SVM classifiers for functional data in monitoring of the Internet users behaviour

Daniel Kosiorowski[1], Dominik Mielczarek[2], Jerzy P. Rydlewski[3]

Abstract

Novel tools offered by functional data analysis offers various methods especially adequate for monitoring phenomena appearing within the new economy, which are described by means of functions of a certain continuum. We mean here, among others, fraud detection in credit card transactions, electricity demand or the Internet traffic monitoring. This paper focuses our attention on a concept of classifier for functional data induced by the general support vector machines methodology. We, among other, study robustness, computational complexity and consistency of the classifiers using analytical as well as empirical arguments. We compare their properties with alternatives presented in the literature using real data set concerning behaviour of users of a big Internet service divided into four subservices.

Keywords: *classifier for functional objects, data depth, the Internet service, management, support vector machine classifier*

JEL Classification: C14, C22, C38

1 Introduction

Many phenomena, we dealt with in the field of economics, takes the form of a function. Note, that different economic phenomena can be formulated in terms of suitable classifier determining. Let us take, for instance, classifying a candidate for a certain position on a labour market or conducting sell/buy decisions within an algorithmic trading. Various research problems occur when functional outliers are present in the considered data set. Moreover, if we do not possess a reliable economic theory on data generating processes, which could be used for describing the economic phenomena, then functional generalizations of well-known statistical procedures are irrelevant (Ramsay & Silverman, 2005; Horvath & Kokoszka, 2012). Main aim of the paper is to propose a novel statistical methodology for classifying functional objects, that can be applied to monitor and manage the Internet users behaviour. This study shows, that a proper Support Vector Machine methods (SVM) (Schoelkopf and Smola, 2002) enable for efficient classification of functional data, appearing

[1] Corresponding author: Cracow University of Economics, Department of Statistics, ul. Rakowicka 27, 31-510 Kraków, e-mail: daniel.kosiorowski@uek.krakow.pl.
[2] AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland, e-mail: dmielcza@wms.mat.agh.edu.pl.
[3] AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland, e-mail: ry@agh.edu.pl.

in modern e-economy (Anagnostopoulos et al., 2012). It is clear, that this method can be used to classify various functional time series objects, e.g. daily concentration of dangerous particles in the atmosphere. Hence, this study proposes a novel nonparametric and robust method for phenomena classification, which may be described as functional objects. The rest of the paper is organized as follows. Section 2 sketches the basic concepts of classification in a functional setup. Section 3 introduces our procedure for classifying a functional data. Section 4 discusses the properties of the procedure through numerical simulations and tests the applicability of the proposed methodology on empirical examples (i.e., activities of Internet users). Section 5 conducts a short sensitivity analysis and comprises a brief summary.

2 Studied data

We consider data from certain Internet service. The service has been divided into four sub-services, called service 1-4, hereafter. Using classifying methods, we would like to classify a new functional object into one of the four considered services. Figure 1 presents four functional principal components for service 1(left) and service 2 (right), for functional PCA methods and its applications see Górecki and Krzyśko (2012). We have used free R packages fda and fda.usc for computations (Febrero-Bande and de la Fuente, 2012).

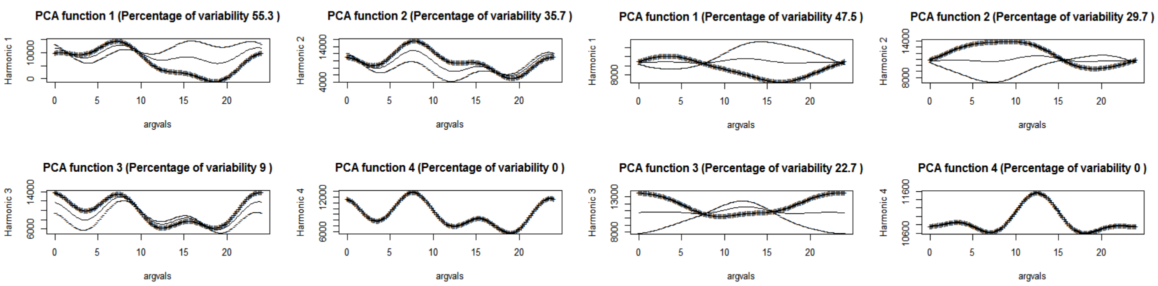


Fig. 1. Functional principal components for service 1(left) and service 2 (right).

3 Basic classification methods of functional objects

In a context of functional data classification, we are given a training sample, that is, at the beginning we are given *n* observations, namely $Z_1, Z_2, \ldots Z_n$ and each observation can be regarded as $Z_i=(f_i, Y_i)$ where $Y_i=0$ or $Y_i=1$, the f_i are called patterns cases, inputs or instances, while Y_i are called labels, outputs or targets. Our aim is to classify a new object *f* into one of the two labelled groups, basing on knowledge included in the training sample. In general, finding a classification rule means to obtain a partition of the feature space into nonempty disjoint subsets, which are summed up to a whole feature space. Partition subsets relate to the appropriate labels. Note, that the most natural method for classifying a functional data is the k-nearest neighbors method. The input set consists of the *k* closest to the classified object training examples in the considered feature space. Note, that many different metrics could be taken into account. The output is then the assignation to a class, which is most common among its *k* nearest neighbors. A crucial and still an opened problem is an appropriate choice of the metric defining neighborhood and the number of neighbors *k* (Ferraty and Vieu, 2006).

The second groups of methods is based on a concept of nearest centroid. The nearest centroid classifiers method assigns to observations the label of the class of training samples whose centroid is closest to the observation. As the centroid we may consider a functional mean, a functional median induced by a certain functional depth.

The third method is based on a concept of local depth for functional data (Paindaveine and Van Bever, 2013) recently used in Kosiorowski et al. (2016) and Kosiorowski et al. (2017). A depth function is a function that assigns to a point a positive number from an interval [0,1] expressing its centrality with respect to the probability measure or a sample, for details see Zuo and Serfling (2000). The local depth concept is a generalization of the concept of global depth. It takes into account the local features of the analysed data, what is especially important in a context of clustering, for example. The object, for which a depth takes a maximal value is called the median induced by this depth. Within Paindaveine and Van Bever (2013) concept, parameter of locality is denoted with β. It ranges from 0 (extreme localization) to 1 (global depth). Figure 2 shows local medians calculated for services 1-4, with locality parameter $\beta = 0,45$.

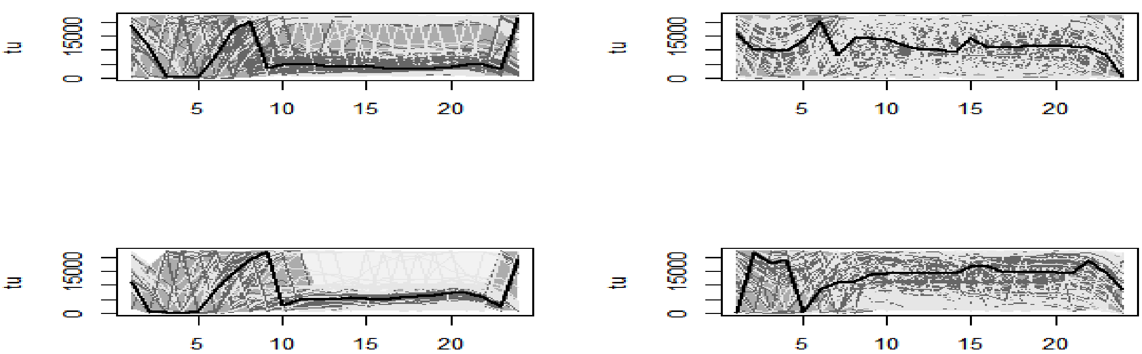


Fig. 2. Local medians calculated for services 1-4, $\beta = 0,45$.

4 Our proposals – Machine learning methods

Our first proposal requires calculation of local depth for functional data. We create then *depth versus depth* plot *(DD*-plot*)*, introduced by Liu et al. (1999). DD–plots show the values of depth function of each observation under two distributions. We use SVM methods in a two-dimensional space of DD-plots in order to divide the data into two groups. Lange et al. (2014) introduced a two-step depth transformation for functional data. Their proposition is to map the functional data into finite-dimensional location-slope space, where each functional observation is treated as a vector consisting of integrals of its levels (location) and first derivatives (slope) over and equally sized subintervals, respectively. The data are transformed then to *DD*-plots with a multivariate depth function. Finally, the data can be discriminated on the *DD*-plot with the well-known classical methods. Another possibility is to consider a depth of the functional data at the beginning of the procedure. Other methods, including Hubert et al. (2016) classifier, basing on so called bag-distance were considered in Kosiorowski and Bocian (2015). The formerly presented classifiers are characterized with a relatively large classification error. That's why we focus our attention on very promising in multivariate case SVM classifier (Schoelkopf and Smola, 2002).

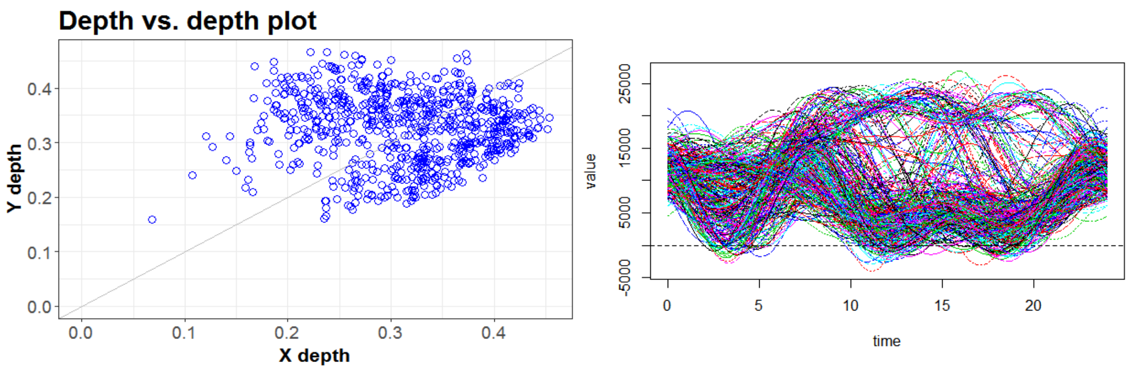
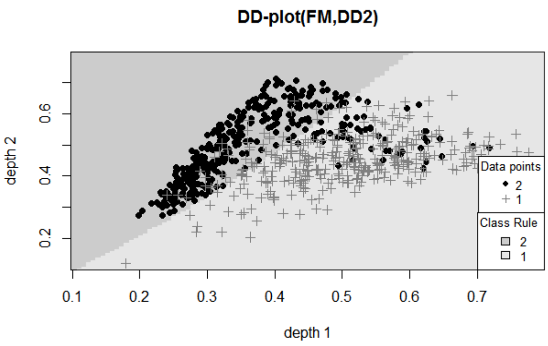
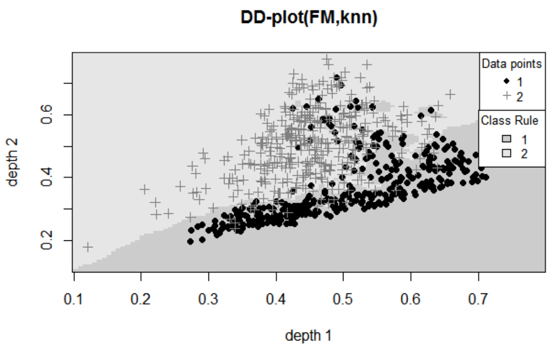
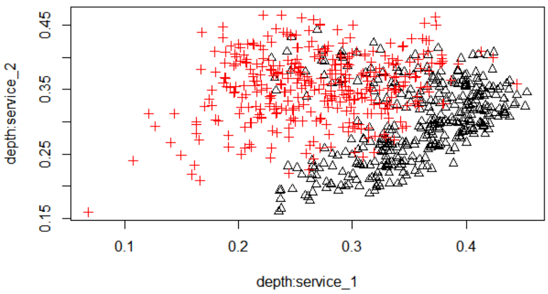


Fig. 3. DD-plot of number of users of service 1 vs. number of users of service 2, for $\beta = 0,45$ (left). Number of service 1 users (right).

Figure 4 shows the DD-plot of service 1 vs. service 2 classifiers based on Fraiman-Muniz depth. The SVM does use a second-order polynomial to make a classification. In such a case a probability of correct classification equals 0.87. Figure 5 shows the DD-plot of service 1 vs. service 2 based on Fraiman-Muniz depth. The k-nearest neighbours algorithm is used to make a classification. A correct classification probability equals 0.9148. The results show that the proposed method are a relevant methods to undertake a classification.







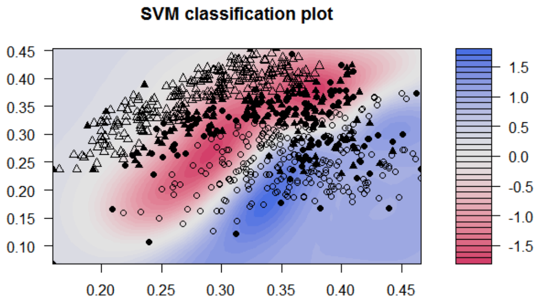


Fig. 4. (top left) DD-plot classifiers of service 1 vs. service 2 based on Fraiman-Muniz depth. The SVM does use a second-order polynomial. A correct classification probability equals 0.87.
Fig. 5. (top right) DD-plot classifier of service 1 vs. service 2 based on Fraiman-Muniz depth. The k-nearest neighbours algorithm is used. A correct classification probability equals 0.9148.
Fig. 6. (bottom left) DD-plot, calculated with modified band depth, service 1 vs. service 2.
Fig. 7. (bottom right) SVM classification used to a local DD-plot of service 1 vs. service 2.

Assume, that we have in disposal a set of points $x_{11},\ldots,x_{1m},x_{21},\ldots,x_{2m},x_{n1},\ldots,x_{nm}$. We need to transform them into functions. In other words, in order to formulate our second proposal we set a basis in the $L^2([0,T])$ integrable space of functions, namely $\phi_1,\ldots,\phi_2,\ldots$ obtain a representation of functions in the assumed space. Let us consider that the sample curves (obtained from the set of points) are centered, i.e. $Ef(t) = 0$, and $E\|f(t)\|^2 < \infty$ and that sample curves belong to the $L^2([0,T])$ space with the usual inner product defined by $< f,g >= \int_0^T f(t)g(t)dt$. We obtain a representation:

$$f_i = \sum_{j=1}^{\infty} c_j \phi_j .$$

We can use spline basis, Fourier basis (which we found the most appropriate) or another properly defined basis. The Fourier basis system is the usual choice for periodic functions, while the spline basis system is a good choice for aperiodic functions. Note, that the fixed basis is usually infinite. For this reason we need to use the representation up to the K*th* number. We determine K_i for each function with AIC (Akaike information criterion) or BIC (Bayes information criterion). We choose a mode of this set, which we denote K and the number determines the degree to which the data are smoothed. Note, that finally we do not consider functions f_i but functions $\overline{f}_i$, i.e.

$$\overline{f}_i = \sum_{j=1}^{K} c_j \phi_j .$$

Coefficients of the expansion c_j are usually estimated with least squares method, i.e. we minimize the least square criterion

$$SMSSE(f \mid c) = \sum_{i=1}^{n} \left[f_i - \overline{f}_i \right]^2 .$$

Next step is to compute principal components for functions $\overline{f}_1, \overline{f}_2, ..., \overline{f}_n$. Functional principal components allow us to make a reduction of the dimension of infinitely dimensional functional data to a smaller finite dimension in an rational way. If functional observations $f_1, f_2, \ldots, f_n$ have the same distribution as a square integrable L^2 [0,T] - valued random function f, we define the functional principal components as the eigenfunctions of the covariance operator C, where $C = E[< f - \mu, \cdot > (f - \mu)]$. The eigenfunctions of the sample covariance operator $\hat{C}(x) = \frac{1}{N}\sum_{i=1}^{n} < f_i - \hat{\mu}, x > (f_i - \hat{\mu})$, where the sample mean function is $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{n} f_i(t)$, are called the empirical functional principal components of the data. Under some regularity conditions the empirical FPC estimate the FPC's up to a sign.

Empirical FPC's may be interpreted as an optimal orthonormal basis with respect to which we can expand the data. The inner product $< f_i, \hat{v}_j > = \int f_i(t)\hat{v}_j(t)dt$ is called the j*th* score of f_i. It can be interpreted as the weight of the contribution of the functional principal components $\hat{v}_j$ to the curve f_i . To sum up, we replace the actual data by the approximation

$$\overline{f}_i = \sum_{j=1}^{p} < \hat{v}_j, \overline{f}_i > \hat{v}_j .$$

We need to determine a small value of p, such that the above approximation is relevant. A pseudo–AIC method and cross–validation have been proposed for this purpose.

The last step to be done is to use SVM methods in order to decide, whether the new observation *f* is assigned to the group with $Y_i=0$ or to the second group with $Y_i=1$. The obtained for the new observation *f* (its truncated version $\overline{f}$) principal components scores, namely $<\hat{v}_j, \overline{f}>$, i=1,2,…,n, are assigned with SVM methods into one of two considered groups. This allows us to decide, to which service we can classify the new observation. Figures 8 and 9 present functional PCA for services 1 and 2.

The quality of the PC estimator may be estimated in the following way. Assume that for any $i \in \{1,\ldots,n\}$ random variable f_i, which belongs to the Hilbert space, is described in a basis $\{\phi_i\}_{i=1}^{\infty}$, namely $f_i = \sum_{j=1}^{\infty} a_j^i \phi_j$. We obtain for any natural number *K*

$$\left\| f_i - \sum_{j=1}^{K} \overline{a}_j^i \phi_j \right\|^2 = \left\| \sum_{j=1}^{K} (a_j^i - \overline{a}_j^i)\phi_j + \sum_{j=K+1}^{\infty} a_j^i \phi_j \right\|^2 = \sum_{j=1}^{K} |a_j^i - \overline{a}_j^i|^2 + \sum_{j=K+1}^{\infty} |a_j^i|^2 \geq \sum_{j=K+1}^{\infty} |a_j^i|^2 .$$

Hence, we get $\min_{i\in\{1,\ldots n\}} \left\| f_i - \sum_{j=1}^{K} \overline{a}_j^i \phi_j \right\|^2 \geq \min_{i\in\{1,\ldots n\}} \sum_{j=K+1}^{\infty} |a_j^i|^2$.

Let $j_0 \in \{1,\ldots n\}$ denote a natural number such, that $\sum_{j=K+1}^{\infty} |a_j^{j_0}|^2 = \min_{i\in\{1,\ldots,n\}} \sum_{j=K+1}^{\infty} |a_j^i|^2$. Hence, we obtain the best estimator $\overline{f} = \sum_{j=1}^{K} a_j^{j_0} \phi_j$ and the estimation error $\left\| f_i - \sum_{j=1}^{K} a_j^{j_0} \phi_j \right\|^2 \geq \sum_{j=K+1}^{\infty} |a_j^i|^2$.

The third method we propose is based on kernel principal components (Schoelkopf and Smola, 2002). In other words, we need to study a map $\Phi : X \to H$, where H is a Hilbert space, such that $k(x, x') = <\Phi(x), \Phi(x')>$. We have to choose a kernel, where k is a real, symmetric and continuous function. Kernel function may be i.e. a Gaussian kernel or polynomial kernel. Next step is to calculate a covariance matrix $C = \frac{1}{N}\sum_{j=1}^{n} \Phi(f_j)\Phi^T(f_j)$. We create a positive definite matrix $K=(k_{ij})=(k(x_i,x_j))_{i,j}$. Consequently, we calculate kernel principal components

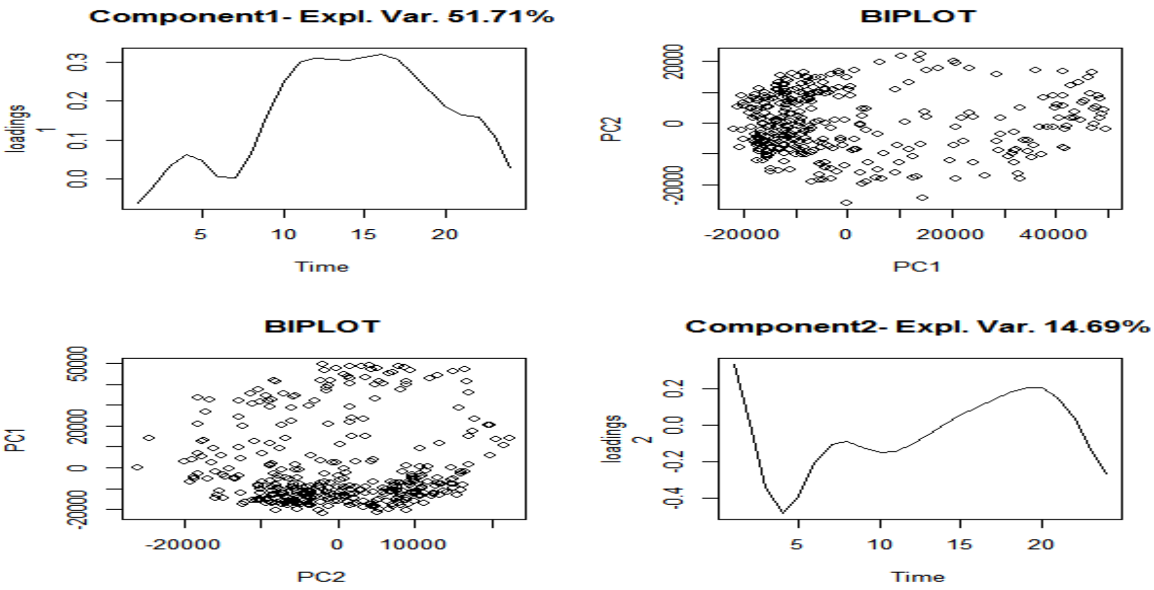


Fig. 8. Functional PCA for service 1.

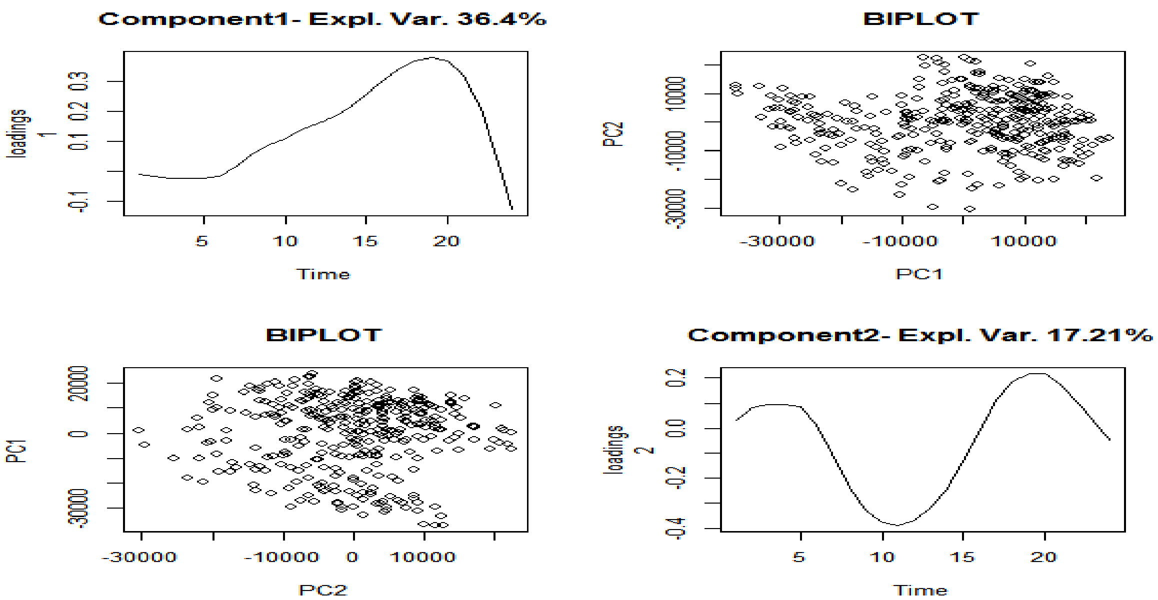


Fig. 9. Functional PCA for service 2.

from the equation $\lambda v=Cv$, where λ denote eigenvalues and v denote eigenvectors. There exist α_j, j=1,2,…,n such that $v=\sum_{j=1}^{n}\alpha_j \Phi(f_j)$. In practice, we cannot calculate vectors $\{\Phi(f_i)\}$, i=1,2,....,n, because we do not know a function Φ. However, we calculate $\hat{\Phi}_i=\Phi_i-\frac{1}{n}\sum_{k=i}^{n}\Phi_k$, where $\Phi_k=\Phi(\bar{f}_k)$ and $\hat{K}=(\hat{k})_{ij}=(<\hat{\Phi}_i,\hat{\Phi}_j>)$. In this setting we obtain a formula

$\hat{K} = PKP$, where $P = (\delta_{ij} - \frac{1}{n})$ and K is a kernel matrix. Hence, we need to solve an equation $\hat{K}\alpha = \lambda\alpha$ in order to obtain principal components. Then, we do proceed as in the preceding method and we use the SVM method in order assign the new observation *f* into one of the considered two groups.

Conclusions

This paper proposes three classification methods for functional data. All of them base on SVM methodology. The first method uses DD-plots, where DD-plots have been calculated using local depth concept. The second method and the third one are using functional principal components methods, but finally SVM is used as well. The presented methodology enables monitoring phenomena appearing within the new economy, which are described by means of functions of a certain continuum. We show on a real data set, related to Internet users behaviour, promising properties of our proposal. In a future, we plan further studies of the proposals using gamma-regression (Rydlewski, 2009) and beta-regression (Rydlewski and Mielczarek, 2012) methods in SVM classifiers development.

Acknowledgements

DK thanks for the support related to CUE grant for the research resources preservation 2017. JPR and DM thanks for the support related to AGH local grant.

References

Anagnostopoulos, C., Tasoulis, D. K., Adams, N. M., Pavlidis, N. G., & Hand, D. J. (2012). Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Statistical Analysis and Data Mining,* 5, 139-166.

Febrero-Bande, M., & de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, *51*(4), 1-28.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Górecki, T., & Krzyśko, M. (2012). Functional principal component analysis. In: Pociecha, J., & Decker, R. (eds.), *Data Analysis Methods and its Applications* (pp. 71-87). Warsaw: Beck.

Horvath, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. Springer.

Hubert, M., Rousseeuw, P. & Segaert P. (2016). Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*. doi: 10.1007/s11634-016-0269-3

Kosiorowski, D., & Zawadzki, Z. (2014). DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena. Retrieved from https://arxiv.org/abs/1408.4542v7

Kosiorowski, D., & Bocian, M. (2015). Functional Classifiers in management of the Internet service. In M. Papież & S. Śmiech (eds.), *The 9th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena*. Cracow.

Kosiorowski, D., Rydlewski, J. P., & Snarska M., (2017). Detecting a structural change in functional time series using local Wilcoxon statistic. *Statistical Papers*. doi: 10.1007/s00362-017-0891-y

Kosiorowski, D., Mielczarek, D., Rydlewski, J. P. & Snarska M., (2016). Generalized Exponential smoothing in prediction of hierarchical time series. Retrieved from https://arxiv.org/abs/1612.02195v1

Lange, T., Mosler, K., & Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1), 49-69.

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, *27*(3), 783-858.

Paindaveine, D., & Van Bever, G. (2013). “From Depth to Local Depth: a Focus on Centrality.” *Journal of the American Statistical Association*, 105, 1105-1119.

Ramsay J., & Silverman, B. (2005). *Functional data analysis*. Springer.

Rydlewski, J. P. (2009). A note on the maximum likelihood estimator in the gamma regression model. *Opuscula Mathematica*, 29(3), 305-312.

Rydlewski, J. P., & Mielczarek D. (2012). On the maximum likelihood estimator in the generalized beta regression model. *Opuscula Mathematica*, 32(4), 761-774.

Schoelkopf, B., & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth functions. *The Annals of Statistics*, 28(2), 461-482.