Multistage classification of signals with the use of multiscale wavelet representation

Urszula Libal
Institute of Computer Engineering, Control and Robotics,
Wroclaw University of Technology,
50-370 Wroclaw, Poland,
urszula.libal@pwr.wroc.pl

***Abstract*—The aim of signal decomposition in wavelet bases is to represent a signal as a sequence of wavelet coefficients sets. There is proposed a multistage classification rule using on every stage only one set of the signal coefficients. The hierarchical construction of wavelet multiresolution analysis was an inspiration for the multistage classification rule. The algorithm makes an optimal decision for every set of coefficients and its main advantage is a smaller dimension of classification problem on every stage.**

***Index Terms*—multistage classification, wavelet decomposition, multiscale representation**

I. Introduction

The aim of this article is to present the multistage classification rule for signals decomposed in wavelet bases. The compactness of wavelet functions support for Daubechies wavelets is used to reduce the amount of wavelet coefficients in signal approximation to a finite number. The classification rule makes a decision basing on one set of wavelet coefficients on every stage, corresponding to the actual scale of decomposition of the signal. The algorithm chooses on every stage a subset of classes (*macro-class*), where the signal came from. The rest of classes is rejected and would not take part in the next step of classification. The algorithm can make mistakes and classifies a signal not to its original class, but to one of the other classes. We measure the error with the risk value.

II. Multiscale wavelet signal representation

According to the *multiresolution analysis*, the Hilbert space of square integrable signals $L^2(\mathbb{R})$ (signals with finite energy) can be presented as an infinite simple sum of approximation subspace V_M and detail subspaces $W_m,\ m = M,\ M+1,\ \ldots,$ i.e.

$$L^2(\mathbb{R}) = V_M \bigoplus_{m=M}^{\infty} W_m. \tag{1}$$

The approximation of signal $s(t;\, K) \in V_K \subset L^2(\mathbb{R})$ according to the multiresolution analysis can be represented in

$$V_K = V_M \bigoplus_{m=M}^{K-1} W_m \tag{2}$$

as a sum of representations in subspaces V_M and $W_m,\ m = M,\ M+1,\ \ldots,\ K-1$.

Let $\phi(t)$ be a *scaling function* and $\psi(t)$ *mother wavelet*. Both functions should be square integrable, have compact support and fulfill additional conditions

$$\int_{\mathbb{R}} \phi(t)\, dt = 1 \tag{3}$$

and

$$\int_{\mathbb{R}} \psi(t)\, dt = 0. \tag{4}$$

The subspaces V_m and W_m have orthonormal bases of scaled and translated in time wavelet functions. We can then denote $V_m = span\{\phi_{mn}(t), n \in \mathbb{Z}\}$ and $W_m = span\{\psi_{mn}(t), n \in \mathbb{Z}\}$, where $\phi_{mn}(t) = 2^{m/2}\phi(2^m t - n)$, $\psi_{mn}(t) = 2^{m/2}\psi(2^m t - n)$. The notation of a signal approximation in subspace V_K can be presented as follows

$$\begin{aligned} s(t;\, K) = & \underbrace{\sum_{n_1\in\mathbb{Z}} \alpha_{Mn_1}\phi_{Mn_1}(t)}_{\in V_M} + \underbrace{\sum_{n_2\in\mathbb{Z}} \beta_{Mn_2}\psi_{Mn_2}(t)}_{\in W_M} \\ & + \underbrace{\sum_{n_3\in\mathbb{Z}} \beta_{M+1,n_3}\psi_{M+1,n_3}(t)}_{\in W_{M+1}} \\ & + \cdots + \underbrace{\sum_{n_N\in\mathbb{Z}} \beta_{K-1,n_N}\psi_{K-1,n_N}(t)}_{\in W_{K-1}}, \end{aligned} \tag{5}$$

where coefficients α_{mn} and β_{mn} are given by the formulas

$$\alpha_{mn} = \int_{\mathbb{R}} s(t)\,\phi_{mn}(t)\, dt \tag{6}$$

and

$$\beta_{mn} = \int_{\mathbb{R}} s(t)\,\psi_{mn}(t)\, dt. \tag{7}$$

A. Compact support of Daubechies wavelets

Daubechies wavelets [1] have special properties, especially useful in applications. We can find in [4] widely described attributes of Daubechies wavelets of order p:

1) Daubechies wavelet of order p has compact support with length $r = 2p - 1$. The length r is of course an odd number. *Scaling function* ϕ for Daubechies wavelet of

978-1-4244-7827-9/10/$26.00 ©2010 IEEE

order p will be denoted by D^p, while *mother-wavelet* ψ by d^p. The support of *scaling function* is

$$supp\{D^p\,(t)\} = [0,\, r] = [0,\, 2p-1]\,,$$

and the support of *mother-wavelet*

$$supp\{d^p\,(t)\} = \left[-\frac{r-1}{2},\, \frac{r+1}{2}\right] = [-p+1,\, p]\,.$$

2) The basis functions $\{D_{mn}\,(t)\}$ and $\{d_{mn}\,(t)\}$ consist on an orthonormal set for each scale m. Their support has length equal to $\frac{r}{2^m}$ and takes the following form

$$supp\{D_{mn}\,(t)\} = \left[\frac{n}{2^m},\, \frac{r+n}{2^m}\right],$$

$$supp\{d_{mn}\,(t)\} = \left[\frac{n}{2^m} - \frac{(r-1)}{2^{m+1}},\, \frac{n}{2^m} + \frac{(r+1)}{2^{m+1}}\right].$$

With the increase of scale (and resolution), the basis wavelets become narrower due to the shortening of the support and higher due to the increase in amplitude.

B. Model decomposed by Daubechies wavelets

Considering properties of Daubechies wavelets, and especially the compactness of their support, we can transform an approximation of signal given by (5) into the form where a finite number of elements will be summed

$$\begin{aligned} s\,(t;\, K) = & \sum_{n=n_{min}(D^p,a,M)}^{n_{max}(D^p,b,M)} \alpha_{Mn} D^p_{Mn}\,(t) \\ & + \sum_{m=M}^{K-1} \sum_{n=n_{min}(d^p,a,m)}^{n_{max}(d^p,b,m)} \beta_{mn} d^p_{mn}\,(t)\,. \end{aligned} \tag{8}$$

The limits in sums depends on the time interval $[a,\, b]$ of signal approximation

$$\begin{aligned} n_{min}\,(D^p,\, a,\, M) &= \lceil 2^M a - r \rceil, \\ n_{max}\,(D^p,\, b,\, M) &= \lfloor 2^M b \rfloor, \\ n_{min}\,(d^p,\, a,\, m) &= \lceil 2^m a - \frac{r-1}{2} \rceil, \\ n_{max}\,(d^p,\, b,\, m) &= \lfloor 2^m b + \frac{r+1}{2} \rfloor, \end{aligned} \tag{9}$$

and coefficients α_{Mn} and β_{mn} are given now by equations

$$\alpha_{Mn} = \int_{n/2^M}^{(r+n)/2^M} s\,(t)\, D^p_{Mn}\,(t)\, dt, \tag{10}$$

$$\beta_{mn} = \int_{\left(n-\frac{(r-1)}{2}\right)/2^m}^{\left(n+\frac{(r+1)}{2}\right)/2^m} s\,(t)\, d^p_{mn}\,(t)\, dt. \tag{11}$$

Such a model has a finite number of coefficients $\{\alpha_{Mn},\, \beta_{mn},\; m = M,\, M+1,\, \ldots,\, K-1\}$, which clearly represent a signal in the approximation subspace V_K with a fixed scale K. This observation is widely used by computational algorithms with wavelet signal representation, and it will be applied to construct a multistage classification rule (recognition algorithm) for signals.

III. Multistage classification

A. Problem statement

Let there be n disjoint classes $i \in \{1,\; 2,\; \ldots,\; n\} = \mathcal{M}$. There should be also known complete information about the *a priori* probability of occurrence p_i and the probability density function $f_i(\underline{x})$ of random features $\underline{x}$, for each class i:

class:	*1*	*2*	...	*n*
	p_1	p_2	...	p_n
	$f_1(\underline{x})$	$f_2(\underline{x})$	...	$f_n(\underline{x})$

The problem is to find a rule classifying an analyzed signal to one of the classes basing on a set of features $\underline{x}$. In statistical classification problem, the true, original class j of the signal is realization of a discrete random variable $\mathbb{J}$, while the vector of features $\underline{x}$ is the realization of a continuous random variable $\mathbb{X}$. *A priori* probability of occurrence of the classes $j \in \mathcal{M}$ is a positive probability of coming the signal from the class j before collecting of any features $\underline{x} \in \mathcal{X}$, i.e. before the experiment:

$$p_j = \mathbb{P}\,(\mathbb{J} = j) > 0. \tag{12}$$

Further, we assume that the signal can be characterized by a vector of N numerical features $\underline{x} \in \mathcal{X} \subseteq \mathbb{R}^N$, and the probability density of the features $\underline{x}$ in classes $j \in \mathcal{M}$ does exist and is known. This is a conditional probability density $f_j\,(\underline{x}) \overset{df}{=} f\,(\underline{x} \mid j)$ in class $j \in \mathcal{M}$, i.e. the probability density of features $\underline{x}$ under the assumption that the signal comes exactly from this class.

B. Classification tree

The classification rule characterizes the gradual use of the individual components $x^{(m)}$ of signal representation $\underline{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(N)})^T$ in successive stages $m = 1, 2, \ldots, N$, (decomposition of the signal) and joining the terminal classes in the transitional *macro-classes* (aggregation of classes). This leads to a reduction of the dimension of signal representation and reduces the number of classes on particular stages. *Macro-classes* are the subsets of the classes identified with the internal nodes of the decision tree structure established in advance. Each *macro-class* is a collection of terminal classes accessible from the node associated with it. The set of terminal classes for each *macro-class* is divided into disjoint subsets that define its child *macro-classes* (direct successors). In Figure 1 we present a tree for two-stage pattern recognition algorithm and four classes.

We will use the following denotations

$\mathcal{M}^{\boldsymbol{i}}$ - a collection of *macro-classes* that are direct predecessors of *macro-class* (node) $\boldsymbol{i}$; ($\mathcal{M}^{\boldsymbol{i_0}}$ means the set of *macro-classes* that are direct predecessors of root $\mathcal{M} = \{1,\, 2,\, \ldots,\, n\}$);

$\mathcal{M}^{\boldsymbol{i_{m-1}}}$ - a collection of *macro-classes* that are direct predecessors of node $\boldsymbol{i_{m-1}}$ (indicated by the algorithm Ψ_{m-1} for stage $m = 2,\, \ldots,\, N$).

And then

$\mathcal{M}_{\boldsymbol{i}}$ - a set of classes (terminal nodes) achievable from *macro-class* (node) $\boldsymbol{i}$; ($\mathcal{M}_{\boldsymbol{i_0}}$ means the set of all classes, i.e. the set of all terminal nodes, $\mathcal{M}_{\boldsymbol{i_0}} = \mathcal{M}$);

$\mathcal{M}_{\boldsymbol{i_{m-1}}}$ - a set of classes (terminal nodes) accessible from the node $\boldsymbol{i_{m-1}}$ (indicated by the algorithm Ψ_{m-1} for stage $m = 2, \ldots, N$).

The conditional *a priori* probability for the specified schema of the decision tree, (the *a priori* probability of *macro-class* $\boldsymbol{j_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}$ appearance on the stage m, provided that in the previous stage $m-1$ *macro-class* $\boldsymbol{i_{m-1}} \in \mathcal{M}^{\boldsymbol{i_{m-2}}}$ was accepted), can be determined from the Bayes' formula of conditional probability as follows

$$p^{(m)}_{\boldsymbol{j_m}} = \frac{\mathbb{P}(\mathbb{J} \in \mathcal{M}_{\boldsymbol{j_m}} \,|\, \mathcal{M}_{\boldsymbol{i_{m-1}}})}{\mathbb{P}(\mathcal{M}_{\boldsymbol{i_{m-1}}})} = \frac{\sum\limits_{j \in \mathcal{M}_{\boldsymbol{j_m}}} p_j}{\sum\limits_{i \in \mathcal{M}_{\boldsymbol{i_{m-1}}}} p_i}, \tag{13}$$

where p_j is an *a priori* probability of class $j \in \mathcal{M}$ appearance for one-stage classification rule Ψ_m (see (21)). On the condition that there has happened $j \in \boldsymbol{i_{m-1}} \in \mathcal{M}^{\boldsymbol{i_{m-2}}}$, the *a priori* conditional probability of class $j \in \mathcal{M}$ appearance on the stage m is

$$p^{(m)}_j = \frac{p_j}{\sum\limits_{i \in \mathcal{M}_{\boldsymbol{i_{m-1}}}} p_i}. \tag{14}$$

It is important that choosing a node $\boldsymbol{i_{m-1}}$ by algorithm Ψ_{m-1} the other branches of decision tree become inactive (compare with another approach in [5]). There are cut off the paths leading to these terminal classes $j \in \mathcal{M}$ that are not accessible from the node $\boldsymbol{i_{m-1}}$ (i.e. $j \notin \mathcal{M}_{\boldsymbol{i_{m-1}}}$). Then the conditional probability for the relevant classes has zero value

$$p^{(m)}_j = 0 \quad for\ j \in \mathcal{M} \backslash \mathcal{M}_{\boldsymbol{i_{m-1}}}. \tag{15}$$

In the construction of a multistage algorithm $\Psi^W = (\Psi_1, \Psi_2, \ldots, \Psi_N)$ we use the conditional probability density function of features $\underline{x} \in \mathcal{X}$ from *macro-class* $\boldsymbol{j_m}$ in the following form

$$f_{\boldsymbol{j_m}}(\underline{x}) \overset{df}{=} f(\underline{x} \,|\, \boldsymbol{j_m}) = \frac{1}{p^{(m)}_{\boldsymbol{j_m}}} \sum_{j \in \mathcal{M}_{\boldsymbol{j_m}}} p^{(m)}_j f_j(\underline{x}). \tag{16}$$

The boundary distribution of feature $x^{(m)} \in \mathcal{X}^{(m)}$, used for classification on m^{th} stage, is given by the following probability density function

$$f^{(m)}(x^{(m)}) = \int\limits_{\mathcal{X}^{(m),C}} f(\underline{x})\, d(x^{(1)} \ldots x^{(m-1)} x^{(m+1)} \ldots x^{(N)}), \tag{17}$$

where

$$\mathcal{X}^{(m),C} = \mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(m-1)} \times \mathcal{X}^{(m+1)} \times \ldots \times \mathcal{X}^{(N)},$$

and the probability density function is

$$f(\underline{x}) = \sum_{j \in \mathcal{M}} p_j\, f_j(\underline{x}), \quad \underline{x} \in \mathcal{X}. \tag{18}$$

The conditional probability density function for feature $x^{(m)} \in \mathcal{X}^{(m)}$ (coming from the *macro-class* $\boldsymbol{j_m}$) on m^{th} stage of classification is then equal to the corresponding mixture of boundary density functions. Based on the formulas (16) and (17) we obtain

$$f^{(m)}_{\boldsymbol{j_m}}(x^{(m)}) \overset{df}{=} f^{(m)}(x^{(m)} \,|\, \boldsymbol{j_m}) = \frac{1}{p^{(m)}_{\boldsymbol{j_m}}} \sum_{j \in \mathcal{M}_{\boldsymbol{j_m}}} p^{(m)}_j f^{(m)}(x^{(m)} \,|\, j). \tag{19}$$

C. Classification rule

Definition 1: [Multistage classification rule] The N-stage classification rule (algorithm)

$$\Psi^W : \quad \mathcal{X} \longrightarrow \mathcal{M}^{\boldsymbol{i_0}} \times \mathcal{M}^{\boldsymbol{i_1}} \times \cdots \times \mathcal{M}^{\boldsymbol{i_N}}$$

classifies signal basing on the features $\underline{x} \in \mathcal{X}$ to the terminal class i_N in the following way

$$\Psi^W(\underline{x}) = (\Psi_1(x^{(1)}), \Psi_2(x^{(2)}), \ldots, \Psi_N(x^{(N)})) = (\boldsymbol{i_1}, \boldsymbol{i_2}, \ldots, i_N). \tag{20}$$

This classification rule is a sequence of N one-stage algorithms $(\Psi_1, \Psi_2, \ldots, \Psi_N)$ such as

$$\Psi_m : \quad \mathcal{X}^{(m)} \longrightarrow \mathcal{M}^{\boldsymbol{i_{m-1}}}.$$

The one-stage classification rule definition is as follows

Definition 2: [One-stage classification rule] For zero-one loss functions (24) algorithm Ψ_m classifies the signal to *macro-class* $\boldsymbol{i_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}$:

$$\Psi_m(x^{(m)}) = \boldsymbol{i_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}, \tag{21}$$

when

$$p^{(m)}_{\boldsymbol{i_m}}\, f^{(m)}_{\boldsymbol{i_m}}(x^{(m)}) = \max_{\boldsymbol{k_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}} p^{(m)}_{\boldsymbol{k_m}}\, f_{\boldsymbol{k_m}}(x^{(m)}).$$

Such a construction of the multistage algorithm $\Psi^W = (\Psi_1, \Psi_2, \ldots, \Psi_N)$ we will call locally optimal because at every stage of the classification the one-stage algorithm Ψ_m is optimal and minimizes the risk. In addition, the usage of zero-one loss functions results in maximization of the *a posteriori* probability of correct classification of the signal to a *macro-class* achievable from the node $\boldsymbol{i_{m-1}}$ on m^{th} stage, for $m = 1, 2, \ldots, N$. The multistage classification rule $\Psi^W = (\Psi_1, \Psi_2, \ldots, \Psi_N)$ for decomposed by wavelets signal is following. The rule Ψ^W classifies signal to the one terminal class basing on the vector of features

$$\underline{x} = (\alpha_M, \beta_M, \ldots, \beta_{K-1})^T, \tag{22}$$

where vector components consist of wavelet coefficient sequences for signal decomposed in the wavelet basis

$$\begin{aligned} s\left(\omega,\, t;\, K\right) = & \underbrace{\sum_{n_1\in\mathbb{Z}} \alpha_{Mn_1}\left(\omega\right)\, D^p_{Mn_1}\left(t\right)}_{\in V_M} \qquad (23)\\ & + \underbrace{\sum_{n_2\in\mathbb{Z}} \beta_{Mn_2}\left(\omega\right)\, d^p_{Mn_2}\left(t\right)}_{\in W_M}\\ & + \underbrace{\sum_{n_3\in\mathbb{Z}} \beta_{M+1,n_3}\left(\omega\right)\, d^p_{M+1,n_3}\left(t\right)}_{\in W_{M+1}}\\ & + \cdots + \underbrace{\sum_{n_N\in\mathbb{Z}} \beta_{K-1,n_N}\left(\omega\right)\, d^p_{K-1,n_N}\left(t\right)}_{\in W_{K-1}}\,, \end{aligned}$$

In (23) we introduced ω (an elementary event) as an additional argument to emphasize random nature of components. We assume the usage of Daubechies wavelets to the signal decomposition for $N = K - M + 1$ components and scales $m = M, M + 1, \ldots, K - 1$. The components of vector $\underline{x}$ consist of

$$\begin{cases} x^{(1)} = \alpha_M \overset{df}{=} \{\alpha_{Mn_1}\left(\omega\right)\},\\ x^{(2)} = \beta_M \overset{df}{=} \{\beta_{Mn_2}\left(\omega\right)\},\\ \vdots\\ x^{(N)} = \beta_{K-1} \overset{df}{=} \{\beta_{K-1,n_N}\left(\omega\right)\} \end{cases}$$

where limits of indexes are defined by (9)

$$\begin{gathered} n_{min}(D^p,\, a,\, M) \leq n_1 \leq n_{max}(D^p,\, b,\, M),\\ n_{min}(d^p,\, a,\, M) \leq n_2 \leq n_{max}(d^p,\, b,\, M),\\ \vdots\\ n_{min}(d^p,\, a,\, K-1) \leq n_N \leq n_{max}(d^p,\, b,\, K-1). \end{gathered}$$

The numbers of wavelet coefficients (the length of wavelet coefficient sequences), counted for the following scales $M,\, M+1,\, \ldots,\, K-1$, are different. The number of stages Ψ^W=$(\Psi_1,\, \Psi_2,\, \ldots,\, \Psi_N)$ depends on the number of wavelet coefficient sequences, to which the signal will be decomposed.

Definition 3: [Multistage classification algorithm] The classification of signal goes in $N = K - M + 1$ stages. There is presented the implementation of *the classification rule* for signal representation $\underline{x} = (\alpha_M,\, \beta_M,\, \ldots,\, \beta_{K-1})^T$ obtained by wavelet decomposition of signal:

Stage 1: The rule Ψ_1 classifies analyzed signal representation $\underline{x}$ to macro-class (node) $\boldsymbol{i_1}$ from the set of the direct successors of root $\mathcal{M}^{\boldsymbol{i_0}}$ based on sequence $\alpha_M \in \mathbb{R}^{d_1}$:

$$\begin{gathered} \Psi_1 : \mathbb{R}^{d_1} \longrightarrow \mathcal{M}^{\boldsymbol{i_0}},\\ \Psi_1(\alpha_M) = \boldsymbol{i_1},\\ \alpha_M \in \mathbb{R}^{d_1},\, \boldsymbol{i_1} \in \mathcal{M}^{\boldsymbol{i_0}}. \end{gathered}$$

Stage 2: The rule Ψ_2 classifies analyzed signal representation $\underline{x}$ to macro-class (node) $\boldsymbol{i_2}$ from the set $\mathcal{M}^{\boldsymbol{i_1}}$ of the direct successors of the node $\boldsymbol{i_1}$ based on sequence $\beta_M \in \mathbb{R}^{d_2}$:

$$\begin{gathered} \Psi_2 : \mathbb{R}^{d_2} \longrightarrow \mathcal{M}^{\boldsymbol{i_1}},\\ \Psi_2(\beta_M) = \boldsymbol{i_2},\\ \beta_M \in \mathbb{R}^{d_2},\, \boldsymbol{i_2} \in \mathcal{M}^{\boldsymbol{i_1}}.\\ \vdots \end{gathered}$$

Stage m: The rule Ψ_m classifies analyzed signal representation $\underline{x}$ to macro-class (node) $\boldsymbol{i_m}$ from the set $\mathcal{M}^{\boldsymbol{i_{m-1}}}$ of the direct successors of the node $\boldsymbol{i_{m-1}}$ based on sequence $\beta_{M+m-2} \in \mathbb{R}^{d_m}$:

$$\begin{gathered} \Psi_m : \mathbb{R}^{d_m} \longrightarrow \mathcal{M}^{\boldsymbol{i_{m-1}}},\\ \Psi_m(\beta_{m+M-2}) = \boldsymbol{i_m},\\ \beta_{m+M-2} \in \mathbb{R}^{d_m},\, \boldsymbol{i_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}.\\ \vdots \end{gathered}$$

Stage N: The rule Ψ_N classifies analyzed signal representation $\underline{x}$ to a class (terminal node) i_N from the set $\mathcal{M}^{\boldsymbol{i_{N-1}}} \subset \mathcal{M}$ of the direct successors of the node $\boldsymbol{i_{N-1}}$ based on sequence $\beta_{K-1} \in \mathbb{R}^{d_N}$:

$$\begin{gathered} \Psi_N : \mathbb{R}^{d_N} \longrightarrow \mathcal{M}^{\boldsymbol{i_{N-1}}},\\ \Psi_N(\beta_{K-1}) = i_N,\\ \beta_{K-1} \in \mathbb{R}^{d_N},\, i_N \in \mathcal{M}^{\boldsymbol{i_{N-1}}}. \end{gathered}$$

The last stage of signal classification ends with choosing the terminal class, i.e. indicating the particular class $i = i_N$, to which finally the representation of signal $\underline{x}$ is assigned.

D. Classification accuracy

The measure of classification accuracy is a *risk* of false classification. Let us define the zero-one loss function for each stage m with following formula

$$L_m(\boldsymbol{i_m},\, \boldsymbol{j_m}) = \begin{cases} 0, & \text{when} \quad \boldsymbol{i_m} \cap \boldsymbol{j_m} \neq \emptyset,\\ 1, & \text{when} \quad \boldsymbol{i_m} \cap \boldsymbol{j_m} = \emptyset. \end{cases} \qquad (24)$$

It should be emphasized that with such a definition (24), the arguments of zero-loss function are the *macro-classes* $\boldsymbol{i_m}$ and $\boldsymbol{j_m}$, which are direct successors of node $\boldsymbol{i_{m-1}}$, i.e.

$$L_m : \quad \mathcal{M}^{\boldsymbol{i_{m-1}}} \times \mathcal{M}^{\boldsymbol{i_{m-1}}} \longrightarrow \{0,\, 1\}.$$

Theorem 1: [Risk of one-stage classification] The risk of classification rule Ψ_m in node $\boldsymbol{i_{m-1}}$ for zero-one loss function $L_m(\boldsymbol{i_m},\, \boldsymbol{j_m})$ is

$$\begin{aligned} R[\Psi_m] &= \mathbb{E}_{\mathbb{X}^{(m)},\mathbb{J}^{(m)}}\left[L_m\left(\Psi_m\left(\mathbb{X}^{(m)}\right),\mathbb{J}^{(m)}\right)\right] \qquad (25)\\ &= \sum_{\boldsymbol{j_m}\in\mathcal{M}^{\boldsymbol{i_{m-1}}}} p^{(m)}_{\boldsymbol{j_m}} \int_{\mathcal{X}^{(m)}\setminus D^{\boldsymbol{j_m}}_{x^{(m)}}} f^{(m)}_{\boldsymbol{j_m}}\left(x^{(m)}\right) dx^{(m)}, \end{aligned}$$

where for each *macro-class* $\boldsymbol{i_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}$ the decision area has form $D^{\boldsymbol{i_m}}_{x^{(m)}} = \{x^{(m)} \in \mathcal{X}^{(m)} : \quad \Psi_m(x^{(m)}) = \boldsymbol{i_m}\}$.

The risk for zero-one loss function has the lowest value. We assume for the multistage classification rule Ψ^W=$(\Psi_1,\, \Psi_2,\, \ldots,\, \Psi_N)$ the form of loss function

$$L^W: \quad \left(\mathcal{M}^{\boldsymbol{i_0}} \times \cdots \times \mathcal{M}^{\boldsymbol{i_{N-1}}}\right) \times \left(\mathcal{M}^{\boldsymbol{i_0}} \times \cdots \times \mathcal{M}^{\boldsymbol{i_{N-1}}}\right) \longrightarrow \{0,\, \frac{1}{N},\, \frac{2}{N},\, \ldots,\, 1\},$$

with the following formula

$$L^W((\boldsymbol{i_1},\, \ldots,\, i_N),\, (\boldsymbol{j_1},\, \ldots,\, j_N)) = \begin{cases} 0, & \text{when} \quad i_N = j_N, \\ \frac{N-m+1}{N}, & \text{when} \quad \boldsymbol{i_m} \cap \boldsymbol{j_m} = \emptyset \\ & \text{and } \boldsymbol{i_{m-1}} \cap \boldsymbol{j_{m-1}} \neq \emptyset. \end{cases} \tag{26}$$

We assume a greater loss if the error is committed at an earlier stage. It is important to construct the decision tree in right way, i.e. to aggregate similar classes in *macro-classes*. The construction of loss function (26) allows us to formulate the following corollary.

Corollary 1: The loss (26) for N-stage algorithm Ψ^W is equal to the average local loss (24) for the subsequent stages of classification, i.e.

$$L^W((\boldsymbol{i_1},\, \ldots,\, i_N),\, (\boldsymbol{j_1},\, \ldots,\, j_N)) = \frac{1}{N}\sum_{m=1}^{N} L_m(\boldsymbol{i_m},\, \boldsymbol{j_m}).$$

The next corollary is the natural consequence of previous fact.

Corollary 2: The risk of a N-stage algorithm Ψ^W is equal to the average local risk (25) for subsequent stages of classification, i.e.

$$R[\Psi^W] = \frac{1}{N}\sum_{m=1}^{N} R[\Psi_m]. \tag{27}$$

To sum up, we state the following theorem:

Theorem 2: [Multistage classification risk] For the loss function (26) the risk of a N-stage algorithm Ψ^W=$(\Psi_1,\, \Psi_2,\, \ldots,\, \Psi_N)$ is

$$R[\Psi^W] = \frac{1}{N}\sum_{m=1}^{N} \sum_{j \in \mathcal{M}_{\boldsymbol{i_{m-1}}}} p_j^{(m)} \int\limits_{D^{j,C}_{x^{(m)}}} f_j\,(\underline{x})\; d\underline{x} \tag{28}$$

$$= \sum_{j \in \mathcal{M}} \left(\frac{1}{N}\sum_{m=1}^{N} p_j^{(m)} \int\limits_{D^{j,C}_{x^{(m)}}} f_j\,(\underline{x})\; d\underline{x} \right), \tag{29}$$

where for each class $j \in \mathcal{M}$ the set $D^{j,C}_{x^{(m)}} = \mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(m-1)} \times [\mathcal{X}^{(m)} \backslash D^j_{x^{(m)}}] \times \mathcal{X}^{(m+1)} \times \ldots \times \mathcal{X}^{(N)}$ is a complement of decision area $D^j_{x^{(m)}}$ to the feature space $\mathcal{X}$. The decision areas

$$D^j_{x^{(m)}} = \{x^{(m)} \in \mathcal{X}^{(m)} : j \in \Psi_m(x^{(m)})\} = \{x^{(m)} \in \mathcal{X}^{(m)} : p^{(m)}_{\boldsymbol{i_m}} f^{(m)}_{\boldsymbol{i_m}}(x^{(m)}) = \max_{\boldsymbol{k_m} \in \mathcal{M}^{\boldsymbol{i_{m-1}}}} p^{(m)}_{\boldsymbol{k_m}} f_{\boldsymbol{k_m}}(x^{(m)})\}$$

are equal for each class $j \in \mathcal{M}_{\boldsymbol{i_m}}$ (achievable from a fixed *macro-class* $\boldsymbol{i_m}$ on the m^{th} stage).

IV. Multistage classification of ECG signal

The decomposed in wavelet bases ECG signals (electrocardiograms), from patients with previously diagnosed heart disease by cardiologists, were examined to the multistage classification algorithm. The main purpose of this experiment was an illustration of the multistage classification rule. Then the identified by the algorithm heart diseases were compared with the diagnoses of doctors what allowed us to evaluate the quality of the algorithm. The ECG signals $s(t,\, K)$ were decomposed into two sequences of wavelet coefficients, i.e. $\underline{x} = (\alpha_{K-1},\, \beta_{K-1})^T$. After decomposition, we reduce the number of wavelet coefficients ($d = 51$) by thresholding to the length $d = 2,\, 3,\, 4$ and 5. There was presented two-stage classification rule Ψ^W=$(\Psi_1,\, \Psi_2)$ and the accuracy for this rule was estimated.

First, based on the data of Eurostat [2] two *macro-classes* were chosen: $\{1,\, 2\}$ *ischemic heart diseases*, $\{3,\, 4\}$ *cardiac arrhythmias*, and there were calculated an *a priori* probabilities for them. ECG signals used in the experiment come from the database PhysioNet [3] and were divided into four classes, corresponding to a different heart diseases:

1) *European ST-T Database* [6] - *48* signals (*class 1*)
2) *MIT-BIH ST Change Database* - *28* signals (*class 2*)
3) *MIT-BIH Arrhythmia Database* - *48* signals (*class 3*)
4) *MIT-BIH Supraventricular Arrhythmia Database* - *78* signals (*class 4*)

Basing on the size of databases we calculated the relative frequencies of terminal classes (compare to (14))

1) $\hat{p}_1 = \frac{48}{202} = 0,238$ for class *1,*
2) $\hat{p}_2 = \frac{28}{202} = 0,139$ for class *2,*
3) $\hat{p}_3 = \frac{48}{202} = 0,238$ for class *3,*
4) $\hat{p}_4 = \frac{78}{202} = 0,386$ for class *4.*

We estimated the *a priori* probabilities of occurrence of *macro-classes* $\{1,\, 2\}$, $\{3,\, 4\}$ (compare to (13)):

- $\hat{p}^{(1)}_{\{1,2\}} = 0,38$ for ischemic heart disease,
- $\hat{p}^{(1)}_{\{3,4\}} = 0,62$ for cardiac arrhythmias.

Each *macro-class* consists of two classes. It is presented on scheme in Figure 1.

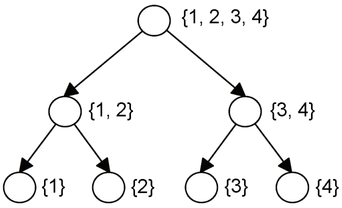

Figure 1. The binary decision tree diagram for the two-stage classification rule with 2 *macro-classes* and 4 terminal classes.

ECG signals in each class of heart disease were divided into two subsets: *learning* and *testing*. Using the electrocardiograms from the first group (a collection of 101 *learning signals*) we estimated the probability density function in each class. Then we examined the quality of the algorithm by comparing the results of classification algorithm for the rest electrocardiograms (101 *testing signals*, not involved in the learning algorithm) with diagnoses of cardiologists.

A. Learning of classification algorithm

The following two steps were taken:

1) The hypothesis that the wavelet coefficient sequences become from the normal probability distribution has been verified. It was rejected by the Shapiro-Wilk and Lilliefors tests at the 5% significance level. It is the cause why we estimated density functions with nonparametric kernel estimators.
2) We used two types of kernel estimators: a histogram and a Rosenblatt-Parzen estimator with the triangular kernel. The kernel functions: uniform and triangular have support with a finite length, so the usage of them reduces the computational complexity of the estimation. There were estimated the density functions for classes and *macro-classes* by analogy to (19). In Figure 2 we show the example of estimated density function.

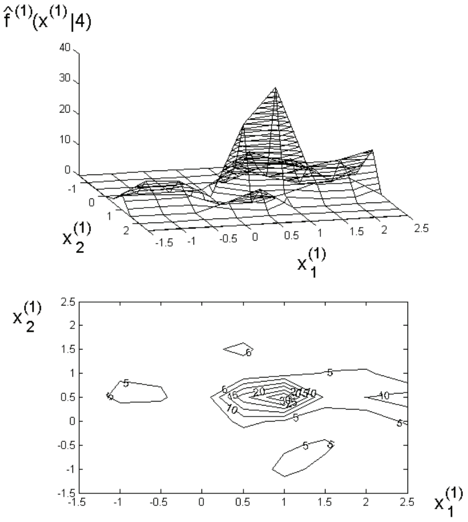

Figure 2. The estimation of density function $\widehat{f}^{(1)}\left(x^{(1)}\,|\,4\right)$ for class 4 calculated by the usage of triangular kernel estimator, basing on approximation coefficients $\{\alpha^1_{Mn}\}^d_{n=1}$, $\{\alpha^2_{Mn}\}^d_{n=1}, \ldots, \{\alpha^{101}_{Mn}\}^d_{n=1}$ thresholded on the level $\lambda_{h1} = 6.0$ (the length of sequence is $d = 2$).

B. Testing of classification algorithm

To estimate the risk of false diagnosis, we counted the differences between classification and medical diagnosis for all signals from *testing set*. Experimental risk was determined from the dependence

$$\widehat{R}[\Psi^{W,d}] = \frac{\sum_{i=1}^{n}(1 - \mathbf{1}_{j_N(i)}\{i_N(i)\})}{n}, \qquad (30)$$

where

n - the number of test ECG signals (here $n = 101$),

$j_N(i)$ - diagnosed by cardiologists class of heart disease for i^{th} signal from *testing set*, $i = 1, 2, \ldots, n$,

$i_N(i)$ - the selected terminal class as the result of classification algorithm $\Psi^{W,d}$,

d - the size of the wavelet coefficient sequence after thresholding (dimension of problem).

The results are presented in Figure 3. The lowest value of experimental risk was gained for the length of coefficient sequences $d = 3$.

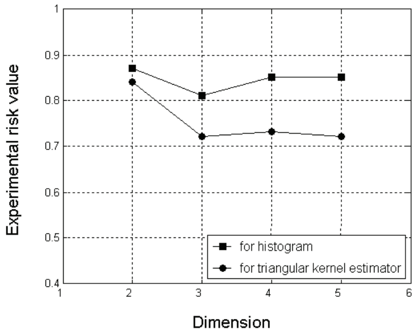

Figure 3. The experimental risk value for the multistage classification rule $\Psi^{W,d}$ depending on the length d of the coefficient sequences (dimension of problem).

V. Final remarks

Generally, in real applications we do not know the probability distributions for signal features in particular, individual classes. It was the reason for development of an empirical classification rule based on the *learning signals*, in addition to the theoretical classification algorithm. Frequency of correct classification has been established on the basis of *testing signals*, that consist of wavelet coefficients extracted from the ECG signals for patients who had the diagnosed heart disease. Experimental classification of ECG signals, pointed out some problems in the practical realization of classification, such as proper selection of features and the need for choosing an appropriate method for estimating probability distributions for features in classes.

The combination of multiscale representation of signal in wavelet bases with a multistage classification rule may lead to the high accuracy with simultaneous reducing the dimension of the problem (the amount of signal representation coefficients in approximation subspace V_m or detail subspace W_m) on each stage of classification.

References

[1] I. Daubechies, *Ten Lectures on Wavelets*, SIAM Edition, Philadelphia (1992).

[2] Eurostat, *Hospital discharges by diagnosis (ISHMT) and region, in-patients, per 100000 inhabitants*, [Data on Hospitalization of Patients Suffering from Heart Disease in Poland in 2004-2005, http://epp.eurostat.ec.europa.eu].

[3] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.* Circulation 101(23): e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215], (2000).

[4] Z. Hasiewicz, P. Śliwiński, *Orthogonal wavelets with compact support. Application to non-parametric identification systems (in Polish)*, Exit, Warsaw, (2005).

[5] M. Kurzyński, *Image recognition. Statistical methods (in Polish)*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (1997).

[6] A. Taddei, G. Distante, M. Emdin, P. Pisani, G.B. Moody, C. Zeelenberg, C. Marchesi. *The European ST-T Database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography.* European Heart Journal No. 13: 1164-1172 (1992).