Unsupervised clustering and symbolic representation of functional data, applied on plethysmograph signals for the study of respiratory behavior

Investigators.

- Dr Charles Truong (charles@doffy.net), Centre Borelli (CNRS, Université de Paris, ENS Paris-Saclay, SSA, INSERM).
- Dr Eric Krejci, Centre Borelli (CNRS, Université de Paris, ENS Paris-Saclay, SSA, INSERM).

Keywords. Unsupervised clustering, functional data, metrics between time series, pattern matching, representation learning.

Presentation of the laboratory. This internship will take place at the Centre Borelli (CNRS, Université de Paris, ENS Paris-Saclay, SSA, INSERM). The Borelli Center is a multidisciplinary research unit focusing on all applications of mathematics, neuroscience and biomedical research. It brings together multidisciplinary teams of mathematicians and experts in physics, mechanics, biology, and engineering, to conduct research actions motivated by real data and use cases. It offers courses of excellence in applied mathematics, neurosciences, ergonomy and bioengineering on promising themes with high technological and societal stakes.

Environnement. The trainee will have the opportunity to be fully integrated into an interdisciplinary team of clinical neurologists, neurophysiologists, ergonomists, statisticians and computer scientists. In case of success, the internship may be the subject not only of scientific publications but also of a valorization project. The internship will take place at University of Paris (45 rue des Saints Pères, 75006 Paris). The trainee will also attend weekly seminars of the MLMDA team.

Motivations. Pesticide exposure, through poisoning by organo-phosphate (OP) compounds and carbamates causes around 100,000 deaths/year worldwide. OP chemical warfare nerve agents, such as sarin, have also intoxicated people during the Iran-Iraq war (mid-1980), and in Syria (2013, 2017). Even low doses and short exposure to these irreversible inhibitors of acetylcholinesterase (AChE) and butyrylcholinesterase (BChE) induce well-characterized cholinergic syndrome including miosis, seizures, and apnea, and finally death. According to the European Medical Agency (2003), treating respiratory failure by controlled ventilation after intubation is the immediate therapeutic priority after poisoning by OP compounds. The objective of researchers is to better understand the respiratory failures caused by OP exposure.
To that end, several dozens mice with different genetic alterations of cholinesterase enzymes were monitored with a plethysmograph that recorded their respiratory airflow (that passes through the nostrum) at a frequency of 2 kHz (see [1] for more information). During the experiments, the mice were exposed to sub-lethal doses of toxic agents antagonist or agonist

of the cholinergic system (and thus affected their breathing). Researchers are particularly interested in the influence of those substances on the breathing behavior of the animals, which in turn can provide insight on which part of the respiratory system is attacked by OP compounds.
It should be noted that, while originally centered on OP exposures, this projects offers numerous perspectives on a general characterization of the breathing behavior in a large range of pathologies that affect the respiratory tract, for instance, infectious diseases such as the COVID19.

Objectives. Using well-known features from the literature (such as respiratory frequency, duration of apneic episodes, pauses after inspiration or before expiration) as well as automatically extracted ones, the objective is (i) to characterize each respiratory cycle, (ii) to quantitatively assess the evolution of the respiratory behavior from the previous characterization and (iii) to propose a relevant representation of this behavior across time. Those tasks require meaningful and efficient machine learning procedures since each collected signal lasts about one hour (~ 7 millions of points) and contains around 10,000 respiratory cycles. In addition to the associated biomedical literature, the candidate will need to extend his search to unsupervised methods for the clustering of functional data, pattern matching procedures, symbolic representation of signals, kernel methods for time series, etc.

Preliminary planning.

1. State of the art for the study of plethysmograph data [1], signal processing analysis (filtering, time-frequency analysis, dynamic time warping and its variants,...), clustering of functional data [2, 3] (1 month)
2. Review and implementation of standard unsupervised clustering procedures (K-means, one-class SVM, spectral clustering, etc.) using non-standard features for functional data (alignement kernels, shapelets, matrix profile, DTW) [4]. Tests on real data and first conclusions on the database. (2 months)
3. Study of the automatic selection of features. Study of the symbolic representation (using the previous clustering) of respiratory cycles and the evolution of the respiratory behavior across time. (2 months)
4. Publication of the results and redaction of the master thesis. (1 month)

References

[1] A. Nervo, A.-G. Calas, F. Nachon, and E. Krejci. Respiratory failure triggered by cholinesterase inhibitors may involve activation of a reflex sensory pathway by acetylcholine spillover. *Toxicology*, 424:152232, 2019.

[2] J. Paparrizos and L. Gravano. k-shape: efficient and accurate clustering of time series. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1855–1870, Melbourne, Victoria, Australia, 2015.

[3] F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011.

[4] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21:1–6, 2020.