Internship - Scalable Time Series Analysis

- **Level**: Masters or Engineer level (specialized in computer science/statistics/applied mathematics)
- **Supervisors**: Florent Forest (forest@lipn.univ-paris3.fr), Hanane Azzag (hanane.azzag@lipn.univ-paris13.fr), Mustapha Lebbah (mustapha.lebbah@lipn.univ-paris13.fr)
- **Location**: Université Paris 13, LIPN - UMR 7030 - CNRS
- **Duration**: 5 or 6 months, starting March/April 2019 (flexible)

Context

Most real-world data has a temporal component, whether it is measurements of natural processes (weather, sound waves) or man-made (sensors, medical data, stock market). Analysis of time series data has been the subject of active research for decades and is considered as one of the top 10 challenging problems in data mining due to its unique properties. Unsupervised machine learning on time series has gained interest recently, but is still a research field, especially concerning multivariate time series. At the same time, new software technologies allow to process huge volumes of data (*Big Data*), and many of them are open-source. We propose to study the interface between these two research fields: large-scale analysis of multivariate time series data using unsupervised learning (clustering in particular).

The skills acquired during this internship will be extremely valuable for working both in research or in industry.

Subject

The internship will consist in 3 phases:

1. Study the current state-of-the-art on time series analysis, with a focus on unsupervised learning, clustering and visualization. Important topics are: distance/similarity measures; dimensionality reduction techniques; traditional mathematical techniques; supervised and unsupervised ML approaches for classification, regression, clustering and/or representation learning; deep learning approaches; etc.
2. Review the current state-of-the-art software tools and architectures for large-scale time series analysis. This includes: useful software packages and libraries to build, train, deploy models on time series data; data visualization tools; large-scale storage and databases. These tools should preferably be compatible with the distributed Hadoop ecosystem.
3. Based on the previous studies, implement one or several algorithms or tools that do not exist yet. Preferred languge is Scala, but Python can also be used. The results of the internship may lead to contributions to open-source, or even a scientific publication, depending on the intern's skills and motivation.

Example projects:

- Implement time series distances in the C4E project (Scala and Spark) such as DTW (Dynamic Time Warping), Complexity-Invariant Distance (CID), Shape-Based Distance (SBD)...
- Design and implement a distributed version of the k-shape clustering algorithm based on Spark.
- Develop an application for efficient time series visualization.

Required skills

- Solid mathematical background (Bs/Ms), in particular applied mathematics and signal processing.

- CS background: algorithms, complexity theory.
- Good imperative and object-oriented programming skills. Functional programming would be a plus. Proficiency in at least one multi-purpose language, including (but not limited to): Scala (preferred), Java, Python, C++, Go…
- Knowledge of one or more of the following technologies: Apache Spark (preferred); Hadoop; SQL/NoSQL databases; version control (git); docker. Scientific computing and ML libraries: Python (numpy, pandas, scikit-learn…), Scala (breeze, smile…), deep learning frameworks (TensorFlow, (Py)Torch, MXNet, DL4J…).

References

Software projects:

- C4E clustering library (developed at LIPN): https://github.com/Clustering4Ever/Clustering4Ever
- spark-timeseries (no longer developed): https://github.com/sryza/spark-timeseries

Datasets:

- UCR Time series data sets: https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

Bibliography:

- Batista, G. E. A. P. A., Wang, X., & Keogh, E. J. (2011). A Complexity-Invariant Distance Measure for Time Series. Proceedings of the 2011 SIAM International Conference on Data Mining, 699–710. https://doi.org/10.1137/1.9781611972818.60
- Giusti, R., & Batista, G. E. A. P. A. (2013). An empirical comparison of dissimilarity measures for time series classification. Proceedings - 2013 Brazilian Conference on Intelligent Systems, BRACIS 2013, 82–88. https://doi.org/10.1109/BRACIS.2013.22
- Paparrizos, J., & Gravano, L. (2015). k-Shape: Efficient and Accurate Clustering of Time Series. Acm Sigmod, 1855–1870. https://doi.org/10.1145/2723372.2737793
- Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42(1), 11–24. https://doi.org/10.1016/j.patrec.2014.01.008
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. https://doi.org/10.1016/S0925-5273(03)00047-1
- Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. Retrieved from http://arxiv.org/abs/1607.00148
- Nguyen, M., Purushotham, S., To, H., & Shahabi, C. (2017). m-TSNE: A Framework for Visualizing High-Dimensional Multivariate Time Series. Retrieved from http://arxiv.org/abs/1708.07942
- and many others…

Contact and other information

To apply, please send a resume and transcripts of your last two years to forest@lipn.univ-paris13.fr.
This is a paid internship: 577.50€/month.