Functional boxplots for summarizing and detecting changes in environmental data coming from sensors

Elvira Romano, Antonio Balzanella
Dipartimento di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli, elvira.romano@unina2.it

Lidia Rivoli
Dipartimento di Matematica e Statistica, Universitá di Napoli Federico II

Abstract: Nowadays, environmental sensor networks produce a large amount of streaming time series whose storage, manipulation and indexing is impractical. In this work, we propose a new strategy for summarizing and describing this kind of data based on functional data representation. It discovers trends and potential anomalies by using an informative exploratory tool: the functional boxplot. Functional boxplots are introduced for conveying location and variability information. In addition, for detecting and illustrating variation a distance among functional boxplots is used.

Keywords: streaming time series, functional data, functional boxplot

1 Introduction

In a wide range of environmental applications, networks of sensors allow to record huge amounts of temporally ordered data. Often, the sampling frequency is very high and the monitored phenomenon is highly evolving. This involves that traditional temporal data mining methods, based on computationally intensive algorithms and requiring the storing of the whole dataset, become ineffective. Especially there is a remarkable delay between the recording of the data and the analysis results which can impact on decisional processes.

In order to deal with this issue, it is necessary to move from the traditional temporal data mining to the data mining of streaming time series which focuses on processing the incoming data on-line without requiring their storage.

Usually, algorithms for data streams mining update, in incremental and online way, the knowledge about data by means of synopses. These provide suitable summaries which are substantially smaller than their base dataset and allow to discard the data once they have been processed. In literature, several summarization techniques for streaming time series have been proposed (a wide review is available in (Mitsa T., 2010)). Some of these transform a streaming time series into a new one of reduced dimensionality, others use sampling, sketches, histograms.

In this paper we introduce an intuitive tool for visualizing and summarizing the behavior of multiple streaming time series, the Functional Box Plot (FBP). Originally defined for functional data, it is considered as a variable and used as synthesis of batches of the incoming multiple streaming time series. The monitoring of the evolution of the data has been performed through the comparison of the FBP variables using an appropriate distance measure, rather than analyzing the incoming recordings.

2 The three steps strategy

Let $y_i(t), \ \ i = 1, \ldots, n, t \in [1, \infty]$ a set of streaming time series made by real valued ordered observations of a variable $Y(t)$ in n sites, on a discrete time grid.

Our aim is to summarize and describe their changes in a streaming fashion by means of a comparison of functional boxplot variables. Functional boxplots are an informative explorative tool for functional data. We use them as variables of synthesis for the set of n streaming time series splitted in non overlapping windows and opportunely approximated by functional data. With this scope a three steps strategy is proposed.

The first step consists in splitting the incoming parallel streaming time series into a set of non overlapping windows $W_j, j = 1, \ldots, \infty$, that are compact subsets of T having size $w \in \Re$ and such that $W_j \bigcap W_{j+1} = \emptyset$. The defined windows frame for each $y_i(t)$ a subset $y_i^{w_j}(t) \ \ t \in W_j$ of ordered values of $y_i(t)$, called subsequence.

Following the FDA approach, we consider each subsequence ${y_i}^{w_j}(t)$ of $y_i(t)$ the raw data which includes noise information (Ramsay, J.E., Silverman, B.W., 2005). Then we determinate a true functional form $f_i^{w_j}(t)$, we call functional subsequence, which describes the trend of the flowing data, by using smoothing spline functions. For each W_j we have that all the subsequences $y_i^{w_j}(t) \ \ i = 1, \ldots, n$ follow the model:

$$y_i^{w_j}(t) = f_i^{w_j}(t) + \epsilon_i^{w_j}(t), \ t \in W_j \ \ i = 1, \ldots, n \tag{1}$$

where $\epsilon_i^{w_j}(t)$ are residuals with independent zero mean and $f_i^{w_j}(\cdot)$ is the mean function which summarizes the main structure of $y_i^{w_j}(t)$.

In a second step since we need to have a summary of the batched streaming time series, we compute functional boxplot variables for each batch. Functional boxplot(box-and-whisker diagram or plot) is an informative graphically tool for depicting functional data through their five-functions summaries. We consider them as a kind of quantitative variables in the functional setting.

In functional data analysis two different definition of boxplot exist. A first one makes use of the first two robust principal component scores, Tukey data depth and highest density regions (Hyndman R.J., Shang, H.L., 2010); a second one is based on center outward ordering induced by band depth for functional data (Sun Y., Genton G., 2011). We makes use of the second boxplot definition, that is a natural extension to the classical boxplot. It is defined starting by a concept which allows to order curves from center outward: the band depth BD (Lopéz-Pintado and Romo 2009).

Let $f_i^{w_j}(t), i = 1, \dots n$ be the collection of functional subsequences in a window W_j, $G(f_i^{w_j}) = \left\{(t, f_i^{w_j}(t)) : t \in W_j\right\}$ be the graph of the function $f_i^{w_j}(t)$, and

$$B\left(f_{i_1}^{w_j}, \dots, f_{i_k}^{w_j}\right) = \left\{\left(t, g_i^{w_j}(t)\right) | t \in W_j, \min_{r=1,\dots,k} f_{ir}^{w_j}(t) \leq g_i^{w_j}(t) \leq \max_{r=1,\dots,k} f_{ir}^{w_j}(t)\right\} \quad (2)$$

be the band in R^2 delimited by the k different curves $\left(f_{i_1}^{w_j}, f_{i_2}^{w_j}, \dots, f_{i_k}^{w_j}\right)$, obtained by computing the minimum and the maximum values for all t. Let $BD_n^{(m)}$ be the portion of bands obtained by $m = 1, \dots, M$ different curves containing the whole graph of $f_i^{w_j}(t)$ expressed by

$$BD_n^{(m)}(f_i^{w_j}) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_m \leq n} I\left\{G\left(f_i^{w_j}\right) \subset B\left(f_{i_1}^{w_j}, f_{i_2}^{w_j}, \dots, f_{i_m}^{w_j}\right)\right\} \quad m \geq 2 \quad (3)$$

where $I\{\cdot\}$ denote the indicator function.

Thus the band depth $BD_{n,M}(f_i^{w_j}(t))$of any of these function $f_i^{w_j}(t)$ is defined as

$$BD_{n,M}(f_i^{w_j}) = \sum_{m=2}^{M} BD_n^{(m)}\left(f_i^{w_j}(t)\right) \quad M \geq 2 \quad (4)$$

Especially let $f_{[i]}^{w_j}(t)$ denote the sample of functional subsequence associated to the ith largest band depth value, the set $f_{[1]}^{w_j}(t) \dots, f_{[n]}^{w_j}(t)$ are order statistics, with $f_{[1]}^{w_j}(t)$ the median curve, that is the most central curve (the deepest), and $f_{[n]}^{w_j}(t)$ is the most outlying curve. Moreover the central region of the boxplot is defined as

$$C_{0.5} = \left\{(t, f^{w_j}(t)) : \min_{r=1,\dots,[n/2]} f_{[r]}^{w_j}(t) \leq f^{w_j}(t) \leq \max_{r=1,\dots,[n/2]} f_{[r]}^{w_j}(t)\right\} \quad (5)$$

where $[n/2]$ is the small integer not less than $n/2$. The border of the 50% central region is defined as the envelope representing the box of the classical boxplot.

Based on the center outwards ordering induced by band depth for functional data, the descriptive statistics of such functional boxplots FBP are: the upper $f_{[u]}^{w_j}(t)$ and lower $f_{[l]}^{w_j}(t)$ curves (boundaries) of the central region, the median curve $f_{[1]}^{w_j}(t)$ and the non-outlying minimum $f_{[b_{min}]}^{w_j}(t)$ and maximum boundaries $f_{[b_{max}]}^{w_j}(t)$.

For each window we have a FBP variable that is considered as a variable compound of five sub functions with the following structure:

$$\left\{f_{[u]}^{w_j}(t), f_{[l]}^{w_j}(t), f_{[1]}^{w_j}(t), f_{[b_{min}]}^{w_j}(t), f_{[b_{max}]}^{w_j}(t)\right\} \quad (6)$$

The third and latest step, consists in monitoring the evolution of the multiple data streams by comparing functional boxplot variables. With this aim we introduce a distance measure between a pair of FBP variables. It is a Manhattan distance which extends the distance for classical boxplot introduced in Arroio J., Mat C., Roque A. (2006) to functional boxplot variables. It is computed by considering that each couple of correspondent functions is compared on the same time interval W by means of a transformation of the functions domain. Thus, the Manhattan distance between a pair of functional boxplot FBP_1, FBP_2 opportunely shifted is:

$$\begin{aligned} d(FBP_1, FBP_2) \; = \; & \left| \int_{t \in W} (f^{'w_1}_{[u]}(t) - f^{'w_2}_{[u]}(t))dt \right| + \left| \int_{t \in W} (f^{'w_1}_{[l]}(t) - f^{'w_2}_{[l]}(t))dt \right| + \\ & + \left| \int_{t \in W} (f^{'w_1}_{[1]}(t) - f^{'w_2}_{[1]}(t)dt) \right| + \left| \int_{t \in W} (f^{'w_1}_{[b_{min}]}(t) - f^{'w_2}_{[b_{min}]}(t))dt \right| + \\ & + \left| \int_{t \in W} (f^{'w_1}_{[b_{max}]}(t) - f^{'w_2}_{[b_{max}]}(t))dt \right| \end{aligned}$$

where $f^{'w_j}_{[u]}(t), f^{'w_j}_{[l]}(t), f^{'w_j}_{[1]}(t), f^{'w_j}_{[b_{min}]}(t), f^{'w_j}_{[b_{max}]}(t)$ are the descriptive functions of the shifted FBP. The synthesis obtained by the FBP allows to have a description of batched streaming time series that can be compared on different time interval, thus this distance can be applied also on different and non consecutive time windows.

3 Concluding remarks

In this paper we have introduced a new strategy for summarizing multiple streaming time series and for monitoring their evolution. Unlike approaches existent in streaming time series literature, we have introduced a tool able also to provide an intuitive graphic summarization of data.

We have performed several tests on climate data in order to assess the effectiveness of the method. Preliminary results are encouraging.

References

Arroyo J., Mat C., Roque A. (2006) *Hierarchical clustering for boxplot variables*, Studies in Classification, Data Analysis, and Knowledge Organization, Part II, 59-66.

Hyndman R.J., Shang, H.L. (2010) Rainbow plots, bagplots and boxplots for functional data, *Journal of Computational and Graphical Statistics*, 19(1), 29-45.

Lopez-Pintado S., Romo, J. (2009). On the Concept of Depth for Functional Data. Journal of the American Statistical Association, 104, 718-734.

Mitsa T. (2010) Temporal Data Mining. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC

Ramsay, J.E., Silverman, B.W. (2005) *Functional Data Analysis* (Second ed.).Springer.

Sun Y., Genton M.G. (2011) Functional boxplots. *Journal of Computational and Graphical Statistics*. To appear.