Functional classifiers in management of the Internet service

Daniel Kosiorowski[1], Mateusz Bocian[2]

Abstract

In this paper we focus our attention on selected classification rules for the functional objects. We consider functional k-nearest neighbour classifier, maximum local depth classifier, minimum bagdistance classifier, and classification based on the depth-depth plot. We study usefulness of these classifiers in a context of management of the Internet service. Theoretical considerations are illustrated by means of real data sets related to day and night the Internet users behaviors in 2013 year for the two big Internet services.

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

JEL Classification: C14, C22, C38

1. Introduction

An analysis of functional data in the statistical science has a long history – it is enough to mention attempts of Gauss and Legendre at the beginning of the 19th century, who have estimated trajectories of comets. Nowadays the Functional Data Analysis (FDA) provides a powerful family of statistical techniques for analysis functional objects appearing within the economics. The FDA undoubtedly brings up a conceptually new areas of economic research and provides new methodology for applications.

Various economic phenomena often directly lead to functional data i.e., e.g., yield curves, income densities, development paths.

It is worth underlying, that many of financial, meteorological or medical data are observed at not equally spaced discrete time points. The time intervals between subsequent observations often differ, what makes applications of well known statistical machinery like moving averages of ARIMA modelling at least problematic. In a context of analysis of the Internet service audience, we consider data on clicks in content, page views, or video materials views, that are performed in different time intervals (especially analyzed from the perspective of a single user). Using FDA techniques we can convert data observed in discrete time points and irregulary spaced time intervals into functional observations and then conduct

[1] Corresponding author: Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: dkosioro@uek.krakow.pl.

[2] Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: matibo@poczta.onet.pl.

further analysis and statistical inference, similarly as in a case of vector data (see Ramsay and Silverman (1997) for details).

The measurement of activity of users in the Internet provides a specific data. The analyst has to be aware of possibility of a presence of both outliers as well as inliers within the data. At a high level of data aggregation, we have to deal with sudden an unusual changes in a process characteristics (mostly increases) caused for example by important media events, unexpected catastrophes and accidents, a sort of promoted content (galleries with many pictures), enormous traffic from social media or other services as a result of a performed marketing campaign. All of the mentioned above situations always strongly influence (in specific days, even hours) a number of page views, and at the same time often resulting in small changes in a number of unique users of service (a unique user of service is counted only once regardless of e.g. a number of displayed pictures in gallery). In a view of described possibilities, classifiers used to the Internet activity data have to ensure the stability of indications, both in a situation of an absence and a presence of outliers in a sample.

In a context of classifying the Internet functional data, we can consider issue of assigning user to specific group of users (resulting e.g. in a specific offer of a content of a service) based on a recorded activity during certain period of time (page views, video views). We also mean here a classification of the Internet users to a specific target advertisement group, based on behaviours and interests of the users in the past.

The rest of the paper is organized as follows: in Section 2, seven classifiers for functional data are briefly described and certain new approach is proposed. In Section 3, results of comparisons of these classifiers are presented. The paper ends with conclusions, plans for a future studies and references.

2. Classifiers for functional data

We assume there exist several classes $C_1, C_2, ..., C_G$, to which the functional data belong to. Our problem to solve is to assign a new functional object to one of these classes basing on a training sample representing the classes being in our disposal. Let us focus our attention on certain new classifiers for functional data known from the literature and certain new classifiers which were proposed within last months.

2.1 The k-nearest neighbors classifier

At first let us consider a classifier based on well known *k-nearest neighbors rule*. In general, this “majority voting rule” classifies the object to the group with the highest number of nearest neighbors. A neighborhood is determined by means of some metric or based on some other procedures (e.g. depth-based neighborhood, see Paindaveine and Van Bever, 2013). In particular, for this purpose we compute distances between curves, using functional L^2-distance. For p-variate curves $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ defined on a set U, this distance is given by:

$$d_2(\mathbf{X},\mathbf{Y})=\sqrt{\int_U \left\|\mathbf{X}(t)-\mathbf{Y}(t)\right\|^2 dt}. \tag{1}$$

With a set of k closest curves of $\mathbf{X}(t)$ in terms of the functional L^2-distance, the k-nearest neighbors classifier for functional data assigns curve $\mathbf{X}(t)$ to the class which is represented in this neighborhood in the highest degree (see Ferraty and Vieu (2006), Rousseeuw, Hubert and Segaert (2014)). The number of neighbors k is chosen using cross-validation minimizing the misclassification rate.

2.2 Minimum distance to central curve classifier

In a context of the univariate functional data, Lopez-Pintado and Romo (2005) introduced the depth based classification rule consisted of the following steps:

1. Estimate the central curve $\mathbf{X}_g$ in each class using β-trimmed functional mean.
2. Compute distance between a new functional observation $\mathbf{X}$ and the central curve of each class $\mathbf{X}_g$ (using e.g. the formula (1)).
3. Classify $\mathbf{X}$ to the class for which the distance is the smallest.

2.3 Depth-based classifiers

The data depth concept (DDC) was originally introduced as a way to generalize the concepts of the median and quantiles to the multivariate framework. A depth function $D(\cdot,F)$ ascribes to a given $\mathbf{x}\in\mathbb{R}^d$ a measure $D(\mathbf{x},F)\in[0,1]$ of its centrality w.r.t. a probability measure $F\in\mathcal{P}$ over $\mathbb{R}^d$ or w.r.t. an empirical measure $F_n\in\mathcal{P}$ calculated from a sample $\mathbf{X}^n=\{\mathbf{x}_1,...,\mathbf{x}_n\}$. The larger the depth of $\mathbf{x}$, the more central $\mathbf{x}$ is w.r.t. to F or F_n. The best known examples of depth functions in the literature are Tukey and Liu depths. The data depth concept provides user friendly and powerful tools for nonparametric and robust multivariate data analysis. The observation of maximal value of depth is called d -variate median. Figure 1 presents a two-dimensional median and an order of the observations determined by the depth contours (points with the same depth value). A theoretical background of data depth concept can be found in Zuo and Serfling (2000) and Kosiorowski (2012).

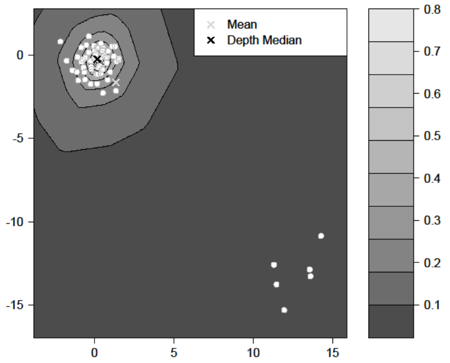

Fig. 1. Sample projection depth contour plot (DepthProc package, see Kosiorowski and Zawadzki (2014)).

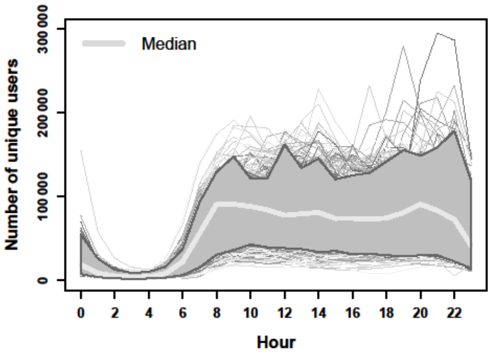

Fig. 2. Number of unique users of the Internet services during the day, with marked 50% central region.

Statistical depth functions can be succesfully defined for functional objects. An application of a functional depth enables us for center outward ordering of curves. A curve with the highest depth value is the most central (w.r.t. to the specific considered set of curves) and is called the functional median, induced by certain depth function (see Figure 2).

One of the most widely used depths for functional data is Fraiman-Muniz' depth (see Fraiman and Muniz (2001), Cuevas, Febrero-Bande and Fraiman (2007)). Let us consider N functions $\left\{X_i(t), t\in\left[t_0,t_L\right]\right\}$, $F_{N,t}(x)=N^{-1}\sum_{i=1}^{N} I\left\{X_i(t)\le x\right\}$. Fraiman and Muniz defined the functional depth by integrating one of the univariate depth (it is possible to use other depths in the below definition):

$$FD_N\left(X_i \mid X^n\right)=\int_{t_0}^{t_L}\left[1-\left|1/2-F_{N,t}\left(X_i(t)\right)\right|\right]dt. \tag{2}$$

In the literature, several robust depth-based classifiers for functional data were proposed. Generally, there are two ways of classifiyng data using depth functions. In the first approach, objects are classified to classes, in which the objects attain maximal depth values. The second approach is to transform the data into low-dimensional space based on their depths (or depth-based distances) and then classify them within the resulting space.

In the context of the first approach, Ghosh and Chaudhuri (2005) proposal of maximum depth classifier should be pointed out. They introduced a classifier assigning an object to the class in which its depth takes the highest value. An illustration of this kind of linear separation is shown in the Figure 3.

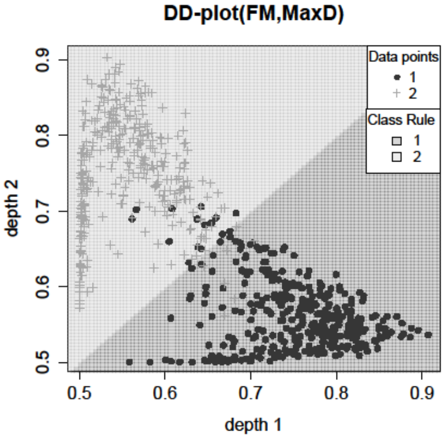

Fig. 3. Maximum depth rule classification illustrated on depth-depth plot - unique users data.

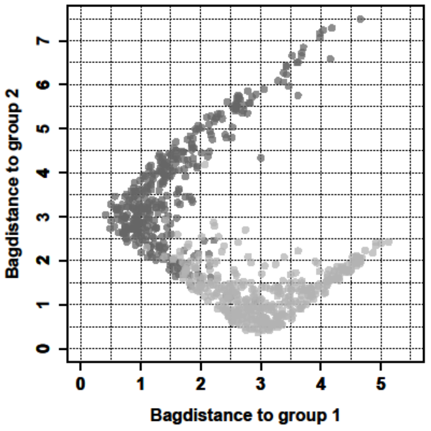

Fig. 4. Distance-distance plot - unique users data.

The second group of functional classifiers is based on *depth versus depth* plot *(DD*-plot*)*, introduced by Liu et al. (1999). It shows the values of depth function of each observation under two distributions (see Figure 3). This method can be used for univariate, multivariate and functional data. In the last two cases, we consider reduction of dimension of the statistical issue to two-dimensional depth space. Classifiers based on this transformation are called *DD*-classifiers. Finally, the observations from different classes can be separated on the *DD*-plot applying techniques known from the classical discriminant analysis such as linear and quadratic discriminant functions, or k-nearest neighbor classifier (see Li et al., 2012).

In turn, Mosler and Mozharovskyi (2014) introduced a new two-step depth transform for functional data. They proposed first 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 L and S equally sized subintervals, respectively. Then, the data in (L, S)-space are transformed to *DD*-plot using a multivariate depth function. Finally, the data can be discriminated on the *DD*-plot using the classical methods.

With a reference to the second approach related to transformation of the data into low-dimensional space, Rousseeuw, Hubert and Segaert (2014) proposed to compute for each observation a *bagdistance* w.r.t. considered distributions (set of objects from different class). In a non-functional case, the bagdistance of an arbitrary point $\mathbf{x} \in \mathbb{R}^p$ to a sample $\mathbf{X}^n$ is defined as $d_b(\mathbf{x}, X_n) = \frac{\|\mathbf{x} - \Theta\|}{\|\mathbf{z} - \Theta\|}$, where Θ is a Tukey median and $\mathbf{z}$ is intersection between the depth contour containing 50% of most central observations and the line connecting $\mathbf{x}$ and Θ. In the functional setting they proposed the *integrated bagdistance*:

$$id_b(\mathbf{X}, \mathbf{Y}) = \sqrt{\int_U d_b(\mathbf{X}(t), \mathbf{Y}(t))^2 \, dt}. \tag{3}$$

For two classes this leads to the *distance-distance plot* (see Figure 4). Finally the original functional data are transformed into so called bagspace (Rousseeuw, Hubert and Segaert, 2014). The proposed minimal bagdistance classifier assignes a new functional object to the class, for which the bagdistance is the smallest. In turn, the bagspace classifier is based on classification in bagspace using such well-known techniques as linear/quadratic discriminant analysis or k-nearest neighbor classifier.

Paindaveine and Van Bewer (2013) proposed an interesting approach to the classification based on the maximization of the local depth, computed w.r.t. some neighborhood of an object (independent for each class). The maximum local depth classifier assigns a new object to the class, in which its local depth is the highest. Extending this concept into functional case, we propose functional maximum β-local depth classifier, according to the following steps:

1. In each class define the neighborhood of the new functional observation $\mathbf{X}$, containing fraction β of observations closest to $\mathbf{X}$ (in a sense of e.g. functional L^2-distance – see formula (1)).
2. Compute the depth of $\mathbf{X}$ with respect to each denoted neighborhood.
3. Classify $\mathbf{X}$ to the class, w.r.t. which neighborhood the depth of $\mathbf{X}$ is the highest.

3. Comparison of classifiers

In order to compare properties of the described classifiers in a context of management of the Internet service, we have taken under our considerations a real data set, containing information about activity of users of two popular Internet services in 2013 year. The activity of users is expressed both using a number of unique users (browsers) and number of page views generated by users on sites of these services. For each of these measures, the data set contains 730 functional observations (365 for each class - service). Each observation (corresponding to one day) was considered as a vector of values at discretized 24 time points (corresponding to consecutive hours from the 1 AM to the 12 PM).

The classification task was to assign a new univariate functional object to one of the services, based on number of unique users (the first part of analysis) and page views during the day and night (the second part). We have conducted such excercise with following classifiers:

1. *kNN* – the functional k-NN classifier based on functional L^2-distance, cross-validated number of neighbors k, see (Ferraty and Vieu, 2006) and {fda.usc} R package (Febrero-Bande and Oviedo de la Fuente, 2012).
2. *maxLD* – the functional maximum β-local depth classifier based on functional L^2-distance and Fraiman-Muniz depth ({fda.usc} R package), cross-validated fraction β of the closest curves, see proposition presented in section 2.3.
3. *maxD* – the functional maximum depth classifier based on Fraiman-Muniz depth, (see Ghosh and Chaudhuri, 2005).
4. *DD* – the functional *DD*-classifier based on Fraiman-Muniz depth and k-NN classification rule, cross-validated number of neighbors k, see (Li, Cuesta-Albertos and Liu, 2012) and {fda.usc} R package.
5. *DD*$_{LS}$ – the functional *DD*-classifier based on location-slope transformation and k-NN classification rule, cross-validated number of neighbors k, see (Mosler and Mozharovskyi, 2014) and {ddalpha} R package (Lange, Mosler and Mozharovskyi, 2014).

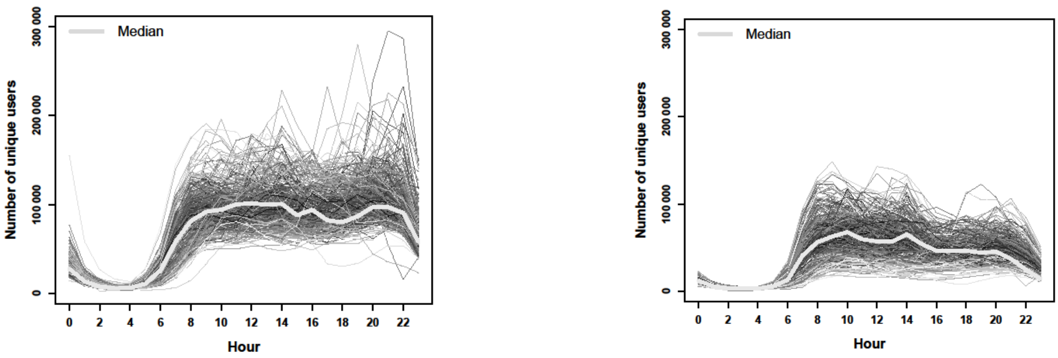

Fig. 5. Daily dynamics of number of unique users for the service 1 (left) and the service 2 (right).

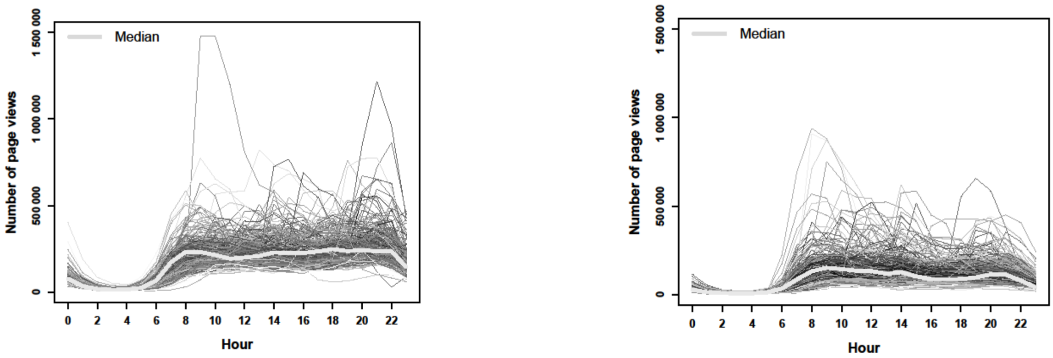

Fig. 6. Daily dynamics of number of page views during the day and night of the service 1 (left) and the service 2 (right).

6. *BagSpace* – the functional bagspace classifier based on k-NN classification rule, cross-validated number of neighbors k, see (Rousseeuw, Hubert and Segaert, 2014).
7. *BagDist* – the functional minimal bagdistance classifier based on functional L^2-distance, see (Rousseeuw, Hubert and Segaert, 2014).
8. *minDist* – the minimum distance to central curve classifier based on functional L^2-distance, cross-validated parameter of trimming β, see (Lopez-Pintado and Romo, 2005).

The quality of the classifiers was evaluated with 10-fold cross validation – the original sample was randomly partitioned into 10 equal size subsamples. For each of the 10 iterations one of the subsamples was considered as a test set and the remaining 9 subsamples as a training set. At the end of cross-validation process, the 10 results from the folds were averaged. Results of the classification for the unique users and the page views data sets are shown in Figures 7 and 8, respectively. Error rates are marked with black dot. Standard deviations of the misclassification rates of each classifier during cross-validation process are represented by symbol "X".

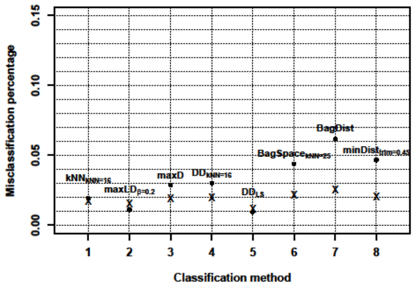

Fig. 7. Results of the conducted classification on the unique users data.

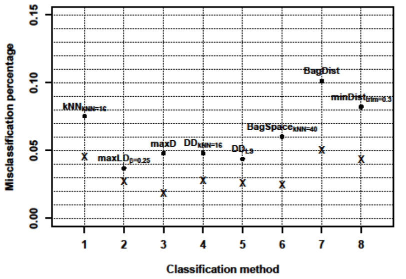

Fig. 8. Results of the conducted classification on the page views data.

Comparison of computational times for all classifiers is presented in Table 1. By a computational time we mean sum of time required for model training (additionally choice of optimal values of the parameters – number of neighbors, fraction β of the closest curves etc.) and time needed for the classification of test set (10% of the original sample size). The page views data set turned out to be more difficult to classify – the error rates and their standard deviations for all methods were higher then in the case of the unique users data. The likely reason of that are significant unusual changes in number of page views per day and even per hour described in Section 1.

kNN	**maxLD**	**maxD**	**DD**	**DD$_{LS}$**	**BagSpace**	**BagDist**	**minDist**
42	141	2	5	32	40	37	71

Table 1. Averaged computational times for the unique users data (in seconds, processor Intel Core 2 Duo 2.00GHz).

Discriminant analysis conducted on the unique users data resulted in the lower misclassification rates. Two classifiers achieved a good results – the *DD$_{LS}$* and the *maxLD*. The last one - proposed functional maximum β-local depth classifier turned out to be very competitive both in terms of accuracy and stability of the results. However, its disadvantage is relatively long time required for optimization of locality parameter β .

Conclusion

Classifiers based on *DD*-plot and maximum (local) depth classification rule achieved satisfactory results during classification both of considered data sets. However, it should be emphasized that particular high level of accuracy in case of the *DD$_{LS}$* and the *maxLD* classifiers is associated with time consuming computations.

In a context of Internet activity data, it turned out, that it is better to use for classification data less exposed to big, unusual changes resulting in significant amount of outlying functional observations. However, alternatively consideration both of mentioned data sets at the same time as a set of multidimensional functional observations may finally resulting in lower misclassification rate of those classifiers. In future studies authors are going to test that approach, especially in a context of proposed functional maximum β-local depth classifier.

Acknowledgements

Authors thank for the financial support related to CUE Faculty of Management fund for the research potential preservation and the Polish National Science Center for financial support from grant UMO-2011/03/B/HS4/01138.

References

Cuevas, A., Febrero, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, *22*(3), 481-496.

Febrero-Bande, M., & Oviedo 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.

Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. *Test*, *10*(2), 419-440.

Ghosh, A. K., & Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, *32*(2), 327-350.

Kosiorowski, D., & Zawadzki, Z. (2014). DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena. *arXiv preprint arXiv:1408.4542*.

Kosiorowski, D. (2014). Functional Regression in Short Term Prediction of Economic Time Series. *Statistics in Transition New Series*, *14*(4).

Kosiorowski, D. (2012). Statystyczne funkcje głębi w odpornej analizie ekonomicznej. *Zeszyty Naukowe/Uniwersytet Ekonomiczny w Krakowie. Seria Specjalna, Monografie*, (208).

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

Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association*, *107*(498), 737-753.

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.

López-Pintado, S., & Romo, J. (2006). Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, *72*, 103.

Mosler, K., & Mozharovskyi, P. (2014). Fast DD-classification of functional data. *arXiv preprint arXiv:1403.1158*.

Paindaveine, D., & Van Bever, G. (2013). From depth to local depth: a focus on centrality. *Journal of the American Statistical Association*, *108*(503), 1105-1119.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional Data Analysis*.

Rousseeuw, P., Hubert, M., & Segaert, P. (2014). *Classification of multivariate functional data based on depth*. Status: published.

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