Eurographics Conference on Visualization (EuroVis) (2015)
R. Maciejewski and F. Marton (Editors)

Enhancing Time Series Segmentation and Labeling Through the Knowledge Generation Model

T. Gschwandtner[1], H. Schumann[2], J. Bernard[3], T. May[3], M. Bögl[1], S. Miksch[1], J. Kohlhammer[3], and M. Röhlig[2]

[1]Vienna University of Technology, [2]University of Rostock, [3]Fraunhofer Institute for Computer Graphics Research (IGD)

Abstract

Segmentation and labeling of different activities in multivariate time series data is an important task in many domains. There is a multitude of automatic segmentation and labeling methods available, which are designed to handle different situations. These methods can be used with multiple parametrizations, which leads to an overwhelming amount of options to choose from. To this end, we present a conceptual design of a Visual Analytics framework (1) to select appropriate segmentation and labeling methods with appropriate parametrizations, (2) to analyze the (multiple) results, (3) to understand different kinds and origins of uncertainties in these results, and (4) to reason which methods and which parametrizations yield stable results and fine-tune these configurations if necessary.

Categories and Subject Descriptors (according to ACM CCS): G.3 [Mathematics of Computing]: Probabilities and Statistics—Time series analysis

1. Introduction

Segmenting and labeling of different activities in large time series data is relevant in many domains, for instance, reconstructing brain or heart activities from EEG or ECG data. To this end, a number of automated techniques exist: *feature-based* techniques [KP97, EAFT12b, EHD*02, KMN09], *pattern-based* techniques [Mül07], *model-based* techniques [EAFT12a, WWW11, BP66], or computational state space models [KNY*14]. However, an appropriate parametrization of these automated techniques is crucial. Sedlmair et al. [SHB*14] and Krause et al. [KPB14] describe the Visual Analytics (VA) approaches to explore such parameter spaces in other contexts. The aspects of (1) selecting appropriate methods, (2) identifying a suited parametrization, and (3) considering the uncertainties and data quality issues associated with these configurations, are strongly interrelated. Existing approaches, however, only focus on selected aspects in isolation. Moreover, domain experts need to investigate and fine-tune the input data, the selected segmentation and labeling methods, the parameters [RLS*14], and the results, in order to obtain a reliable segmentation.

Our contributions are (1) the conceptual design of a VA framework to integrate all important aspects, (2) the design in accordance with the knowledge generation model (KGM) for VA [SSS*14], and (3) the outline of interrelated aspects and important problems that need special consideration.

2. VA for Segmentation and Labeling of Time Series

To tackle the challenge of identifying an appropriate method with an appropriate parametrization as well as the corresponding uncertainties for a given time series and task, we apply the feedback loop of the KGM [SSS*14]. In a first step (i.e., **data preparation**), data must be gathered, cleansed [BRG*12, GAM*14], and an appropriate time scale must be selected. In the context of this abstract, however, we focus on the tasks of (1) **model building**:

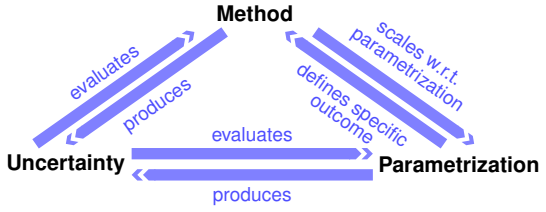

Figure 1: *Interdependencies of the chosen method, its parametrization, and resulting uncertainties.*

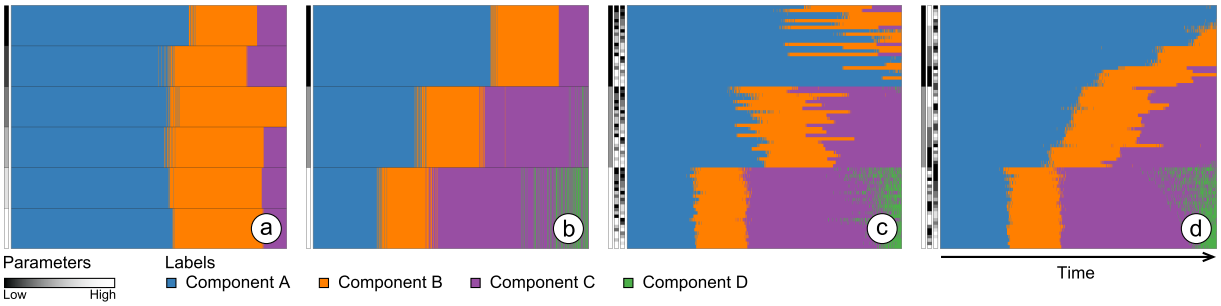

Figure 2: *This figure shows a time series derived from a simulation model with three parameters describing bio-chemical reactions [MRU11]. Different segments are distinguished by color and each horizontal line represents a segmentation and labeling result. (a) Changing parameter 3 only slightly influences the results. (b) Different values for parameter 1 lead to quite different results. However, in both cases, all components are predominant for a certain period of time. This is not true for testing parameter combinations as shown in (c). Arranging the results by similarity (d) illustrates the complex influences of individual parameters.*

selecting appropriate segmentation and labeling methods, (2) **model usage**: identifying appropriate parametrizations of these methods, and (3) **model-vis mapping**: identifying different degrees of uncertainty that come with different methods and parametrizations (see Figure 1). A **visual mapping** is needed to communicate these aspects for a better informed decision process. **Manipulation of the visualization** is needed to support the user in steering the configuration as well as in **inspecting** the corresponding result. Interactions are needed to retrace which segmentation and labeling methods and which parametrizations significantly influenced the result, and which uncertainties are introduced at different stages.

2.1. Model Building: Selection of Methods

For selecting appropriate segmentation and labeling methods we face the following three problems: (1) data preprocessing and selecting appropriate methods need to be coordinated as they influence each other, (2) the number of segmentation and labeling methods and their implementations induces a great diversity of steerable parameters, and (3) the selection of effective segmentation and labeling methods requires the user to anticipate how they operate on the data. Thus, relating the output of respective methods with the complex, multivariate time series data input is desirable.

2.2. Model Usage: Parametrization

For finding an appropriate parametrization of the model, the user needs to (1) investigate the influence of different parameters on the result, (2) verify the parametrization's suitability for changed conditions, and (3) explore introduced uncertainties to evaluate the quality of the segmentation and labeling results. Figure 2 represents parameter combinations. It clearly shows that the global optimum based on parameter combinations does not necessarily comply with the local optimum that is based on only one parameter change.

2.3. Model-Vis Mapping: Uncertainties

One important aspect that needs to be considered is that the results may comprise different kinds of uncertainties at different levels: (1) some segmentation and labeling methods, such as HMMs [BP66], provide a number of alternative segmentation labels of the time series together with their probabilities, (2) uncertainties introduced when composing alternative results from different methods and parametrizations into a final result (3) uncertainties of temporal boundaries of segments, and (4) uncertainties of which segmentation and labeling methods and which parametrizations have caused the result to what extend. These uncertainties need to be communicated to the user for him/her to be able to evaluate and fine-tune the used methods and parametrizations according to the KGM feedback loop.

3. Discussion and Further Work

We describe the challenges of time series segmentation and labeling which comprises different tasks: identifying one or more appropriate segmentation and labeling methods, instantiating them with appropriate parametrizations, investigating different kinds of associated uncertainties, and reasoning about the results. While existing approaches focus only on selected aspects, we argue for a VA approach to integrate them according to the KGM [SSS*14]. Our conceptual framework combines these tasks into one workflow with immediate visual feedback, which is essential to conduct an informed steering of the configuration, and eventually, to achieve an appropriate labeled segmentation. We envision many potential applications for our approach, including bio-chemical analysis, medical analysis, factory sensor analysis, or cyber-security.

References

[BP66] BAUM L. E., PETRIE T.: Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics 37*, 6 (1966), 1554–1563. doi:10.1214/aoms/1177699147. 1, 2

[BRG*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-Interactive preprocessing of time series data. In *Proc. of SIGRAD 2012: Interactive Visual Analysis of Data* (2012), pp. 39–48. URL: http://www.ep.liu.se/ecp/081/006/ecp12081006.pdf (accessed: 2015-02-18). 1

[EAFT12a] ESMAEL B., ARNAOUT A., FRUHWIRTH R., THONHAUSER G.: Improving time series classification using Hidden Markov Models. In *Proc. of the 12th International Conference on Hybrid Intelligent Systems (HIS)* (2012), pp. 502–507. doi:10.1109/HIS.2012.6421385. 1

[EAFT12b] ESMAEL B., ARNAOUT A., FRUHWIRTH R., THONHAUSER G.: Multivariate time series classification by combining trend-based and value-based approximations. In *Computational Science and Its Applications (ICCSA 2012)*, vol. 7336 of *Lecture Notes in Computer Science*. 2012, pp. 392–403. doi:10.1007/978-3-642-31128-4_29. 1

[EHD*02] EADS D. R., HILL D., DAVIS S., PERKINS S. J., MA J., PORTER R. B., THEILER J. P.: Genetic algorithms and support vector machines for time series classification. In *Proc. of SPIE 4787: Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation V* (2002). doi:10.1117/12.453526. 1

[GAM*14] GSCHWANDTNER T., AIGNER W., MIKSCH S., GÄRTNER J., KRIGLSTEIN S., POHL M., SUCHY N.: TimeCleanser: A visual analytics approach for data cleansing of time-oriented data. In *Proc. of the 14th International Conference on Knowledge Technologies and Data-Driven Business (i-KNOW 2014)* (2014), pp. 18:1–18:8. doi:10.1145/2637748.2638423. 1

[KMN09] KAMPOURAKI A., MANIS G., NIKOU C.: Heartbeat time series classification with support vector machines. *IEEE Transactions on Information Technology in Biomedicine 13*, 4 (2009), 512–518. doi:10.1109/TITB.2008.2003323. 1

[KNY*14] KRÜGER F., NYOLT M., YORDANOVA K., HEIN A., KIRSTE T.: Computational state space models for activity and intention recognition. A feasibility study. *PLoS ONE 9*, 11 (2014), e109381. doi:10.1371/journal.pone.0109381. 1

[KP97] KEHAGIAS A., PETRIDIS V.: Predictive modular neural networks for time series classification. *Neural Networks 10*, 1 (1997), 31–49. doi:10.1016/S0893-6080(96)00040-8. 1

[KPB14] KRAUSE J., PERER A., BERTINI E.: INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1614–1623. doi:10.1109/TVCG.2014.2346482. 1

[MRU11] MAUS C., RYBACKI S., UHRMACHER A. M.: Rule-based multi-level modeling of cell biological systems. *BMC Systems Biology 5*, 1 (2011), 166. doi:10.1186/1752-0509-5-166. 2

[Mül07] MÜLLER M.: Dynamic time warping. In *Information Retrieval for Music and Motion*. 2007, pp. 69–84. doi:10.1007/978-3-540-74048-3_4. 1

[RLS*14] RÖHLIG M., LUBOSCHIK M., SCHUMAN H., BÖGL M., ALSALLAKH B., MIKSCH S.: Analyzing parameter influence on time-series segmentation and labeling. In *Poster Proc. of the IEEE Visualization Conference (VIS)* (2014). URL: http://publik.tuwien.ac.at/files/PubDat_230765.pdf (accessed: 2015-02-18). 1

[SHB*14] SEDLMAIR M., HEINZL C., BRUCKNER S., PIRINGER H., MÖLLER T.: Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2161–2170. doi:10.1109/TVCG.2014.2346321. 1

[SSS*14] SACHA D., STOFFEL A., STOFFEL F., KWON B. C., ELLIS G., KEIM D.: Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1604–1613. doi:10.1109/TVCG.2014.2346481. 1, 2

[WWW11] WANG P., WANG H., WANG W.: Finding semantics in time series. In *International Conference on Management of Data (SIGMOD)* (2011), pp. 385–396. doi:10.1145/1989323.1989364. 1