Classifying Functional Data from Orthogonal Projections – model, properties and fast implementation

Ewa Skubalska-Rafajłowicz[1][0000−0002−2795−3835] and
Ewaryst Rafajłowicz[2][0000−0003−4347−1358]

[1] Dept. of Engineering Informatics, ewa.rafajlowicz@pwr.edu.pl
[2] Dept. of Control Systems and Mechatronics, ewaryst.rafajlowicz@pwr.edu.pl
Wroclaw University of Science and Technology, Wyb. Wyspianskiego 27, 50 370 Wrocław, Poland

Abstract. We consider the problem of functional, random data classification from equidistant samples. Such data are frequently not easy for classification when one has a large number of observations that bear low information for classification. We consider this problem using tools from the functional analysis. Therefore, a mathematical model of such data is proposed and its correctness is verified. Then, it is shown that any finite number of descriptors, obtained by orthogonal projections on any differentiable basis of $L_2(0,\, T)$, can be consistently estimated within this model.
Computational aspects of estimating descriptors, based on the fast implementation of the discrete cosine transform (DCT), are also investigated in conjunction with learning a classifier and using it on-line. Finally, the algorithm of learning descriptors and classifiers were tested on real-life random signals, namely, on accelerations, coming from large bucket-wheel excavators, that are transmitted to an operator's cabin. The aim of these tests was also to select a classifier that is well suited for working with DCT-based descriptors.

Keywords: functional data classification, random element, bias, functional data model, classifying signals, DCT

1 Introduction

Tasks of classifying functional data are difficult for many reasons. The majority of them seems to concern a large number of observations, frequently having an unexpectedly low information content from the point of view of their classification. This kind of difficulty arises in many industrial applications, in which sensors may provide thousands of samples per second (see, e.g., our motivation example at the end of this section). We focus our attention on classifying data from repetitive processes, i.e., on stochastic processes that have a finite and the same duration $T > 0$ and after time T the process, denoted as $\mathbf{X}(t)$, $t \in [0,\, T]$ is repeated with the same or different probability measures. For simplicity of the exposition, we confine ourselves to two such measures and the problem

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

is to classify samples from $\mathbf{X}(t)$, $[0, T]$ to two classes, having at our disposal a learning sequence $\mathbf{X}_n(t)$, $[0, T]$, $n = 1, 2, \ldots, N$ of correctly classified subsequences. An additional requirement is to classify newly incoming samples almost immediately after the present $[0, T]$ is finished, so as to be able to use the result of classification for making decisions for the next period (also called a pass). This requirement forces us to put emphasis not only on the theoretical but also on the computational aspects of the problem.

An outline of the approach and the paper organization. It is convenient to consider the whole $\mathbf{X}$ and $\mathbf{X}_n(t)$'s as random elements in a separable Hilbert space. We propose a framework (Section 2) that allows us to impose probability distributions on them in a convenient way, namely, by attaching them to a finite number of orthogonal projections, but the residuals of the projections definitely do not act as white noise, since the samples are highly correlated, even when they are far in time within $[0, T]$ interval. After proving the correctness of this approach (Section 2, Lemma 1) , we propose, in Section 3, the method of learning descriptors, which are projections of $\mathbf{X}$ and $\mathbf{X}_n(t)$'s on a countable basis of the Hilbert space. We also sketch proof of the consistency of the learning process in a general case and then, we concentrate on the computational aspect of learning the descriptors (Section 4) by the fast discrete cosine transform (DCT) and its joint action together with learning and using a classifier of descriptors. Finally,in Section 5, the proposed method was intensively tested on a large number of augmented data, leading to the selection of classifiers that cooperate with the learning descriptors in the most efficient way, from the viewpoint of the classification quality measures.

Motivating case study. Large mechanical constructions such as bucket-wheel excavators, used in open pit mines, undergo repetitive excitations that are transmitted to an operator's cabin, invoking unpleasant vibrations, which influence the operator's health in the long term. These excitations can be measured by accelerometers, as samples from functional observations that repeatedly occur after each stroke of the bucket into the ground. Roughly speaking, these functional observations can be classified into two classes, namely, to class I, representing typical, heavy working conditions and to class II, corresponding to less frequent and less heavy working conditions, occurring, e.g., when a sand background material is present (see Fig. 1 for an excerpt of functional data from Class I and II, a benchmark file is publicly available from the Mendeley site [28], see also [29] for its detailed description).

Proper and fast classification can be useful for decision making whether to use more or fewer vibrations damping in the next period between subsequent shocks, invoked by strokes of the bucket into the ground. We refer the reader to [25] to the study on a control system based on magneto-rheological dampers, for which the classifier proposed here can be used as an upper decision level.

Previous works. Over the last twenty years the problems of classifying functions, curves and signals using methods from functional analysis has attracted considerable attention from researchers. We refer the reader to the fundamental paper [10] on (im-)possibilities of classifying proba-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

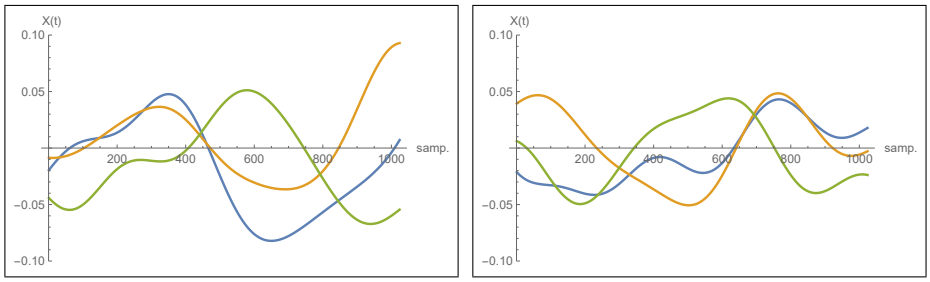

Fig. 1. An excerpt of functional data, representing accelerations vs sample number of an operator's cabin in bucket-wheel excavators. Left panel – curves from Class I (heavy working conditions), right panel – curves from Class II (less onerous working conditions)

bility density functions with (or without) certain qualitative properties. Function classification, using a functional analogue of the Parzen kernel classifier is developed in [6], while in [12], [5] generalizations of the Mahalanobis distance are applied. Mathematical models of functional data are discussed in [18]. The reader is also referred to the next section for citations of related monographs and to [21].

All the above does not mean that problems of classifying functions, mainly sampled signals, were not considered earlier. Conversely, the first attempts at classifying electrocardiogram (ECG) signals can be traced back, at least, to the 1960s, see [1] for the recent review and to [20] for feature selection using the FFT.

The recognition problems for many other kinds of bio-medical signals have been extensively studied. We are not able to review all of them, therefore, we confine ourselves to recent contributions, surveys, and papers more related to the present one.

Electroencephalogram (EEG) signals are rather difficult for an automatic classification, hence the main effort is put on a dedicated feature selection, see [14], [13] and survey papers [4], [19] the latter being of special interest for human-computer interactions. In a similar vein, in [2] the survey of using electromyography (EMG) signals is provided. For a long time, also studies on applying the EMG signals classification for control of hand prosthesis had been conducted. We refer the reader to recent contributions [30], [17] and to [8] for a novel approach to represent a large class of signals arising in a health care system.

Up to now, problems of classifying data from accelerometers, as those arising in our motivating case study, have not received too much attention (see [22], where the recognition of whether a man is going upstairs or downstairs is considered).

Our derivations are based on orthogonal projections. One should notice that classifiers based on orthogonal expansions were studied for a long time, see [15] for one of the pioneering papers on classifiers based on probability density estimation and [9] for the monograph on probabilistic approaches to pattern recognition. Observe, however, that in our problem we learn the expansion coefficients in a way that closer to non-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

parametric estimation of a regression function with non-random (fixed design) cases (see, e. g., [23]). Furthermore, in our case observation errors are correlated, since they arise from the truncation of the orthogonal series with random coefficients.

2 Model of random functional data and problem statement

Constructing a simple mathematical description of random functional data, also called random elements, is a difficult task, since in infinite dimensional Hilbert spaces an analogue of the uniform distribution does not exists (see monographs: [16], [11], [27], [3] for basic facts concerning probability in spaces of functions). Thus, it is not possible to define probability density function (p.d.f.) with respect to this distribution. As a way to get around this obstacle, we propose a simple model of random elements in the Hilbert space $L_2(0,\, T)$ of all squared integrable functions, where $T > 0$ is the horizon of observations.

V1) Let us assume that $\mathbf{v}_k(t)$, $t \in [0,\, T$, $k = 1,\, 2, \ldots$ is a selected orthogonal and complete, infinite sequence of functions in $L_2(0,\, T)$, which are additionally normalized, i.e., $||\mathbf{v}_k|| = 1$, $k = 1,\, 2, \ldots$, where for $g \in L_2(0,\, T)$ its squared norm $||g||^2$ is defined as $< g,\, g >$, while $< g,\, h >= \int_0^T g(t)\, h(t)\, dt$ is the standard inner product in $L_2(0,\, T)$.

Within this framework, any $g \in L_2(0,\, T)$ can be expressed as

$$g = \sum_{k=1}^{\infty} < g,\, \mathbf{v}_k > \mathbf{v}_k, \tag{1}$$

where the convergence is understood in the L_2 norm. For our purposes we consider a class of random elements, denoted further as $\mathbf{X}$, $\mathbf{Y}$ etc. that can be expressed as follows

$$\mathbf{X} = \sum_{k=1}^{K} \theta_k\, \mathbf{v}_k \,+\, \sum_{k=K+1}^{\infty} \alpha_k\, \mathbf{v}_k, \tag{2}$$

where

- $1 \leq K < \infty$ is a preselected positive integer that splits[3] the series expansion of $\mathbf{X}$ into two parts, namely, the first one that we later call an informative part and the second one, which is either much less informative or noninformative at all from the point of view of classifying $\mathbf{X}$,,
- coefficients θ_k , $k = 1,\, 2 \ldots, K$ are real-valued random variables that are drawn according to exactly one of cumulative, multivariate distribution functions $F_I(\bar{\theta})$ or $F_{II}(\bar{\theta})$, $\bar{\theta} \stackrel{def}{=} [\theta_1,\, \theta_2,\, \ldots,\, ,\theta_K]$,
- α_k, $k = (K+1), (K+2), \ldots$ are also random variables (r.v.'s), having properties that are specified below.

[3] For theoretical purposes K is assumed to be fixed and a priori known. Later, we comment on the selection of K in practice.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

We shall write

$$\mathbf{X}(t) = \sum_{k=1}^{K} \theta_k \, \mathbf{v}_k(t) + \sum_{k=K+1}^{\infty} \alpha_k \, \mathbf{v}_k(t), \quad t \in [0, T], \tag{3}$$

when the dependence of $\mathbf{X}$ on t has to be displayed.
Distribution functions $F_I(\bar{\theta})$ and $F_{II}(\bar{\theta})$, as well as those according to α_k, $k = (K+1), (K+2), \ldots$ are drawn, are not known, but we require that the following assumptions hold.

R1) The second moments of θ_k, $k = 1, 2 \ldots, K$ exist and they are finite. The variances of θ_k's are denoted as σ_k^2.

R2) The expectations $\mathbb{E}(\alpha_k) = 0$, $k = (K+1), (K+2), \ldots$, where $\mathbb{E}$ denotes the expectations with respect to all θ_k's and α_k's. Furthermore, there exists a finite constant $0 < C_0 < \infty$, say, such that

$$\mathbb{E}(\alpha_k^2) \leq \frac{C_0}{k^2}, \quad k = (K+1), (K+2), \ldots. \tag{4}$$

R3) Collections θ_k, $k = 1, 2 \ldots, K$ and α_k, $k = (K+1), (K+2), \ldots$ are mutually uncorrelated in the sense that $\mathbb{E}(\theta_k \, \alpha_l) = 0$ for all $k = 1, 2 \ldots, K$ and $l = (K+1), (K+2), \ldots$. Furthermore, $\mathbb{E}(\alpha_j \, \alpha_l) = 0$ for $j \neq l$, $j, l = (K+1), (K+2), \ldots$.

To motivate assumption R2), inequality (4), notice that expansion coefficients of smooth, e.g., continuously differentiable, functions into the trigonometric series decay as $O(k^{-1})$, while the second order differentiability yields $O(k^{-2})$ rate of decay.
To illustrate the simplicity of this model, consider $\bar{\theta}$ that drawn at random from the K-variate normal distribution with the expectation vector $\bar{\mu}_c$ and the covariance matrix Σ_c^{-1}, where c stands for class label I or II. Consider also sequence α_k, $k = (K+1), (K+2), \ldots$ of the Gaussian random variables, having the zero expectations, that are mutually uncorrelated and uncorrelated also with $\bar{\theta}$. Selecting the dispersions of α_k's of the form: σ_0/k, $0 < \sigma_0 < \infty$, we can draw at random $\bar{\theta}$ and α_k's for which R1)–R3) hold. Thus, it suffices to insert them into (3). We underline, however, that in the rest of the paper, the gaussianity of $\bar{\theta}$ and α_k's are not postulated.

Lemma 1 (Model correctness). *Under V1), R1) and R2) model* (2) *is correct in the sense that* $\mathbb{E}||\mathbf{X}||^2$ *is finite.*

Indeed, applying V1), and subsequently R1) and R2), we obtain

$$\mathbb{E}||\mathbf{X}||^2 = \sum_{k=1}^{K} \mathbb{E}(\theta_k^2) + \sum_{k=(K+1)}^{\infty} \mathbb{E}(\alpha_k^2) \leq \sum_{k=1}^{K} \mathbb{E}(\theta_k^2) + C_0 \, \gamma_K, \tag{5}$$

where $\gamma_K \stackrel{def}{=} \sum_{k=(K+1)}^{\infty} k^{-2} < \infty$, since this series is convergent. •

Lemma 2 (Correlated observations). *Under V1), R1) and R2) observations* $\mathbf{X}(t')$ *and* $\mathbf{X}(t'')$ *are correlated for every* t', $t'' \in [0, T]$ *and for their covariance we have:*

$$\mathbb{C}ov(\mathbf{X}(t'), \mathbf{X}(t'')) = \sum_{k=(K+1)}^{\infty} \mathbb{E}(\alpha_k^2) \, \mathbf{v}_k(t') \, \mathbf{v}_k(t''), \tag{6}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

and, for commonly bounded basis functions, its upper bound is given by

$$|\mathbb{C}ov(\mathbf{X}(t'), \mathbf{X}(t''))| \leq c_0^2 \gamma_K, \quad c_0 \stackrel{def}{=} \sup_k \sup_{t\in[0,\,T]} |\mathbf{v}_k(t)|. \tag{7}$$

Problem statement. Define a residual random element $\mathbf{r}_K$ as follows: $\mathbf{r}_K = \sum_{k=(K+1)}^{\infty} \alpha_k \mathbf{v}_k$ and observe that (by R2)) $\mathbb{E}(\mathbf{r}_K) = 0$, $\mathbb{E}||(\mathbf{r}_K)||^2 \leq C_0 \gamma_K < \infty$. Define also an informative part of $\mathbf{X}$ as $\mathbf{f}_{\bar\theta} = \sum_{k=1}^{K} \theta_k \mathbf{v}_k$ and assume that we have observations (samples) of $\mathbf{X}$ at equidistant points $t_i \in [0,\, T]$, $i = 1,\, 2, \ldots,\, m$ which are of the form

$$x_i = \mathbf{X}(t_i) = \mathbf{f}_{\bar\theta}(t_i) + \mathbf{r}_K(t_i), \quad i = 1,\, 2, \ldots,\, m. \tag{8}$$

Having these observations, collected as $\bar x$, at our disposal, the problem is to classify $\mathbf{X}$ to class I or II. These classes correspond to unknown information on whether $\bar\theta$ in (8) was drawn according to F_I or F_{II} distributions, which are also unknown.
The only additional information is that contained in samples from learning sequence $\{(\mathbf{X}^{(1)},\, j_1),\, (\mathbf{X}^{(2)},\, j_2),\, \ldots,\, (\mathbf{X}^{(N)},,\, j_N)\}$. The samples from each $\mathbf{X}^{(n)}$ have exactly the same structure as (8) and they are further denoted as $\bar x^{(n)} = [x_1^{(n)},\, x_2^{(n)},\, \ldots x_m^{(n)}]^{tr}$, while $j_n \in \{I,\, II\}$, $n = 1,\, 2,\, \ldots,\, N$ are class labels attached by an expert.
Thus, the learning sequence is represented by collection $\mathcal{X}_N \stackrel{def}{=} [\bar x^{(n)},\, n = 1,\, 2,\, \ldots,\, N]$, which is an $m \times N$ matrix and the sequence of labels $\mathcal{J} \stackrel{def}{=} \{j_n \in \{I,\, II\},\, n = 1,\, 2,\, \ldots,\, N\}$ Summarizing, our aim is to propose a nonparametric classifier that classifies random function $\mathbf{X}$, represented by $\bar x$, to class I or II and a learning procedure based on $\mathcal{X}_N$ and the corresponding j_n's.

3 Learning descriptors from samples and their properties

The number of samples in $\bar x$ and $\bar x_n$'s is frequently very large (when generated by electronic sensors, it can be thousands of samples per second). Therefore, it is impractical to build a classifier directly from samples. Observe that the orthogonal projection of $\mathbf{X}$ on the subspace spanned by $\mathbf{v}_1,\, \mathbf{v}_2,\, \ldots \mathbf{v}_K$ is exactly $\mathbf{f}_{\bar\theta}$. Thus, the natural choice of descriptors of $\mathbf{X}$ would be $\bar\theta$, but it is not directly accessible. We do not have also a direct access to $\bar\theta^{(n)}$'s constituting $\mathbf{X}^{(n)}$'s. Hence, we firstly propose a nonparametric algorithm of learning $\bar\theta$ and $\bar\theta^{(n)}$'s from samples. We emphasize that this algorithm formally looks like as well known algorithms of estimating regression functions (see, e.g., [23], [26]), but its statistical properties require re-investigation, since noninformative residuals $\mathbf{r}_N$ have a different correlation structure than that which appears in classic nonparametric regression estimation problems.
Denote by $\hat\theta_k$ the following expression

$$\hat\theta_k = \frac{T}{m} \sum_{i=1}^{m} x_i \mathbf{v}_k(t_i), \quad k = 1,\, 2,\, \ldots,\, K \tag{9}$$

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

further taken as the learning algorithm for $\theta_k =< \mathbf{X}, \mathbf{v}_k >$.
Asymptotic unbiasedness. It can be proved that for continuously differentiable $\mathbf{X}(t)$, $t \in [0, T]$ and $\mathbf{v}_k$'s we have

$$|\mathbb{E}_{\bar{\theta}}(\hat{\theta}_k) - \theta_k| \leq \frac{T\, L_1}{m} \tag{10}$$

where $\mathbb{E}_{\bar{\theta}}$ is the expectation with respect to α_k's, conditioned on $\bar{\theta}$ and $L_1 > 0$ is the maximum of $|\mathbf{X}'(t)|$ and $|\mathbf{v}'_k(t)|$, $t \in [0, T]$.
One can largely reduce errors introduced by approximate integration by selecting $\mathbf{v}_k$'s that are orthogonal in the summation sense on sample points, i.e.,

$$\frac{T}{m} \sum_{i=1}^{m} \mathbf{v}_l(t_i)\, \mathbf{v}_k(t_i) \;=\; 0 \text{ for } k \neq l, \quad k, l \;=\; 1, 2, \ldots . \tag{11}$$

The well known example of such basis is provided by the cosine series

$$\mathbf{v}_1(t) = 1,\, \mathbf{v}_2(t) = \sqrt{2}\, \cos(\pi\, t/T),\, \mathbf{v}_3(t) = \sqrt{2}\, \cos(2\, \pi\, t/T),\, \ldots \tag{12}$$

computed at equidistant t_i's.

Lemma 3 (Bias). *For all $k = 1, 2, \ldots, K$ we have: 1) if $\mathbf{X}(t)$ and $\mathbf{v}_k(t)$'s are continuously differentiable $t \in [0, T]$, then $\hat{\theta}_k$ is asymptotically unbiased, i.e., $\mathbb{E}_{\bar{\theta}}(\hat{\theta}_k) \to \theta_k$ as $m \to \infty$,*
2) if for $\mathbf{v}_k$, $k = 1, 2, \ldots, K$ and m conditions(11) hold, then $\hat{\theta}_k$ is unbiased for m finite, i.e., $\mathbb{E}_{\bar{\theta}}(\hat{\theta}_k) = \theta_k$.

Variance and mean square error (MSE). Analogously, assuming that $\mathbf{v}_k$'s and $\mathbf{X}(t)|$ are twice continuously differentiable, we obtain:

$$\mathbb{V}\mathrm{ar}_{\bar{\theta}}(\hat{\theta}_k) \leq \frac{T\, L_2}{m^2}\, \gamma_K, \tag{13}$$

where L_2 is the maximum of $\mathbf{X}''(t)|$ and $|\mathbf{v}''_k(t)|$, $t \in [0, T]$. Thus, the conditional mean squared error of learning $\hat{\theta}_k$ is not larger than $\frac{T\, L_1}{m} + \frac{T\, L_2}{m^2}\, \gamma_K$ and it can be reduced by enlarging m.

Lemma 4 (Consistency). *For all $k = 1, 2, \ldots, K$ we have:*

$$\mathbb{E}_{\bar{\theta}} \left(\hat{\theta}_k - \theta_k\right)^2 \;\to\; 0, \;\; \textit{as } m \;\to\; \infty, \tag{14}$$

i.e., $\hat{\theta}_k$ is consistent in the MSE sense, hence also in the probability.

Notice also that this is the worst case analysis in the class of all twice differentiable functions $\mathbf{X}(t)|$ and $|\mathbf{v}_k(t)|$, which means that L_1 and L_2 depend on k.
Observe that replacing x_i's in (9) by $x_i^{(n)}$'s we obtain estimators $\hat{\theta}_k^{(n)}$ of the descriptors $\theta_k^{(n)}$ in the learning sequence. Obviously, the same upper bounds (10) and (13) hold also for them.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

4 A fast algorithm for learning descriptors and classification

The above considerations are, to a certain extent, fairly general. By selecting (12) as the basis, one can compute all $\hat{\theta}_k$'s in (9) simultaneously by the fast algorithm, being the fast version of the discrete cosine transform (see, e.g., [7] and [24]). The action of this algorithm on $\bar{x}$ (or on $\bar{x}^{(n)}$'s) is further denoted as $\mathcal{FDCT}(\bar{x})$. Notice, however, that for vector $\bar{x}$, containing m samples, also the output of the $\mathcal{FDCT}(\bar{x})$ contains m elements, while we need only $K < m$ of them, further denoted as $\hat{\bar{\theta}} = [\hat{\theta}_k,\, k = 1, 2, \ldots, K]^{tr}$. Thus, if $\texttt{Trunc}_K[.]$ denotes the truncation of a vector to its K first elements, then

$$\hat{\bar{\theta}} = \texttt{Trunc}_K[\mathcal{FDCT}(\bar{x})], \tag{15}$$

is the required version of the learning of all the descriptors at one run, at the expense of $O(m\,\log(m))$ arithmetic operations.

Remark 1. If K is not known in advance, it is a good point to select it by applying $\texttt{Trunc}_K[\mathcal{FDCT}(.)]$ to $\bar{x}^{(n)}$'s together with the minimization of one of the well known criterions such as the AIC, BIC etc. Notice also that K plays the role of a smoothing parameter, i.e., smaller K provides a less wiggly estimate of $\mathbf{X}(t)$, $t \in [0, T]$.

A projection-based classifier for functional data. The algorithm: $\texttt{Trunc}_K[\mathcal{FDCT}(.)]$ is crucial for building a fast classifier from projections, since it will be used many times both in the learning phase as well as for fast recognition of forthcoming observations of $\mathbf{X}$'s. The second ingredient that we need is a properly chosen classifier for K dimensional vectors $\bar{\theta}$. Formally, any reliable and fast classifier can be selected, possibly excluding the nearest neighbors classifiers, since they require to keep and look up the whole learning sequence, unless its special edition is not done. For the purposes of this paper we select the support vector machine (SVM) classifier and the one that is based on the logistic regression (LReg) classifier. We shall denote by $\texttt{Class}[\bar{\theta}, \{\bar{\theta}^{(n)}, \mathcal{J}\}]$ the selected classifier that – after learning it from the collection of descriptors $\{\bar{\theta}^{(n)}\}$, $n = 1, 2, \ldots, N$ and correct labels $\mathcal{J}$ – classifies descriptor $\bar{\theta}$ of new $\mathbf{X}$ to I or II class.

A projection-based classification algorithm (PBCA)
Learning

1. Convert available samples of random functions into descriptors:

$$\bar{\theta}^{(n)} = \texttt{Trunc}_K[\mathcal{FDCT}(\bar{x}^{(n)})], \quad n = 1, 2, \ldots, N$$

   and attach class labels j_n to them in order to obtain $(\bar{\theta}^{(n)}, j_n)$, $n = 1, 2, \ldots, N$.
2. Split this sequence into the learning sequence of the length $1 < N_l < N$ with indexes selected uniformly at random (without replacements) from $n = 1, 2, \ldots, N$. Denote the set of this indexes by $\mathcal{J}_l$ and its complement by $\mathcal{J}_v$.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

3. Use $\bar{\theta}^{(n)}$, $n \in \mathcal{J}_l$ to learn classifier $\texttt{Class}[., \{\bar{\theta}^{(n)}, \mathcal{J}_l\}]$, where dot stands for a dummy variables, representing a descriptor to be classified.
4. Verify the quality of this classifier by testing it on all descriptors with indexes from $\mathcal{J}_v$, i.e., compute

$$\hat{j}_{n'} = \texttt{Class}[\bar{\theta}^{(n')}, \{\bar{\theta}^{(n)}, \mathcal{J}_l\}], \quad n' \in \mathcal{J}_v. \tag{16}$$

5. Compare the obtained class labels $\hat{j}_{n'}$ with proper ones $j_{n'}$, $n' \in \mathcal{J}_v$ and count the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases. Use them to compute the classifier quality indicators such as *accuracy, precision, F1 score, ...* and store them.
6. Repeat steps 2-5 a hundred times, say, and assess the quality of the classifier, using the collected indicators. If its quality is satisfactory, go to the on-line classification phase. Otherwise, repeat steps 2-5 for different K.

On-line classification

Acquisition: collect samples $x_i = \mathbf{X}(t_i)$, $i = 1, 2, \ldots, m$ of the next random function and form vector $\bar{x}$ from them.

Compute descriptors: $\bar{\theta} = \texttt{Trunc}_K[\mathcal{FDCT}(\bar{x})]$.

Classification: Compute predicted class label $\hat{j}$ for descriptors $\bar{\theta}$ as $\hat{j} = \texttt{Class}[\bar{\theta}, \{\bar{\theta}^{(n)}, \mathcal{J}_l\}]$.

Decision: if appropriate, make a decision corresponding to class $\hat{j}$ and go to the Acquisition step.

Even for a large number of samples from repetitive functional random data the PBCA is relatively fast for the following reasons.

- The most time-consuming Step 1 is performed only once for each (possibly long) vector of samples from the learning sequence. Furthermore, the fast $\mathcal{FDCT}$ algorithm provides the whole vector of m potential descriptors in one pass. Its truncation to K first descriptors is immediate and it can be done many times, without running the $\mathcal{FDCT}$ algorithm. This advantage can be used for even more advanced task of looking for a sparse set of descriptors, but this topic is outside the scope of this paper.
- Steps 2-5 of the learning phase are repeated many times for the validation and testing reasons, but this is done off-line and for descriptor vectors of the length $K << m$. The total execution time of the validation and testing phase depends on the time of learning $\texttt{Class}[., \{\bar{\theta}^{(n)}, \mathcal{J}_l\}]$ that depends on a particular choice of the classifier $\texttt{Class}$. For the SVM and LogReg classifiers and for K about dozens, it takes a few seconds on a standard PC with 3 GHz CPU clock.
- The execution time of the on-line usage phase is fast, since it uses the fast version of DCT only once for the incoming vector of samples $\bar{x}$ at the expence of $O(m \log(m))$ operations, while the already trained recognizer has to classify $\bar{\theta}$ of the length K only.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

5 Testing on accelerations of the operator's cabin

The PBCA was tested on samples of a function (signal), representing the accelerations of an operator's cabin (see Fig. 1), mounted on a bucket-wheel excavator. The aim of testing was not only to check the correctness of the algorithm, but also to select a suitable classifier.

We had 44 000 samples, acquired with the rate 512 Hz and grouped into portions of $T = 2$ sec. duration each. The resulting $\bar{x}^{(n)}$'s of the length $m = 1024$ samples, representing the learning sequence $\mathbf{X}_n$, $n = 1, 2, \ldots, N = 43$, were extended by adding labels of their proper classifications. A low-pass filter with the cutoff[4] frequency 5 Hz was applied before using $\mathcal{FDCT}$. The number of $K = 16$ of estimated descriptors $\hat{\theta}_k^{(n)}$, $k = 1, 2, \ldots, K$ was selected as the first K elements of $\mathcal{FDCT}$ sequences.

As one can notice, 44 000 samples occurred to be low informative for functional data classification. Therefore, for the aim of our tests, we had to use augmented data. In the augmentation process we used a silent, nice feature of the projection-based descriptors and the linearity of (9) with respect to samples. Namely, instead of augmenting raw samples, we augmented $\hat{\theta}_k^{(n)}$, $k = 1, 2, \ldots, K$ by adding to each of them pseudo-random errors that had Gaussian distribution with zero mean and dispersion $\sigma_a = 0.018$. Taking into account that most of $\hat{\theta}_k^{(n)}$'s was of the order ± 0.5, the interval $\pm 3\,\sigma_a$ has the length of 10.8 % of their amplitudes. In this way the augmented testing sequence, containing $N' = 43\,000$ examples, having $K = 16$ descriptors , was generated.

The following classifiers were tested as part of the PBCA:

LogR – the logistic regression classifier,
SVM – the support vector machine,
DecT – the decision tree classifier,
gbTr – the gradient boosted trees,
RFor – the random forests classifier,
5NN – the 5 nearest neighbors[5] classifier.

The results of learning and testing are summarized in Table 1. Its right panel contains just one example of the confusion matrix – for illustration only. The left panel summarizes all the extensive simulations. It contains the values of indicators that are the most frequently used for assessing the quality of classifiers.

The analysis of these quality indicators allows recommending the SVM and the LogR classifiers as the decision unit, applied after learning descriptors. Also the CPU time of $7.5\,10^{-6}$ sec., used for the SVM and LogR classifier to recognize a new example, was slightly better than for the rest

[4] From earlier experiments [25], it was known that frequencies of importance are less than 2.5 Hz.

[5] The 5 NN classifier was tested for comparisons only. We do not recommend its usage with the PCBA, since it requires storing all the learning sequence, unless its editing (condensation) is not done.

[6] MCC is the abbreviation for the Matthews Correlation Coefficient.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

Classifier	$LogR$	SVM	$DecT$	$gbTr$	$RFor$	$5\,NN$
Accuracy	0.91	**0.94**	0.84	0.92	0.91	0.90
Cohen κ	0.76	**0.82**	0.60	0.77	0.73	0.70
MCC [6]	0.76	**0.82**	0.61	0.77	0.73	0.71
Precision	**0.96**	0.94	0.93	0.94	0.92	0.90
Recall	0.92	**0.98**	0.86	0.96	0.96	**0.98**
Specificity	**0.88**	0.80	0.60	0.79	0.72	0.65
FScore	0.94	**0.96**	0.89	0.95	0.94	0.94

	Pred. class	
	I	*II*
I	30303	2697
II	1220	8780

Table 1. Left table – a summary of learning and testing the PBCA on the augmented acceleration data for different classifiers (abbreviations explained in the text). Right table – an example of the confusion matrix when the LogR classifier was used.

of classifiers displayed in Table 1, which needed about $10 - 15\,10^{-6}$ sec., as the average of 30000 simulation experiments.

6 Concluding remarks

The mathematical model of random infinite-dimensional data is proposed that allows us to impose arbitrary probability distribution on a finite dimensional space of descriptors. Its correctness is proved and the learning algorithm for these descriptors is proposed and investigated. In particular, it was shown that the learning algorithm is consistent in the MSE sense for any finite number of the descriptors.

The fast version of the learning algorithm is tested from the view point of its cooperation with a finite dimensional classifier. The winners are the SVM and logistic regression classifiers, as tested on augmented real data.

By passing, a new approach to data augmentation is proposed. Namely, instead of augmenting raw observations, we use random perturbation of estimated descriptors, which leads to essential computational savings. On the other hand, the descriptors estimated from the raw learning sequence are sufficient for learning the classifiers, which means a kind of raw data compression when they are disregarded.

Further research in this direction is desirable. One can consider extending them by including ensembles of classifiers and neural network-based recognizers.

From the practical point of view, it would be also of interest to consider the classification of signals from accelerometers to more than two classes, taking into account the kind of background that is met by a bucket-wheel excavator. This is, however, outside the scope of this paper, since it requires cumbersome data labeling by experts.

Further directions of research may include also other applications, e.g., a human motion classification, based on a motion capture cameras, a computer-aided laparoscopy training and theoretical aspects such as classifying random elements by learning their derivatives.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

References

1. Abdulla, L., Al-Ani, M.: A review study for electrocardiogram signal classification. UHD Journal of Science and Technology **4**(1), 103–117 (2020). https://doi.org/10.21928/uhdjst.v4n1y2020.pp103-117
2. Ahsan, M.R., Ibrahimy, M.I., Khalifa, O.O., et al.: EMG signal classification for human computer interaction: a review. European Journal of Scientific Research **33**(3), 480–501 (2009)
3. Aneiros, G., Bongiorno, E.G., Cao, R., Vieu, P., et al.: Functional statistics and related fields. Springer (2017)
4. Azlan, W.A., Low, Y.F.: Feature extraction of electroencephalogram (EEG) signal - a review. In: 2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES). pp. 801–806 (2014). https://doi.org/10.1109/IECBES.2014.7047620
5. Berrendero, J.R., Bueno-Larraz, B., Cuevas, A.: On Mahalanobis distance in functional settings. Journal of Machine Learning Research **21**(9), 1–33 (2020)
6. Biau, G., Bunea, F., Wegkamp, M.H.: Functional classification in Hilbert spaces. IEEE Transactions on Information Theory **51**(6), 2163–2172 (2005). https://doi.org/10.1109/TIT.2005.847705
7. Britanak, V., Yip, P.C., Rao, K.R.: Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations. Elsevier (2010)
8. Cyganek, B., Woźniak, M.: Tensor based representation and analysis of the electronic healthcare record data. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1383–1390 (2015). https://doi.org/10.1109/BIBM.2015.7359880
9. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, vol. 31. Springer Science & Business Media (2013)
10. Devroye, L., Lugosi, G.: Almost sure classification of densities. Journal of Nonparametric Statistics **14**(6), 675–698 (2002). https://doi.org/10.1080/10485250215323
11. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Science & Business Media (2006)
12. Galeano, P., Joseph, E., Lillo, R.E.: The Mahalanobis distance for functional data with applications to classification. Technometrics **57**(2), 281–291 (2015)
13. Gandhi, T., Panigrahi, B.K., Anand, S.: A comparative study of wavelet families for EEG signal classification. Neurocomputing **74**(17), 3051–3057 (2011)
14. Garrett, D., Peterson, D.A., Anderson, C.W., Thaut, M.H.: Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Transactions on Neural

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

Systems and Rehabilitation Engineering **11**(2), 141–144 (2003). https://doi.org/10.1109/TNSRE.2003.814441

15. Greblicki, W., Pawlak, M.: Classification using the Fourier series estimate of multivariate density functions. IEEE Transactions on Systems, Man and Cybernetics **11**, 726–730 (1981)
16. Horváth, L., Kokoszka, P.: Inference for functional data with applications, vol. 200. Springer Science & Business Media (2012)
17. Kurzynski, M., Wolczowski, A.: EMG and MMG signal recognition using ensemble of one-feature classifiers with pruning via clustering method. In: 2019 International Conference on Advanced Technologies for Communications (ATC). pp. 38–43. IEEE (2019)
18. Ling, N., Vieu, P.: Nonparametric modelling for functional data: selected survey and tracks for future. Statistics **52**(4), 934–949 (2018). https://doi.org/10.1080/02331888.2018.1487120
19. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering **4**(2), R1–R13 (Jan 2007). https://doi.org/10.1088/1741-2560/4/2/r01
20. Mironovova, M., Bíla, J.: Fast Fourier transform for feature extraction and neural network for classification of electrocardiogram signals. In: 2015 Fourth International Conference on Future Generation Communication Technology (FGCT). pp. 1–6 (2015). https://doi.org/10.1109/FGCT.2015.7300244
21. Mueller, H.G., et al.: Peter Hall, functional data analysis and random objects. The Annals of Statistics **44**(5), 1867–1887 (2016)
22. Preece, S. J., Goulermas, J. Y., Kenney, L. P. J.: A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Transactions on Biomedical Engineering **56**(3), 871–879 (2009). https://doi.org/10.1109/TBME.2008.2006190
23. Rafajłowicz, E.: Nonparametric orthogonal series estimators of regression: a class attaining the optimal convergence rate in L2. Statistics and Probability Letters **5**, 219–224 (1987)
24. Rafajłowicz, E., Skubalska-Rafajłowicz, E.: FFT in calculating nonparametric regression estimate based on trigonometric series. Journal of Applied Mathematics and Computer and Computer Science **3**(4), 713–720 (1993)
25. Rafajłowicz, W., Więckowski, J., Moczko, P., Rafajłowicz, E.: Iterative learning from suppressing vibrations in construction machinery using magnetorheological dampers. Automation in Construction **119**, 103326 (2020)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3

26. Rutkowski, L., Rafajłowicz, E.: On optimal global rate of convergence of some nonparametric identification procedures. IEEE Trans. Automatic Control **AC-34**, 1089–1091 (1989)
27. Srivastava, A., Klassen, E.P.: Functional and shape data analysis, vol. 1. Springer (2016)
28. Więckowski, J.: Data from vibration in SchRs1200, Mendeley Data, V1. `http://dx.doi.org/10.17632/htddgv2p3b.1`, accessed: January 2021
29. Więckowski, J., Rafajlowicz, W., Moczko, P., Rafajlowicz, E.: Data from vibration measurement in a bucket wheel excavator operator's cabin with the aim of vibrations damping. Data in Brief **106836** (2021)
30. Wozniak, M., Połap, D., Nowicki, R.K., Napoli, C., Pappalardo, G., Tramontana, E.: Novel approach toward medical signals classifier. In: 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–7 (2015). https://doi.org/10.1109/IJCNN.2015.7280556

Acknowledgements. The authors express their thanks to Professor P. Moczko and Dr. J. Więckowski from the Faculty of Mechanical Engineering, Wroclaw University of Science and Technology for permission to use data from the bucket-wheel excavator.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_3