Data-Adaptive Approximation Selection for Large Time-Series Visualization

Alberto Baggio A.BAGGIO@UMAIL.LEIDENUNIV.NL
Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Ugo Vespier UVESPIER@LIACS.NL
Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Arno Knobbe KNOBBE@LIACS.NL
Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Keywords: Model Selection, Time Series Visualization, Dimensionality Reduction

When dealing with large amounts of data measured from complex time-evolving systems, interactive time series visualization is an effective way to perform exploratory analysis and form an intuition of the system's behavior. Indeed, the human ability to process visual information helps to identify structure and patterns and permits to exploit prior knowledge when applying machine learning and data mining algorithms.

The main challenge when visualizing large time series is to maintain interactivity while allowing the user to quickly zoom in and retrieve detailed portions of the data. Moreover, since the data is plotted in a viewport with a pixel width typically smaller than then the number of points in the time series, some sort of approximation of the original data needs to be performed. The best approximation for this task is the one with the best trade-off between compression ratio and ability to preserve the important perceptual features in the data.

The literature contains many examples of approximation algorithms for time series (Fu, 2011) from frequency-domain methods, such as DFT and the DWT, to time-domain methods such as PAA and APCA. These are however focused on minimizing the Euclidean distance between the original data and the reduced one. There are few algorithms which actually take care of preserving the perceptual features of the original data. Douglas-Peucker (Hao & et al, 2011), PIP (Son & Anh, Oct) and Important Extrema (Fink & Gandhi, 2011) are the most widely cited. However, we show that, while at low compression ratios they model the data pretty well, their L_2 error becomes consistent at high compression ratios.

Considering these limitations, we claim that there is not a single existent technique which behaves well under the interactive visualization requirements depicted above. A good compromise would result from a smart combination of two or more approximations techniques.

We propose a method to select a data-adaptive hybrid approximation obtained by composing diverse techniques. In order to evaluate the reliability of our method, we also define a quality measure for the approximation which keeps into account both the $L2$ error and the capability of preserving important perceptual features. We evaluate our method over 3 months of sensor data (around 500 millions measurements) collected by a sensor network installed on the Hollandse Brug, in the context of the InfraWatch project[1].

References

Fink, E., & Gandhi, H. S. (2011). Compression of time series by extracting major extrema. *J. Exp. Theor. Artif. Intell.*, *23*, 255–270.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*, 164 – 181.

Hao, M. C., & et al (2011). A visual analytics approach for peak-preserving prediction of large seasonal time series. *Computer Graphics Forum*, *30*, 691–700.

Son, N. T., & Anh, D. T. (Oct.). An improvement of pip for time series dimensionality reduction and its index structure. 47–54.

[1] http://www.infrawatch.com