EuroVis Workshop on Visual Analytics (2014)
M. Pohl and J. Roberts (Editors)

A Visual Analytics Approach to Segmenting and Labeling Multivariate Time Series Data

Bilal Alsallakh[1], Markus Bögl[1], Theresia Gschwandtner[1], Silvia Miksch[1],
Bilal Esmael[2], Arghad Arnaout[2], Gerhard Thonhauser[2,3], Philipp Zöllner[2]

[1]Vienna University of Technology, Austria
[2]TDE Thonhauser Data Engineering GmbH, Austria
[3]University of Leoben, Austria

Abstract

Many natural and industrial processes such as oil well construction are composed of a sequence of recurring activities. Such processes can often be monitored via multiple sensors that record physical measurements over time. Using these measurements, it is sometimes possible to reconstruct the processes by segmenting the respective time series data into intervals that correspond to the constituent activities. While automated algorithms can compute this segmentation rapidly, they cannot always achieve the required accuracy rate e.g. due to process variations that need human judgment to account for. We propose a Visual Analytics approach that intertwines interactive time series visualization with automated algorithms for segmenting and labeling multivariate time series data. Our approach helps domain experts to inspect the results, identify segmentation problems, and correct mislabeled segments accordingly. We demonstrate how our approach is applied in the drilling industry and discuss its applicability to other domains having similar requirements.

Categories and Subject Descriptors (according to ACM CCS): G.3 [Probabilities and Statistics]: —Time series analysis I.5.2 [Pattern Recognition]: Design Methodology—Classifier design and evaluation

1. Introduction

A common problem in time series analysis is segmenting and labeling a composite time series into the sequence of units that compose it. This problem arises in several areas such as speaker diarization [TR06], brain activity analysis [PMML95, BMM11], and industrial process reconstruction [EAFT11]. For example, the units can represent different activities performed while drilling a borehole, where the time series data is composed of multiple sensor measurements recorded over the drilling process. The sequence of activities in such processes can be reconstructed by segmenting the sensor data into labeled intervals, enabling several possibilities for process analysis and optimization [Tho04].

Automated segmentation approaches (Sect. 2) often model each of the labels as a class using certain rules or parameters. As example, a rule-based system to segment drilling data involves multiple rules that determine the activity for a small time interval based on the respective sensor values. The designer of such algorithms needs to take several decisions about segmentation parameters and thresholds. Additionally, it is not always feasible to cover all possible cases that might take place during actual execution, especially when handling industrial processes that often exhibit new variations and exceptions. Moreover, automated methods might fail to handle missing data and outliers in the data, which impacts their accuracy.

We propose a Visual Analytics approach for improving automated segmentation and labeling of multivariate time series. Our approach (Sect. 3) uses familiar time-oriented visualizations and interactions to enable end users to inspect and correct segmentation results computed by automated algorithms. Furthermore, it allows running these algorithms with appropriate thresholds and parameters, based on the actual data behavior in specific time intervals. We demonstrate how this approach is implemented in a productive system and applied successfully in the drilling industry (Sect. 4). In Sect. 5 we discuss potential improvements and scalability limitations that need to be addressed in future work.

2. Related Work

Time series analysis has been a central topic in data mining and information visualization. We provide an overview of machine-learning algorithms proposed for segmenting multivariate time series data as well as visualization techniques for this type of data.

2.1. Time series segmentation and labeling

Existing segmentation methods can be divided into three main categories [XPK10]. *Feature-based methods* treat each data point or data window individually using the respective time series values as data features. Several classification techniques have been applied to assign labels to these data points such as decision trees and rule induction [EAFT12b], artificial neural networks [KP97], genetic algorithms [EHD*02], and support vector machines [KMN09]. *Pattern-based methods* use similarity measures to match a given time series against a group of predefined templates such as measures based on Euclidean distance and dynamic time warping [MÖ7, JJO11]. *Model-based methods* use generative models such as Hidden Markov Models (HMM) to model the temporality of the data as a sequence of observations. Several HMM-based techniques have been proposed to compute this sequence [WWW11, EAFT12a].

Automated techniques often depend on appropriate parameterization to produce the desired segmentation results. Several visualization techniques assist in choosing parameter values by providing insights into the time series behavior and actual value ranges and distribution.

2.2. Multivariate time series visualization

Line and area charts are among the most common representations of numeric time series [AMST11]. Several techniques were proposed to visualize multivariate time series data for different purposes. Aigner et al. [AMST11] provide an extensive survey of these techniques. ThemeRiver [HHWN02], Stacked Graphs [BW08], and Braided Graphs [JME10] display multiple time series in one plot using either stacking or superimposition. Horizon Graph [Rei08, HKA09] uses a compression technique to show area charts in a compact vertical space, while preserving the value resolution. This allows showing multiple time series in separate plots below each other. Beside visual representations, several interaction techniques were proposed to explore time series data such as VisuExplore [RMA*10]. ChronoLenses [ZCPB11] and SignalLens [Kin10] provide interactive lens techniques to support fluid Focus+Context exploration in high-frequency time series.

Machine learning experts use visualization mainly in the design phase to validate their assumption about the data and select reliable features for their algorithms. We propose using visualization in the runtime to allow end users to inspect and improve the results, as we explain next.

3. Visual Analytics Approach

The basic idea of our approach is to visualize the segmentation results computed by automated algorithms along with the time series data in one view (Fig. 1). This enables inspecting the results in relation with the data and investigating possible reasons for segmentation problems. Detailed analysis of the data and manipulation of the results are possible through interaction with this view and intertwining it with automated analysis, as we show next.

3.1. Automated analysis

As discussed in Sect. 2.1, several machine learning algorithms can be used for segmenting and labeling time series. Our approach is independent of the actual algorithms used, as it is mainly concerned with the raw data and segmentation results without imposing restriction on how these results are computed. Nevertheless, algorithm-specific parameters can still be adjusted via a dedicated view (Fig. 1c).

3.2. Interactive visualization

Our approach shows the *time series data* in multiple plots using appropriate techniques such as line charts (Fig. 1b). To save vertical space or to better reveal correlations, two variables can be superimposed in one plot if they are semantically related. Additional variables can be depicted in the plots besides or instead of the raw time series such as time-varying features extracted from the data.

The *segmentation results* are depicted as a sequence of colored stripes that encode the respective labels over time (Fig. 1d). The colors are chosen from a categorical color scale and, when applicable, assigned to match existing color conventions in the problem domain. As we discuss in Sect. 5, a dedicated rendering algorithm is needed to ensure minimum visibility of individual stripes.

A *time slider* (Fig. 1a) indicates the time interval being displayed. The above-mentioned views and the slider are synchronized to show the same time interval up on zooming and panning in the plots area or up on moving the slider.

Conventional interactions with the plots are possible such as reading values via tooltips and selecting a time range. Also, the vertical value ranges and aspect ratios of the plots can be adjusted individually to emphasize certain ranges.

The user can click on a segment to highlight the respective time interval in the raw data plots (Fig. 1e). This allows examining the time series behavior in this interval in order to check if the computed label is correct. The user can change this label manually using appropriate interaction. Moreover she may inquire why the automated methods computed a certain label. For example, if the segments are labeled using a rule-based classifier, she can check which rule was applied to compute this label. This enables matching this rule against the actual time series values in the respective interval.

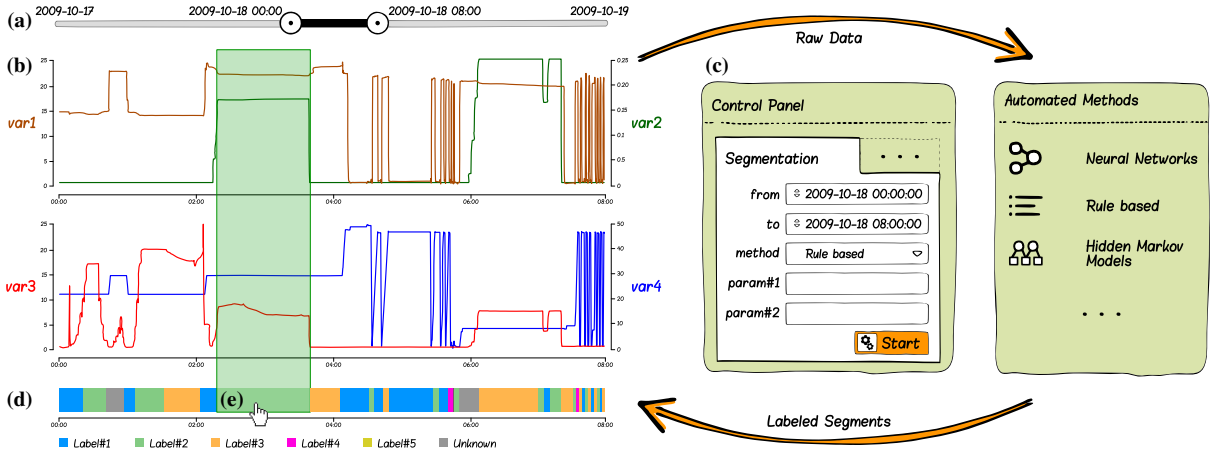

Figure 1: *The Visual Analytics approach to time series segmentation: both raw data (b) and results (d) are depicted for a certain time interval (a). Several interactions allow inspecting the results (e). The user can steer automated segmentation (c) for selected time interval. [Single elements used from Basiliq UI images kit by Cloud Castle, licensed under Creative Commons.]*

3.3. Steering automated time series segmentation

The proposed interactive visual interface provides several possibilities for inspecting the segmentation results and identifying mislabeled segments in the time series. Such labeling errors happen for various reasons. For example, a miscalibrated sensor might deliver values beyond the default thresholds, causing some classification rules to fail. Instead of rectifying these labels manually, it is possible to reapply the segmentation algorithm using different algorithms or thresholds. The user can adjust these thresholds based on gained insights into the erroneous data. Such adjustment can be entered in text fields, or using more advanced input methods, depending on the specific algorithm used. The visualization is updated interactively with the new segmentation results. This process can be repeated, until the user is satisfied with the results. Appropriate visual comparison techniques are needed to compare past and new results (Sect. 5). To avoid impacting previously correctly-labeled segments, changes to the segmentation results can be restricted to a certain time interval. This also reduces the computation time, especially for large time series data.

Integrating the segmentation algorithm with the interactive visual interface enables a Visual Analytics process following Keim's mantra [KMS*08]: *Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand*. After applying automated segmentation, the important is shown as colored segment with interactive means for gaining more details on demand. Further analysis is possible by interactively re-applying the automated algorithms. The approach presented so far is generic, as it imposes no restriction on applicable automated methods or visualizations. Additional components and design decisions might be needed for specific use cases, as we show next.

4. Use Case: Reconstructing Drilling Processes

A successful well construction process requires multiple activities to be performed by the on-site field personnel (drilling crew) in a specific sequence. For this purpose, the drilling rig runs through a certain sequence of states such as "drilling rotating", "reaming the hole", or "in slips". Logging these activities is crucial for many tasks such as planning and auditing, process optimization, as well as analyzing and predicting drilling problems [Tho04]. Manual logging is both unfeasible and inaccurate due to various circumstances involved in the drilling process and sometimes due to biased or misinformed human judgment. A feasible alternative is to reconstruct these activities based on measurements of several sensors mounted on the drilling rig, such as the depth of the drilling bit, the pressure of the mud pumps, or the position of the hook in the derrick. For this purpose, a set of 10 standard sensors are probed at a frequency of 1 to 0.1 Hz, resulting in multivariate numeric time series data containing about 10,000 samples every work day.

The Operations Detection system (ADPM) automatically detects rig activities out of available sensor data [Tho04]. In addition, the system enables drilling specialists to process daily drilling data efficiently using various software tools that implement industry standards to acquire the data. One of the major tasks these users perform over night is reconstructing drilling activities of the previous work day. For this purpose, they run ADPM's rule-based segmentation and labeling algorithm over the data, which computes the results in about one minute on an average computer. Due to frequent issues with the quality of the sensor data, such as dysfunctional or miscalibrated sensors, the automatic results need to be quality-controlled. Another reason for this are unexpected events such as stuck pipes or blowout prevention [AAF*12].

Automated algorithms can detect some of these issues, such as missing sensor data, and produce segments of unknown label accordingly. To enable users to examine such time intervals and identify further issues with the segmentation results, the ADPM system is extended with an interactive interface as described in Sect. 3. This interface enables expert users to employ their domain knowledge to inspect the results and identify unexpected events or erroneous labels.

A typical work day for Jane, an expert user of the system, involves loading specific data snapshots for inspection. Besides the charts of sensor data, she uses additional components and charts such as a time×depth plot of the drilling process. This chart helps her both to comprehend the data and to select certain ranges such as time intervals with slow or no drilling. The visualization provides her with information on segmentation results including intervals with unknown labels or with highly changing labels which indicate high uncertainty in the results. She checks these intervals in detail by inspecting the respective sensor values and decides to change some labels and assign "unknown" to certain intervals to avoid false positive detections. Jane also notices that certain activities are too long or too short for the specific drilling session and rigs she is analyzing, or appear in unusual times. By inspecting the sensor data, she noticed that "block position" values are drifted downwards in later work hours. Therefore, she decides to restart the segmentation with a different threshold for this sensor. Inspecting data of one day takes her a few minutes in which she adjusts the labeling for four hours of drilling (about 25% of the data). After Jane saves her changes, the quality-assurance manager loads the data and checks the new results to approve them, possibly asking for some modifications.

The described visualization has been applied successfully in the drilling industry in the past decade both to process and to analyze the sensor data involved [Tho]. It has been efficiently employed to produce daily activity reports and identify potential for optimizing drilling processes accordingly. Furthermore, a variety of subtle drilling problems were revealed by means of interactive visual analysis.

5. Discussion and Future Work

Several application domains can profit from the generic Visual Analytics approach to time series segmentation we propose. This applies when domain knowledge is essential for improving the results, but cannot be easily embedded in the automated methods. In such cases, appropriate visualizations and interactions enable domain experts to incorporate their knowledge by inspecting and adjusting the segmentation results. For example, a multi-speaker diarization application can present the results along with the signal features employed by the diarization algorithm in a graphical interface. The user can inspect the recorded conversation and adjust wrongly assigned speakers or misaligned segments, with help of computed features for each speaker.

Our approach provides basis for additional components to support analyzing and adjusting segmentation results. Possible extensions include visualizations that provide overview of data-quality issues such as outliers and missing values in multivariate time series data, as well as interactive methods to manipulate such data values before starting automated segmentation. Also, the segmentation stripe can be extended to show uncertainty in the result or to compare results from multiple algorithms. In case of rule-based segmentation, a dedicated view can be developed to enable domain experts to adjust existing rules or create new ones. Such a view should allow interactively changing the conditions or thresholds used in the rules and visually inspecting the results.

The proposed visualizations have certain scalability limitations in the number of variables, labels, and data points. About ten variables can be visualized as line charts in single or shared plots. Handling a larger number of variables requires space-efficient visualizations such as horizon graphs [JME10]. Conventional line charts with interactive zooming can handle time series having thousands of data points. To handle larger time series data, computational aggregation as well as Focus+Context exploration techniques are needed. Encoding segment labels in color allows for distinguishing up to 20 labels. Additional or alternative visual encodings are needed to handle larger number of labels. Furthermore, an appropriate visual aggregation technique is needed to handle time intervals exhibiting frequent changes in segment labels. This is important to insure the visibility of all labels appearing in a time interval when the number of segments is close to or exceed the interval's pixels.

6. Conclusion

Segmenting multivariate time series into labeled segments is a fundamental data analysis problem in several applications domains. Machine-learning algorithms can compute this segmentation efficiently for large time series data, but might produce erroneous segments and labels. Visualizing the segmentation results along with the raw time series data allows inspecting the results and analyzing why certain labels were assigned. Furthermore, integrating interactive visualization with automated segmentation allows steering the algorithms by choosing appropriate thresholds and parameter values for certain time intervals. We demonstrated how this approach is used to inspect and improve segmentation results in the drilling industry. This is done by enabling domain experts to incorporate their knowledge in order to adjust the segment labels or reject uncertain results. We also discussed how our approach can be extended in future work with additional and more scalable visual components.

Acknowledgement We thank Wolfgang Aigner for contributing several ideas to this work, and other colleagues at TDE for their cooperation. This work was supported by the Austrian Federal Ministry of Economy, Family and Youth via CVAST, a Laura Bassi Centre of Excellence (No. 822746).

References

[AAF*12] ARNAOUT A., ALSALLAKH B., FRUHWIRTH R., THONHAUSER G., ESMAEL B., PROHASKA M.: Diagnosing drilling problems using visual analytics of sensors measurements. In *IEEE Intl. Instrumentation and Measurement Technology Conference* (2012), IEEE, pp. 1750–1753. doi:10.1109/I2MTC.2012.6229708. 3

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*. Springer, London, UK, 2011. doi:10.1007/978-0-85729-079-3. 2

[BMM11] BRUNET D., MURRAY M. M., MICHEL C. M.: Spatiotemporal analysis of multichannel EEG: CARTOOL. *Computational intelligence and neuroscience 2011* (2011), 2. doi:10.1155/2011/813870. 1

[BW08] BYRON L., WATTENBERG M.: Stacked graphs – geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1245–1252. doi:10.1109/TVCG.2008.166. 2

[EAFT11] ESMAEL B., ARNAOUT A., FRUHWIRTH R. K., THONHAUSER G.: Automated system for drilling operations classification using statistical features. In *International Conference on Hybrid Intelligent Systems* (2011), IEEE, pp. 196–199. doi:10.1109/HIS.2011.6122104. 1

[EAFT12a] ESMAEL B., ARNAOUT A., FRUHWIRTH R. K., THONHAUSER G.: Improving time series classification using hidden markov models. In *International Conference on Hybrid Intelligent Systems (HIS)* (2012), IEEE, pp. 502–507. doi:10.1109/HIS.2012.6421385. 2

[EAFT12b] ESMAEL B., ARNAOUT A., FRUHWIRTH R. K., THONHAUSER G.: Multivariate time series classification by combining trend-based and value-based approximations. In *International Conference on Computational Science and Its Applications*. Springer, 2012, pp. 392–403. doi:10.1007/978-3-642-31128-4_29. 2

[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 *International Symposium on Optical Science and Technology* (2002), International Society for Optics and Photonics, pp. 74–85. doi:10.1117/12.453526. 2

[HHWN02] HAVRE S., HETZLER E., WHITNEY P., NOWELL L.: ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (Jan 2002), 9–20. doi:10.1109/2945.981848. 2

[HKA09] HEER J., KONG N., AGRAWALA M.: Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), ACM, pp. 1303–1312. doi:10.1145/1518701.1518897. 2

[JJO11] JEONG Y.-S., JEONG M. K., OMITAOMU O. A.: Weighted dynamic time warping for time series classification. *Pattern Recognition 44*, 9 (2011), 2231–2240. doi:10.1016/j.patcog.2010.09.022. 2

[JME10] JAVED W., MCDONNEL B., ELMQVIST N.: Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 927–934. doi:10.1109/TVCG.2010.162. 2, 4

[Kin10] KINCAID R.: Signallens: Focus+ context applied to electronic time series. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 900–907. doi:10.1109/TVCG.2010.193. 2

[KMN09] KAMPOURAKI A., MANIS G., NIKOU C.: Heartbeat time series classification with support vector machines. *Information Technology in Biomedicine, IEEE Transactions on 13*, 4 (2009), 512–518. doi:10.1109/TITB.2008.2003323. 2

[KMS*08] KEIM D. A., MANSMANN F., SCHNEIDEWIND J., THOMAS J., ZIEGLER H.: Visual analytics: Scope and challenges. In *Visual Data Mining*, Simoff S., Böhlen M., Mazeika A., (Eds.), vol. 4404 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008, pp. 76–90. doi:10.1007/978-3-540-71080-6_6. 3

[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. 2

[MÖ07] MÜLLER M.: Dynamic time warping. In *Information retrieval for music and motion* (2007), Springer, pp. 69–84. doi:10.1007/978-3-540-74048-3_4. 2

[PMML95] PASCUAL-MARQUI R., MICHEL C., LEHMANN D.: Segmentation of brain electrical activity into microstates: model estimation and validation. *Biomedical Engineering, IEEE Transactions on 42*, 7 (July 1995), 658–665. doi:10.1109/10.391164. 1

[Rei08] REIJNER H.: The development of the horizon graph. In *Electronic Proceedings of the VisWeek Workshop From Theory to Practice: Design, Vision and Visualization* (2008). 2

[RMA*10] RIND A., MIKSCH S., AIGNER W., TURIC T., POHL M.: Visuexplore: Gaining new medical insights from visual exploration. In *Proceedings of the International Workshop on Interactive Systems in Healthcare (WISH)* (2010), Hayes G. R., Tan D. S., (Eds.), pp. 149–152. doi:10.1007/978-3-642-25364-5_22. 2

[Tho] THONHAUSER G.: proNova. [Online; accessed Mar. 2014]. URL: http://www.tde.at. 4

[Tho04] THONHAUSER G.: Using real-time data for automated drilling performance analysis. *OIL GAS European Magazine, Edition 4* (2004), 170–173. 1, 3

[TR06] TRANTER S. E., REYNOLDS D. A.: An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on 14*, 5 (2006), 1557–1565. doi:10.1109/TASL.2006.878256. 1

[WWW11] WANG P., WANG H., WANG W.: Finding semantics in time series. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (2011), ACM, pp. 385–396. 2

[XPK10] XING Z., PEI J., KEOGH E.: A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter 12*, 1 (2010), 40–48. doi:10.1145/1882471.1882478. 2

[ZCPB11] ZHAO J., CHEVALIER F., PIETRIGA E., BALAKRISHNAN R.: Exploratory analysis of time-series with chronolenses. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (Dec 2011), 2422–2431. doi:10.1109/TVCG.2011.195. 2