5$. The problem of smoothing parameter selection is much more difficult in the nonstationary case and is a subject of the future research.

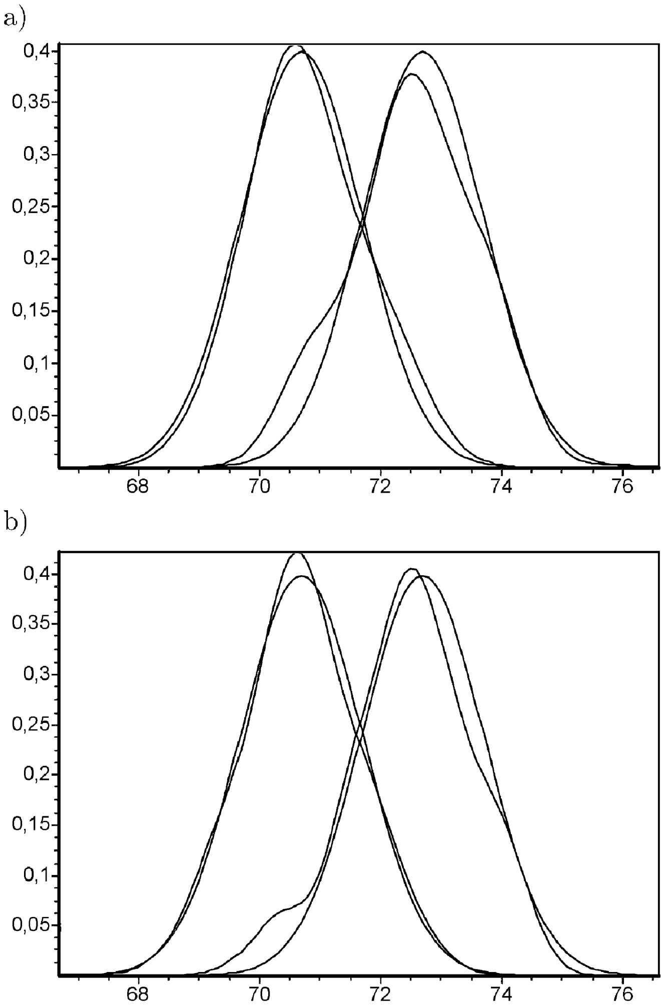

Fig. 11. Time-varying discriminant functions for $t = 0.5$, (a) $a = 0.3$, $H = 0.2$, $k = 3$; (b) $a = 0.4$, $H = 0.3$, $k = 5$.

B. Classification

Consider a two-category classification problem with $p_{1n} = p_{2n} = 1/2$ and

$$f_{1n}(x) = f_1(x - n^t), \quad f_{2n}(x) = f_2(x - n^t)$$

where

$$f_1(x) = N(0, 1), \quad f_2(x) = N(2, 1).$$

In this case the minimum probability of error is given by [9, p. 73]

$$P_e = \frac{1}{\sqrt{2\pi}} \int_1^\infty e^{-\frac{u^2}{2}} du = 0,159.$$

We will use procedures (58) and (62) for classification and show that empirical probability of missclassification approaches the above minimum probability. In Fig. 11 we present the results of simulations for $n = 5000$ (for each class) and $t = 0.5$. Two cases are considered: $a = 0.3$, $H = 0.2$, $k = 3$ and $a = 0.4$, $H = 0.3$, $k = 5$. We observe that the decision boundary is almost perfectly estimated. Tables VI and VII show empirical probability of missclassification approaching the minimum probability of error P_e as n grows. For each n (varying from

TABLE VI
MISCLASSIFICATION VERSUS n: $t = 0.5, k = 3, a = 0.3, H = 0.2$

n	error
1000	0.1766
2000	0.1820
3000	0.1666
4000	0.1791
5000	0.1706
6000	0.1771
7000	0.1670
8000	0.1645
9000	0.1640
10000	0.1645

TABLE VII
MISCLASSIFICATION VERSUS n: $t = 0.5, k = 3, a = 0.4, H = 0.3$

n	error
1000	0.176
2000	0.182
3000	0.166
4000	0.179
5000	0.170
6000	0.177
7000	0.167
8000	0.164
9000	0.164
10000	0.162

1000–10 000) we test the PNNs on 1000 samples from class 1 and 1000 from class 2.

XI. SUMMARY AND DISCUSSIONS

The presented pattern classification procedures are asymptotically optimal in the sense of Definitions 1 and 2. These properties are true with certain assumptions concerning densities in classes $f_{mn}, m = 1, \ldots, M, n = 1, 2, \ldots$.

In the stationary case, analogous properties are true with no assumptions concerning densities in classes [13], [26], but we should remember that the problems considered in this paper are much more difficult.

In Corollaries 3–6 concerning asymptotic optimality of rule (58), the type of nonstationarity was not specified, which enables us to use the obtained results for classifications of patterns characterized by various types of nonstationarity, but at the cost of the clarity of the respective conditions. These conditions, as it was shown in Tables I–IV, are clear and understandable in the case of the "movable argument"-type of the nonstationarity. Using Tables I–IV, it is possible to design a system realizing the pattern classification algorithm (58), i.e., properly select sequences a_n, h_n or $q(n)$, when nonstationary densities in classes are of the "movable argument" type, i.e.,

$$f_{mn}(x) = f_m(x - c_{mn}), \quad m = 1, \ldots, M, \quad n = 1, 2, \ldots.$$

For example, if $c_{mn} = n^{t_m}$ then neither the knowledge of function f_m nor the knowledge of parameters t_m is necessary in order to design algorithm (58). In spite of this, our algorithm will possess asymptotically-optimal properties in the sense of Definitions 1 and 2.

The comparison of algorithms based on the Parzen kernel with algorithms constructed on the basis of the orthogonal series method, carried out in Section IX, was undoubtedly more favorable to the former. Their application requires weaker assumptions concerning smooth properties of the density function and they allow the tracking of more significant changes of these functions (Table V). Finally, we note that combining the results of several classifiers may lead to improved performance of classification. Therefore, it would be interesting to investigate other soft computing techniques, e.g., SVM [42] or fuzzy methods [31], to classify patterns in a time-varying environment.

APPENDIX

Proof of Theorems 1 and 2

Of course,

$$\int \left(\widehat{R}_n(x) - R_n(x)\right)^2 dx \le 2 \int \left(\widehat{R}_n(x) - r_n(x)\right)^2 dx + 2 \int_A (r_n(x) - R_n(x))^2 dx. \tag{124}$$

Using argumentation similar to that in [30], we obtain

$$\begin{aligned} E \int_A & \left[\left(\widehat{R}_{n+1}(x) - r_{n+1}(x)\right)^2 | X_1, Y_1, \ldots, X_n, Y_n\right] \\ &\le (1 - a_{n+1}(1 - c_1)) \int_A \left(\widehat{R}_n(x) - r_n(x)\right)^2 dx \\ &+ a_{n+1}^2 \int \operatorname{var}\left[Y_{n+1} K_{n+1}(x, X_{n+1})\right] dx \\ &+ c_2 a_{n+1}^{-1} \int (r_{n+1}(x) - R_{n+1}(x))^2 dx \\ &+ c_3 a_{n+1}^{-1} \int (R_{n+1}(x) - R_n(x))^2 dx \\ &+ c_4 a_{n+1}^{-1} \int (r_n(x) - R_n(x))^2 dx. \end{aligned} \tag{125}$$

Application of the appropriate lemma in [3] to (125) concludes the proof. ■

Proof of Theorem 3

The theorem is a consequence of the application of Chung's [5] lemma (for $0 < a < 1$) or Watanabe's [44] (for $0 < a \le 1$) to expression (125). ■

Proof of Theorem 4

Slightly modifying the proof of theorem in work [47], we obtain

$$\begin{aligned} 0 &\le L_{n+k}(\widehat{\varphi}_n) - L_{n+k}\left(\varphi_{n+k}^*\right) \\ &\le \sum_{m=1}^{M} \int_{\chi_n} \left|\widehat{d}_{mn}(x) - d_{m,n+k}(x)\right| dx \\ &+ \sum_{m=1}^{M} \int_{\chi_n} d_{m,n+k}(x) dx. \end{aligned} \tag{126}$$

Under Schwartz's inequality

$$E\sum_{m=1}^{M}\int_{\chi_n}\left|\widehat{d}_{mn}(x)-d_{m,n+k}(x)\right|dx$$
$$\leq \mu^{\frac{1}{2}}(\chi_n)\sum_{m=1}^{M}\left[\int E\left(\widehat{d}_{mn}(x)-d_{m,n+k}(x)\right)^2\right]^{\frac{1}{2}}.$$

Thus, from the inequality

$$\sum_{m=1}^{M}\int_{\chi_n} d_{m,n+k}(x)dx \leq \frac{\varepsilon}{2}$$

follows the first part of the theorem.

The second part can be proved in a similar way. ■

Proof of Corollary 1

From the obvious inequality

$$I_{n,k} \leq 2I_n + 2\int\left(R_{n+k}(x)-R_n(x)\right)^2 dx \tag{127}$$

it follows that in order to ensure convergence of procedure (30) for the prediction problem, the conditions of theorems presented in Section IV should be supplemented with:

$$\int\left(R_{n+k}(x)-R_n(x)\right)^2 dx \xrightarrow{n} 0. \tag{128}$$

It is easily seen that for $k = 1$, condition (128) is implied by assumption (35).

For $k \geq 2$, the following holds

$$\begin{aligned}&|R_{n+k}(x)-R_n(x)|\\ &\quad\leq |R_{n+k}(x)-R_{n+k-1}(x)|\\ &\qquad+|R_{n+k-1}(x)-R_{n+k-2}(x)|\\ &\qquad+\ldots+|R_{n+1}(x)-R_n(x)|\,.\end{aligned} \tag{129}$$

Moreover, applying many times inequality $(a+b)^2 \leq 2a^2+2b^2$, we obtain

$$\begin{aligned}&\int\left(R_{n+k}(x)-R_n(x)\right)^2 dx\\ &\leq c_1\int\left(R_{n+k}(x)-R_{n+k-1}(x)\right)^2 dx\\ &\quad+c_2\int\left(R_{n+k-1}(x)-R_{n+k-2}(x)\right)^2 dx\\ &\quad+\ldots+c_k\int\left(R_{n+1}(x)-R_n(x)\right)^2 dx.\end{aligned} \tag{130}$$

It means that for $k \geq 2$, condition (128) is implied by assumption (35) which concludes the proof. ■

Proof of Corollary 2

The corollary follows from Theorem 3 and inequality (127).■

Proof of Theorem 5

From inequality (126) it follows that:

$$\begin{aligned}0 &\leq EL_{n+k}(\widehat{\varphi}_n)-L_{n+k}\left(\varphi^*_{n+k}\right)\\ &\leq \mu^{\frac{1}{2}}(\chi_n)\sum_{m=1}^{M}\left[\int E\left(\widehat{d}_{mn}(x)-d_{m,n+k}(x)\right)^2\right]^{\frac{1}{2}}\\ &\quad+\sum_{m=1}^{M}\int_{\chi_n} f_{m,n+k}(x)dx.\end{aligned}$$

Let χ_n be a p-dimensional cube with the middle in point zero and let the length of its side be e_n. Then, for $s > 0$,

$$\int_{\chi_n} f_{m,n+k}(x)dx \leq e_n^{-sp}t_{m,n+k}.$$

Consequently,

$$\begin{aligned}0 &\leq EL_{n+k}(\widehat{\varphi}_n)-L_{n+k}\left(\varphi^*_{n+k}\right)\\ &\leq \text{const. } e_n^{\frac{p}{2}}u_n^{\frac{1}{2}}+e_n^{-sp}t_{n+k}.\end{aligned}$$

Choosing $e_n = 0(u_n^{-1/(2s+1)p})$, we obtain the conclusion of Theorem 5. Let us point out that if $f_{mn} = f_m(x-c_{mn})$, then as χ_n we may choose a p-dimensional cube with the middle in c_{mn}. ■

Proof of Corollaries 3 and 4

Let us point out that

$$ET_{mn}K_n(x,X_n) = p_{mn}\int K_n(x,u)f_{mn}(u)du$$

Hence

$$\begin{aligned}&\int\left(ET_{mn}K_n(x,X_n)-d_{mn}(x)\right)^2 dx\\ &=p_{mn}^2\int\left(\int K(u)(f_{mn}(x-uh_n)-f_{mn}(x))du\right)^2 dx.\end{aligned} \tag{131}$$

Assuming that function K is of the type (83) and conditions (84)–(87) are satisfied, we obtain

$$\int\left(ET_{mn}K_n(x,X_n)-d_{mn}(x)\right)^2 dx \leq \text{const.}p_{mn}^2h_n^{2r}\delta_{mn}^{\underline{i}}(x).$$

Moreover,

$$\begin{aligned}&\int \text{var}\left[T_{mn}K_n(x,X_n)\right]dx\\ &\quad\leq h_n^{-p}\int\int K^2(z)f_n(x-zh_n)dxdz\\ &\quad\leq h_n^{-p}\int K^2(z)dz\end{aligned}$$

and

$$\begin{aligned}&\int\left(d_{m,n+1}(x)-d_{mn}(x)\right)^2 dx\\ &\quad\leq 2(p_{m,n+1}-p_{mn})^2\int f_{m,n}^2(x)dx\\ &\qquad+2p_{m,n+1}^2\int\left(f_{m.n+1}(x)-f_{mn}(x)\right)^2 dx.\end{aligned}$$

Now, Corollaries 3 and 4 are a direct consequence of Theorems 1 and 2. ■

Proof of Corollaries 5 and 6

Taking into consideration the following facts:

$$\mathrm{var}\left[T_{mn}K_n(x,X_n)\right] \leq \left(\sum_{j=0}^{q(n)} G_j^2\right)^p ,$$

$$\int \left(ET_{mn}K_n(x,X_n) - d_{mn}(x)\right)^2 dx = p_{mn}^2 S_{mn}.$$

we proceed in a similar manner like in the previous proof. ■

REFERENCES

[1] A. E. Albert and L. A. Gardner, *Stochastic Approximation and Nonlinear Regression*. Cambridge, MA: MIT Press, 1967.

[2] J. S. Bendat and A. G. Piersol, *Random Data Analysis and Measurement Procedures*. New York: Wiley, 1971.

[3] E. M. Braverman and L. I. Rozonoer, "Convergence of random processes in machine learning theory," *Autom. Remote Control*, vol. 30, pp. 44–64, 1969.

[4] P. Burrascano, "Learning vector quantization for the probabilistic neural network," *IEEE Trans. Neural Networks*, pp. 458–461, Mar. 1991.

[5] K. L. Chung, "On a stochastic approximation method," *Ann. Math. Statist.*, vol. 25, pp. 463–483, 1954.

[6] R. J. P. de Figueiredo, "Convergent algorithms for pattern recognition in nonlinearly evolving nonstationary environment," *Proc. IEEE*, vol. 56, pp. 188–189, 1968.

[7] L. P. Devroye, "Universal Consistency in Nonparametric Regression and Nonparametric Discrimination," School Comp. Science, McGill Univ, 1978.

[8] L. P. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. New York: Wiley, 1983.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[10] S. Efromovich, *Nonparametric Curve Estimation. Methods, Theory and Applications*. New York: Springer-Verlag, 1999.

[11] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, 1988.

[12] W. Greblicki, "Asymptotically optimal pattern recognition procedures with density estimates," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 250–251, 1978.

[13] W. Greblicki and L. Rutkowski, "Density-free Bayes risk consistency of nonparametric pattern recognition procedures," *Proc. IEEE*, vol. 69, pp. 482–483, Apr. 1981.

[14] S. Haykin and T. K. Bhattacharya, "Modular learning strategy for signal detection in a nonstationary environment," *IEEE Trans. Signal Proceesing*, vol. 45, pp. 1619–1637, June 1997.

[15] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag, 2002.

[16] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidh selection for density estimation," *J. Amer. Statist. Assoc.*, vol. 91, pp. 401–407, 1996.

[17] S. C. Kenyon, "Hyperspace organization for classification of nonstationary patterns," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, vol. 1, 1991, pp. 185–190.

[18] C. Kramer, B. Mckay, and J. Belina, "Probabilistic neural network array architecture for ECG classification," in *Proc. Annu. Int. Conf. IEEE Engineering Medicine Biology*, 17, 1995, pp. 807–808.

[19] K. Z. Mao, K.-C. Tan, and W. Ser, "Probabilistic neural-network structure determination for pattern classification," *IEEE Trans. Neural Networks*, vol. 11, pp. 1009–1016, July 2000.

[20] M. T. Musavi, K. H. Chan, D. M. Hummels, and K. Kalantri, "On the generalization ability of neural-network classifier," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 659–663, June 1994.

[21] A. Pagan and A. Ullah, *Nonparametric Econometrics*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[22] P. P. Raghu and B. Yegnanarayana, "Supervised texture classification using a probabilistic neural network and constraint satisfaction model," *IEEE Trans. Neural Networks*, vol. 9, pp. 516–522, May 1998.

[23] P. V. Rao and J. I. Thornby, "A robust point estimate in a generalized regression model," *Ann. Math. Statist.*, vol. 40, pp. 1784–1790, 1969.

[24] G. Roberts, A. M. Zoubir, and B. Boashash, "Time-frequency and time-scale analysis," in *Proc IEEE-SP Int. Symp.*, 1996, pp. 245–248.

[25] R. D. Romero, D. S. Touretzky, and G. H. Thibadeau, "Optical Chinese character recognition using probabilistic neural netwoks," *Pattern Recognit.*, vol. 3, pp. 1279–1292, 1997.

[26] L. Rutkowski, "Sequential estimates of probability densities by orthogonal series and their application in pattern classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, pp. 918–920, Dec. 1980.

[27] ——, Sequential estimates of a regression function by orthogonal series with applications in discrimination, in Lectures Notes in Statistics, New York-Heidelberg-Berlin, vol. 8, pp. 236–244, 1981.

[28] ——, "On Bayes risk consistent pattern recognition procedures in a quasistationary environment," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 84–87, Jan. 1982.

[29] ——, "Sequential pattern recognition procedures derived from multiple Fourier series," *Pattern Recognit. Lett.*, vol. 8, pp. 213–216, 1988.

[30] ——, "Nonparametric learning algorithms in the time-varying environments," *Signal Process.*, vol. 18, pp. 129–137, 1989.

[31] L. Rutkowski and K. Cpałka, "Flexible neuro-fuzzy systems," *IEEE Trans. Neural Networks*, vol. 14, pp. 554–574, May 2003.

[32] L. Rutkowski, *New Soft Computing Techniques for System Modeling, Pattern Classification and Image Processing*. New York: Springer-Verlag, 2004.

[33] D. F. Specht, "Probabilistic neural networks and the polynomial adaline as complementary techniques for classification," *IEEE Trans. Neural Networks*, pp. 111–121, Jan. 1990.

[34] ——, "Probabilistic neural networks," *Neural Netw.*, vol. 3, pp. 109–118, 1990.

[35] ——, "A general regression neural network," *IEEE Trans. Neural Networks*, pp. 568–576, Mar. 1991.

[36] ——, "Enhancements to the probabilistic neural networks," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Baltimore, MD, 1992, pp. 761–768.

[37] R. L. Streit and T. E. Luginbuhl, "Maximum likelihood training of probabilistic neural network," *IEEE Trans. Neural Networks*, vol. 5, pp. 764–783, Sept. 1994.

[38] J. R. Thompson and R. A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*. Philadelphia, PA: SIAM, 1990.

[39] H. G. C. Traven, "A neural-network approach to statistical pattern classification by semiparametric estimation of a probability density functions," *IEEE Trans. Neural Networks*, vol. 2, pp. 366–377, May 1991.

[40] J. Z. Tzypkin, *Introduction to the Self-Learning Systems Theory*. Moscow: Nauka Publishers, 1970.

[41] I. Vajda, L. Györfi, and Z. Györfi, "A strong law of large numbers and some applications," *Studia Scient. Math. Hung.*, vol. 12, pp. 233–244, 1977.

[42] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[43] G. G. Walter, "Properties of Hermite series estimation of probability density," *Ann. Statist.*, vol. 5, pp. 1258–1264, 1977.

[44] W. Watanabe, "On convergence of asymptotically optimal discriminant functions for pattern classification problems," *Bull. Math. Statist.*, vol. 16, pp. 23–34, 1974.

[45] A. R. Webb, *Statistical Pattern Recognition*, New York: Wiley, 2002.

[46] W. Wertz, *Statistical Density Estimation: A Survey*, Germany: Vandenhoeck & Ruprecht, 1978.

[47] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal discriminant functions for pattern classification," *IEEE Trans. Inform. Theory*, vol. 15, pp. 258–265, 1969.

[48] A. Zaknich, "A vector quantization reduction method for the probabilistic neural network," in *Proc. IEEE Int. Conf. Neural Networks*, Piscataway, NJ, 1997, pp. 1117–1120.

Leszek Rutkowski (M'92–SM'94) was born in Wrocław, Poland, in 1952. He received the M.Sc., Ph.D., and D.Sc. degrees in 1977, 1980, and 1986, respectively, all from the Technical University of Wrocław.

Since 1980, he has been with the Technical University of Częstochowa, where he is currently a Professor and Chairman of the Computer Engineering Department. From 1987 to 1990, he held a visiting position with the School of Electrical and Computer Engineering, Oklahoma State University. His research interests include neural networks, fuzzy systems, computational intelligence, pattern recognition, and systems identification. He published more than 100 technical papers including 16 in various series IEEE TRANSACTIONS. He is the author of the books *New Soft Computing Techniques For System Modeling, Pattern Classification, and Image Processing* (New York: Springer), *Flexible Neuro-Fuzzy Systems* (Norwell, MA: Kluwer Academic Publishers), and *Adaptive Filters and Adaptive Signal Processing* (in Polish), and is the coauthor of two others (in Polish) *Neural Networks, Genetic Algorithms and Fuzzy Systems* and *Neural Networks for Image Compression*.

Dr. Rutkowski is also President and Founder of the Polish Neural Networks Society. He organized and served as General Chair of the Polish Neural Networks Conferences held in 1996, 1997, 1999, 2000, and 2002. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS.