A Clustream strategy for Functional Boxplots on multiple streaming time series

Antonio Balzanella and Elvira Romano

Abstract In this paper we propose a micro-clustering strategy for Functional Boxplot variables defined on multiple streaming time series splitted in non overlapping windows. It is a two step strategy. In the first step it performs an on-line summarization keeping updated the set of functional data structures, named Functional Boxplot micro-clusters; in the second step it reveals the final summarization by processing, off-line, the Functional Boxplot micro-clusters. Thus a new definition of micro-cluster based on using the Functional Boxplot as the centroid is proposed. Moreover a proximity measure which allows to allocate the data to the new defined micro-clusters is defined. This will allow to get a graphical summarization of the streaming time series by five functional basic statistics. The obtained synthesis will be able to keep track of the dynamic evolution of the streams.

Key words: Streaming time series, CluStream, Functional Boxplot

1 Introduction

Data stream mining has gained a lot of attention due to the development of applications where sensor networks are used for monitoring physical quantities such as electricity consumptions, environmental variables, computer network traffic. In these applications it is necessary to analyze potentially infinite flows of temporally ordered observations which cannot be stored and which have to be processed using reduced computational resources. The on-line nature of these data streams require the development of incremental learning methods which update the knowledge

Antonio Balzanella
Department of European and Mediterrean Studies, Second University of Naples, Caserta, Italy

e-mail: antonio.balzanella@unina2.itElvira Romano
Department of European and Mediterrean Studies, Second University of Naples, Caserta, Italy,e-mail: elvira.romano@unina2.it

about the monitored phenomenon every time a new observation is collected. Among the exploratory tools for data stream processing, clustering methods are widely used in order to get a summarization of data. The idea is to store for each group of similar observations, a single p-dimensional (where p is the number of data streams) data point which is, usually, the cluster centroid. According to this approach, the data collected by sensors are at first processed in order to update the cluster centroids and then discarded. Following this schema, the CluStream algorithm proposed in [1], provides a two-step strategy. The first one performs a first on-line summarization keeping updated a set of data structures named micro-clusters; the second one reveals the final summarization by processing, off-line, the micro-clusters. The CluStream provides only a rough summarization of the data coming from sensors since it only records the average and the variance of groups of similar multidimensional items.

In this paper we extend the Clustream algorithm by using the concept of Functional Boxplot variables (originally introduced in the framework of streaming time series in [4]). It is a micro-clustering strategy on Functional Boxplot variables defined on multiple streaming time series splitted in non overlapping windows. A new definition of micro-cluster based on the use the Functional Boxplot as the centroid of the cluster is proposed. Moreover a proximity measure which allows to allocate the data to the new defined micro-clusters is defined. The advantage of the proposed method consists to gain knowledge from multiple streaming time series by a periodical synthesis of five basic functional statistics (first and third quartile, median, maximum and minimum value) that can be graphically represented. This will allow to get a not only a clear compact description of the stream trends but also finer summarization of the entire set of streaming time series.

2 CluStream of Functional Boxplots

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 main aim is to supply a compact data description or synopsis to reduce dimensionality, and process each example in constant time and memory, in order to keep track of the dynamic evolution of the streams. Thus, we propose a Clustream algorithm on functional boxplots of a set of n streaming time series splitted in non overlapping windows.

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 functional subsequence.

Following the FDA approach, a summary of the batched streaming time series is given by the functional boxplot variables [5] for each batch. Especially let $y_{[i]}^{w_j}(t)$ denote the sample of functional subsequence associated to the ith largest band depth value, the set $y^{w_j}_{[1]}(t)\dots,y^{w_j}_{[n]}(t)$ are order statistics, with $y^{w_j}_{[1]}(t)$ the median curve, that is the most central curve (the deepest), and $y^{w_j}_{[n]}(t)$ is the most outlying curve. Moreover the central region of the boxplot is defined as

$$C_{0.5} = \left\{(t, y^{w_j}(t)) : \min_{r=1,\dots,[n/2]} y^{w_j}_{[r]}(t) \leq y^{w_j}(t) \leq \max_{r=1,\dots,[n/2]} y^{w_j}_{[r]}(t)\right\} \qquad (1)$$

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 outward ordering induced by band depth for functional data, the descriptive statistics of a functional boxplot are: the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope. Thus, the descriptive statistics of such functional boxplots FBP are: the upper $y^{w_j}_{[u]}(t)$ and lower $y^{w_j}_{[l]}(t)$ curves (boundaries) of the central region, the median curve $y^{w_j}_{[1]}(t)$ and the non-outlying minimum $y^{w_j}_{[b_{min}]}(t)$ and maximum boundaries $y^{w_j}_{[b_{max}]}(t)$.

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

$$\left\{y^{w_j}_{[u]}(t), y^{w_j}_{[l]}(t), y^{w_j}_{[1]}(t), y^{w_j}_{[b_{min}]}(t), y^{w_j}_{[b_{max}]}(t)\right\} \qquad (2)$$

Our method is based on allocating the functional boxplots computed on each window to specific data structures which we name FBP-micro-clusters.

A FBP-micro-cluster, similarly to the micro-cluster in CluStream, stores several statistics about data.

In particular it stores the set of functions $\left\{y^{k}_{[u]}(t), y^{k}_{[l]}(t), y^{k}_{[1]}(t), y^{k}_{[b_{min}]}(t), y^{k}_{[b_{max}]}(t)\right\}$ (where k is the index of the FBP-micro-cluster) defining the functional boxplot which assumes the role of centroid, the number of allocated functional boxplots n^* and a threshold value th. In the on-line step, every time the data of a new window W_j become available, a Functional Boxplot is constructed and then allocated to a FBP-micro-cluster. The allocation is obtained evaluating the distance between the FBP and the centroid of each FBP-micro-cluster so that if the minimum value of distance is lower than a threshold value th, the allocation is performed to the corresponding FBP-micro-cluster, otherwise a new one is started setting the functional boxplot of the window as centroid and n^* to 1.

The allocation is based on the definition of an appropriate distance measure for comparing functional boxplots. We propose to use an Euclidean distance for functional boxplot variables. It is computed by considering that each couple of correspondent functions is compared on the same time interval by means of a transformation of the functions domain. Thus, the distance between a pair of functional boxplots FBP_1, FBP_2 is:

$$
\begin{aligned}
d(FBP_1, FBP_2) = & \sqrt{\int_{t\in W} (y_{[u]}^{'w_1}(t) - y_{[u]}^{'w_2}(t))^2 dt} + \sqrt{\int_{t\in W} (y_{[l]}^{'w_1}(t) - y_{[l]}^{'w_2}(t))^2 dt} + \\
& + \sqrt{\int_{t\in W} (y_{[1]}^{'w_1}(t) - y_{[1]}^{'w_2}(t)^2 dt)} + \sqrt{\int_{t\in W} (y_{[b_{min}]}^{'w_1}(t) - y_{[b_{min}]}^{'w_2}(t))^2 dt} + \\
& + \sqrt{\int_{t\in W} (y_{[b_{max}]}^{'w_1}(t) - y_{[b_{max}]}^{'w_2}(t))^2 dt}
\end{aligned}
$$

where $\left\{ y_{[u]}^{'w_j}(t), y_{[l]}^{'w_j}(t), y_{[1]}^{'w_j}(t), y_{[b_{min}]}^{'w_j}(t), y_{[b_{max}]}^{'w_j}(t) \right\}$ are the corresponding shifted functions.

The consequences of an allocation are the unitary increment of n^* and the computation of the FPB-micro-cluster centroid. The latter is performed so that the average is kept for each of the five functions which define the FBP. This can be obtained starting from the information stored in the FBP-micro-cluster self and the just allocated FBP.

In order to reveal the final summarization of the streams, the on-line process can be stopped at any time so that the status of the FBP-micro-clusters can be used as input of the off-line phase.

The centroids of the FBP-micro-clusters become the items to be processed by a k-means like algorithm which uses the distance defined above and the average of FBP as centroid. These ones are the final output of the whole procedure.

3 Concluding remarks

In this paper we have introduced a new Clustream strategy for multiple streaming time series. Unlike the existent CluStream strategy in streaming time series literature, we have introduced a tool able also to provide a graphic synthesis.

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

References

1. Aggarwal, C. C., Han J., Wang, J., Yu, P. S.: A Framework for Clustering Evolving Data Stream. In Proc. of the 29th VLDB Conference.(2003)
2. Lopez-Pintado S., Romo, J. On the Concept of Depth for Functional Data. Journal of the American Statistical Association, **104**, 718–734, (2009).
3. Ramsay, J.E., Silverman, B.W.: Functional Data Analysis (Second ed.).Springer. (2005)
4. Romano, E., Balzanella, A., Rivoli, L.: Functional boxplots for summarizing and detecting changes in environmental data coming from sensors. In Electronic Proceedings of Spatial 2, Spatial Data Methods for Environmental and Ecological Processes 2nd Edition. Foggia, 1-3 Settembre 2011. .
5. Sun Y., Genton, M.G.: Functional boxplots. Journal of Computational and Graphical Statistics, **20**, 316-334. (2011).