Methods and tools for mining multivariate time series

De Gouveia da Costa Cachucho, R.E.

Chapter 3

Bipeline: a Web-based Visualization Tool for Biclustering of Multivariate Time Series

Ricardo Cachucho, Kaihua Liu, Siegfried Nijssen, Arno Knobbe

in Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2016

Abstract

Large amounts of multivariate time series data are being generated every day. Understanding this data and finding patterns in it became a contemporary and relevant task. To find prominent patterns present in multivariate time series, one can use biclustering, that is looking for patterns both in subsets of variables that show coherent behavior and in a number of time periods. For this, an experimental tool is needed.
Here, we present Bipeline*, a web-based visualization tool that provides both experts and non-experts with a pipeline for experimenting with multivariate time series biclustering. With* Bipeline*, it is straightforward to save experiments and try different biclustering algorithms, enabling users to intuitively go from pre-processing to visual analysis of biclusters.*

3.1 Introduction

The development of sensing technology lead to an explosion of sensor-based applications, commonly known as the internet of things. Such applications strive mainly due to factors such as, flexibility of design (smaller sensors), lower costs of production (cheaper sensors) and ease in terms of deployment and communication (interactive sensors). In many cases, such sensor systems measure complex phenomena without any sort of supervision, where variables are collected synchronously over time. As a result, there is an explosion of unsupervized multivariate time series.

As a motivating example take the case of a monitoring project for a highway bridge in the Netherlands [52, 102]. In this project, about 150 sensors were deployed on one span of the Hollandsebrug bridge, during an overall refurbish procedure to increase the life time of an important highway bridge. The intent of the project is to develop structural health monitoring methodologies for such a concrete bridge and find key performance indicators (KPIs) that could lead to a predictive maintenance. The fact is that the bridge is exposed to environmental elements (temperature, wind, rain, salt...) and is also subject to multiple events due to the traffic passing on the bridge. All these factors put together, result in a complex phenomena that were measured and materialized as multivariate time series. To discover patterns in such a multivariate and unsupervized setting, one would need sophisticated pattern recognition methods.

Pattern mining of multivariate time series is becoming highly relevant, both in scientific research and industrial applications. There are multiple tasks to deal with pattern mining for time series, such as segmentation and motif discovery. In the case of motif discovery, the task is set to find recurrent patterns over time. The motif discovery solutions normally focuses on the univariate case. Note that in the multivariate setting, not only recurrent patterns in one variable over time are relevant, but also relationships between multiple variables could provide useful insights. This task, is both clustering for time periods and variables, also know as biclustering [19, 61, 15].

Given a multivariate time series, it could be useful to try different biclustering algorithms. Also, one needs to optimize parameters across different steps, such as pre-processing, segmentation and biclustering itself. For each of these steps, there are many parameters to be optimized, leading to a large number of experiments. Furthermore, at each step, visual inspection is highly important for researchers to validate their findings. However, there is a lack of tools for this process.

We propose *Bipeline*, a web-based visualization tool that provides a pipeline for applying biclustering to multivariate time series. This tool is readily accessible to anyone via a web-based interface, allowing them to navigate through multiple experimental settings. Parameters can be interactively tuned, with web components such as checkboxes, sliders and drop-down menus. At each step of the biclustering process, feedback is provided be means of visualizations, with plots such as pre-processed time series, segmentation boundaries and biclusters. One or more biclusters can be plotted with a simple selection procedure.

3.2 Related Work

Until now, biclustering software tools with a graphical user interface have been developed to deal with biological gene expression data. BicOverlapper [82] is a tool for visual inspection of gene expression biclusters, introducing a novel visualization algorithm *Overlapper* to represent biclusters. Similarly, BiCluster Viewer [34] is a visualization tool for efficient and interactive analysis of large gene expression datasets. BicAT [7] implements multiple biclustering algorithms, for visualization and analysis of biclusters for expression data. BiGGEsTS [29] provides an environment for biclustering time series gene expression data.

All tools mentioned above integrate techniques for pre-processing and biclustering analysis, specifically for gene expression data. Their main purpose is to support biologists with the analysis and exploration of the gene expression data. However, these tools do not support biclustering analysis for multivariate time series. Also, most of them do not provide a pipeline experiment environment. Bipeline provides such a pipeline, where intermediate results can be inspected and saved. Using a friendly and interactive plotting environment, both non-experts and experts can pre-process, segment and analyze biclusters for multivariate time series.

Another class of tools are the machine learning experimental tool, where one can compare algorithms and decide on which is the best solution for a particular dataset. Examples of such tools are Weka [32], Moa [10] or KNIME [8]. From the multivariate time series perspective, the setback of such tools is that they are not tailored to experiment on biclustering or not tailored to analyze time series. On the other hand, Bipeline is an environment where compare multiple traditional biclustering algorithms and compare them to the biclustering algorithm proposed in the previous chapter.

3.3 Tool Overview

Bipeline is a web-based application that provides a pipeline to pre-process, segment and bicluster multivariate time series. An online version is available[1], which is compatible with all modern web browsers and across different client platforms. Both the user interface in the web browser and the server are implemented using R Shiny package [17]. In Figure 3.1, the system architecture illustrates the experimental pipeline and how each individual step relates to the other steps:

Importing Users can upload datasets and have a first view of the data table and descriptive statistics (*minimum*, *maximum*, *mean*, ...). This first inspection, although useful, is not enough to assess the quality of the data.

Plotting To gain further insight into the time series, it is crucial to have a visual inspection of the time series. The plotting panel includes multiple interactive plotting views, using a plotting R package dygraphs [97]. An example of these plots is illustrated in Figure 3.2. These interactive plots allow zoom in and out functionality, which is a highly desirable functionality for visual inspection of large time series.

This process of visual inspection is important across multiple phases of the biclustering process. This need for plotting is specially important in tasks that need human evaluation. Considering that there are multiple steps involved (pre-processing, segmentation and biclustering), Bipeline gives the user the flexibility to vizualize the data multiple times across different stages.

[1] http://fr.liacs.nl:7000

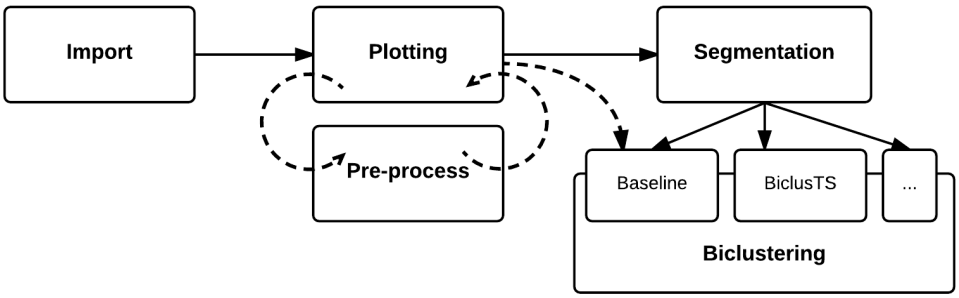

Figure 3.1: A overview of *Bipeline* architecture.

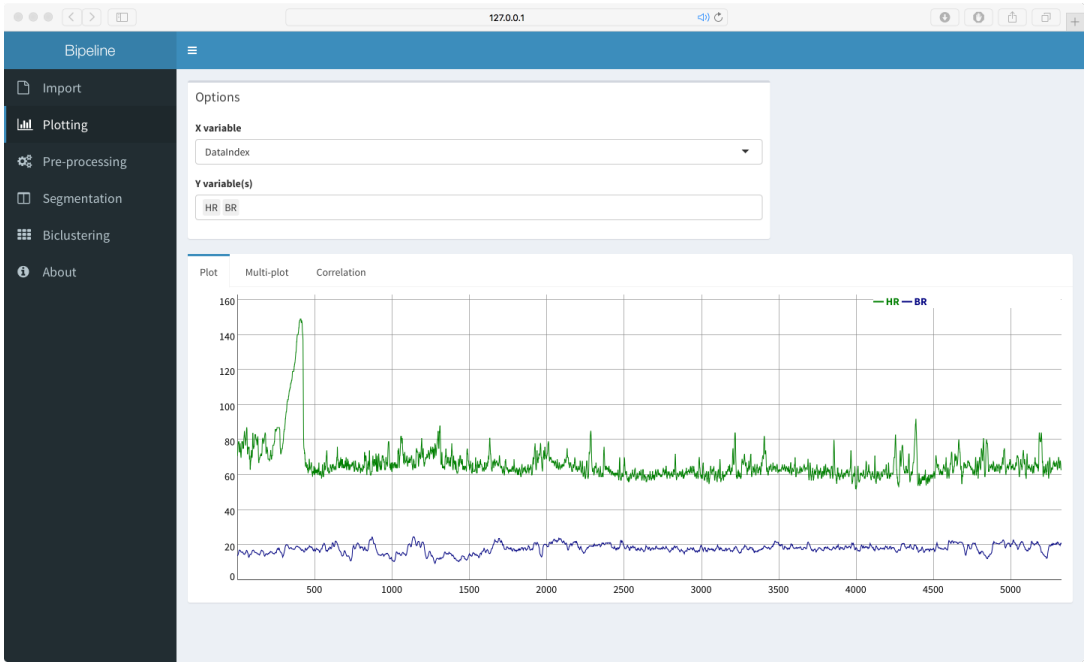

Figure 3.2: *Bipeline* user interface: Plotting menu.

Figure 3.3: *Bipeline* user interface: Pre-processing menu.

Pre-processing This panel allows preliminary handling of data such as: excluding variables, normalization, conditional removal and replacement of data, and outlier removal. Users can alternate between *plotting* (Figure 3.2), and *pre-processing* (Figure 3.3 until satisfied, then export the pre-processed data by clicking the *Save* button.

Segmentation This allows segmentation of the data, one of the steps necessary for the biclustering as suggested by the method presented in the previous chapter. By default, all variables share the same parameter settings: *window size*, *overlap* and *threshold* can be easily tuned. For greater flexibility, the user can dynamically create new tabs to set the parameters for individual variables. Additionally, a *minimum segment size* is customizable, and the tool will merge short segments to its most similar contiguous segment. Segmentation results can be visualized (Figure 3.4, saved and (re-)loaded, allowing the results to be used during the next step (biclustering) and over multiple user sessions.

Biclustering In *Bipeline*, we implement a number of biclustering algorithms, grouped in three categories. The *baseline algorithms* allow users to try well-known biclustering algorithms (e.g., Cheng & Church) [19, 61], that have been implemented using R package biclust [44]. *Segmentation + Baseline* biclusters the time series using an average representation of each segment, instead of using individual rows. *Segmentation + BiclusTS* is a novel algorithm [15] introduced to recognize similarities between segments, using probability density-difference estimation [86]. All biclusters are plotted in colored blocks, as shown in Fig.3.5. Users can select the biclusters they want to see, and the plot will respond with a real-time update.

The consideration of multiple biclustering algorithms in the tool, give the possibility to experiment different settings and interpret the different results visually. These different experimental options, place *Bipeline* in the group of tools that can be used for extensive biclustering experiments.

To allow extensive experiments, multiple features are shared by both *Segmentation* and *Biclustering*. Plots and parameter tables from different experiments are kept in history, allowing users to navigate back and forth to compare results and optimize parameters. During computationally expensive tasks, the front-end displays a progress bar, while the back-end server is busy carrying out the calculations. Furthermore, interactive web components can be saved into images with a single click.

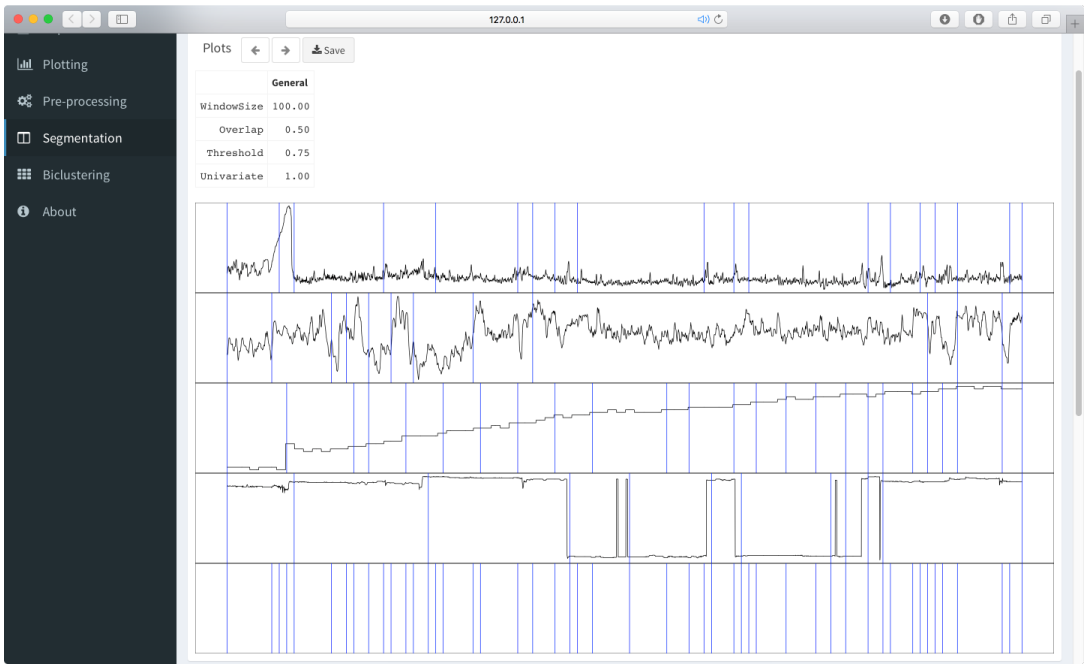

Figure 3.4: *Bipeline* user interface: Segmentation menu.

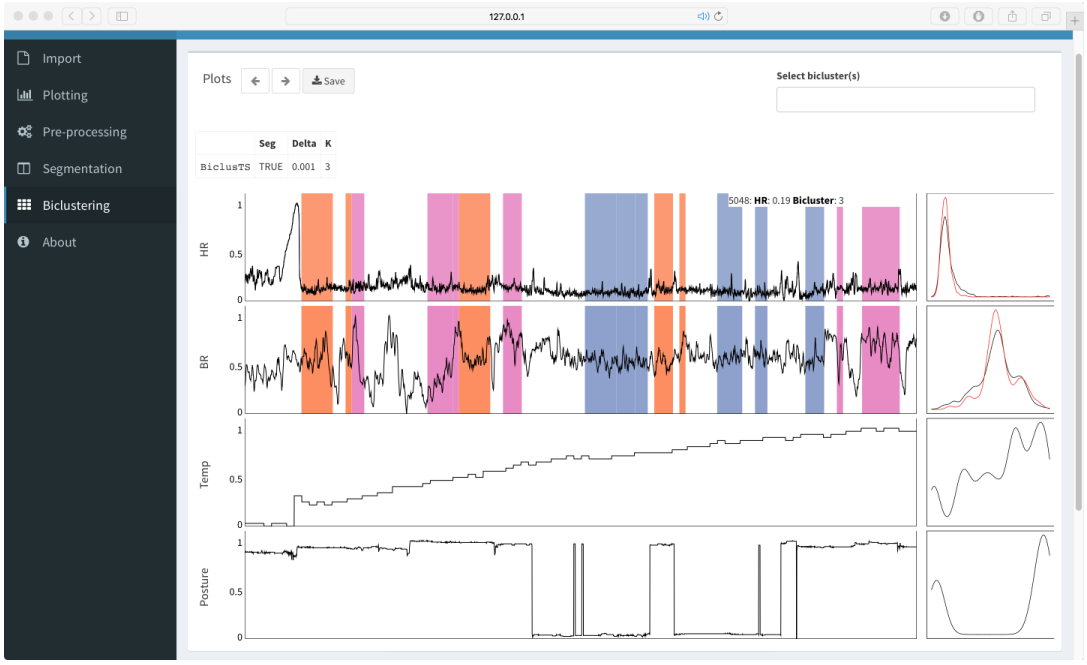

Figure 3.5: *Bipeline* user interface: Biclustering menu.

3.4 Conclusion

In this chapter we propose *Bipeline*, a web-based visualization tool, which provides a pipeline for applying biclustering to multivariate time series. Its main features include: visual inspection at multiple stages, interactive zoom in and out plotting, easy navigation, storage of results, and saving plots and experimental settings using a single click.

Bipeline's intuitive web-based design, makes it accessible both to experts and non-experts, and compatible across platforms. From a user perspective, the implementation of this tool can be found online[2]. Additionally to the online tool, for the users with a computer science background, both the biclusTS algorithm (see Chapter 2) and the tool implementation are open sourced and freely available[3].

[2]http://fr.liacs.nl:7000
[3]https://github.com/kainliu/Bipeline