Temporal phenotyping of patients from EHR data based on tensor decomposition

Master 2 Internship subject

2021–2022

Supervising environment

The project is proposed to contribute to the chair AI-RACLES funded by Inria-APHP-CS. Inria is the French national institute for digital science. APHP is the greater Paris university Hospital. And Central Supelec (CS) is a prestigious engineering school. AI-RACLES aims at developing artificial intelligence techniques to better exploit the APHP data lake to improve healthcare system and practices, especially for fragile patients.

The internship is proposed by two chair holders of AI-RACLES (Thomas Guyet and Pr. Etienne Audureau) and it will be supervised by:

- Thomas Guyet, Inria, Lyon `mailto:thomas.guyet@inria.fr`
- Pr. Etienne Audureau, APHP/UPEC, CEpiA (Clinical Epidemiology and Ageing), CHU Henri Mondor, `mailto:etienne.audureau@aphp.fr`
- Romain Tavenard, Univ. Rennes/LETG, `mailto:romain.tavenard@univ-rennes2.fr`

There will be opportunities for a funded PhD position after the internship.

Context

The APHP data lake is a huge Electronic Health Records (EHR) repository of the patients being admitted in one of the hospitals located in the greatest Paris. The database contains information about patient visits, including the care and drugs delivered along each of their visit (with their timestamps). For example, the APHP identified a cohort of more than 20,000 patients hospitalized during the Covid-19 crisis. A dataset was thus created from information on their condition and the care they received. This information constitutes their care pathway.

The main objective of the chair AI-RACLES is to develop new artificial intelligence techniques to analyze this data lake in order to address health questions. The context of this internship is to investigate how to support the evaluation of health care pathways. The notion of health care pathways denotes the sequence of cares of a patient being cured for a given disease. Quality assessment aims to identify the key characteristics of pathways which may likely leads to a positive outcome for the patient. For example, in the case of the Covid-19 crisis, it is interesting to identify the care strategies that would prevent patients from requiring intensive cares [3].

The first step to achieve this objective is to describe the actual care pathways. The APHP data lake gives us the opportunity to describe objectively the care pathways of patients from historical data. This internship aims to contribute to identifying the care pathways through the unsupervised or semi-supervised machine learning techniques.

Temporal phenotyping through tensor decomposition

The proposed research direction is the use of a powerful unsupervised machine learning technique called tensor factorization (or tensor decomposition). This generic techniques consists in decomposing a tensor $\mathcal{X}$ of dimension n into a collection of lower dimensional tensors $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ such that $X \approx \mathcal{Y}_1 \otimes \cdots \otimes \mathcal{Y}_k$ where $\otimes$ is a matrix product.

In the context of EHR data analysis, $\mathcal{X}$ is seen as a three-dimensional tensor whose dimensions are the patient identifier, the time and the medical events (procedures, labtests, drugs delivered. The decomposition of two dimensional tensors allow the identification of typical patient profiles (the medical events per patients), which are called phenotypes. A care pathway is then represented by the sequence of the phenotypes.

The problem of tensor decomposition is an old statistical problem for which statistical approaches have been proposed [6] since the early years of the past century. But in recent years, this problem is renewed on the light of machine learning, and neural networks. Several recent neural networks architecture have been proposed [7, 1, 9].

They proved the feasibility of the approach to decompose efficiently large and complex tensors. In parallel, the interest of phenotyping from EHR data has also been highlighted in the biomedical literature [3, 4, 8].

In this internship, we would like to investigate the notion of temporal phenotypes, and temporal phenotyping. Contrary to a phenotype that gives a combination of medical events at one time instant, a temporal phenotype describes a temporal arrangement of medical events. It is thus more expressive and may be useful to identify short term procedures that make the care pathways.

A similar objective is targeted by Emonet et al. [5, 2] with Temporal Analysis of Motif Mixtures (TAMM). The problem of identifying temporal phenotypes (topic models) is addressed by a non-parametric Bayesian model fitted using Gibbs sampling. One of the limitation of the proposal is the slowness and resources consumption of the solving technique, and a rigid model (modifying the model requires to derive a new sampler).

A starting point of the internship will be to adapt the model of TAMM to solve it using machine learning techniques and to evaluate it (from the efficiency and accuracy points of view). Then, the implemented model will be applied to extract temporal patient phenotypes from the APHP Covid-19 cohort data. and contribute to 1) describing Covid-19 patients, possibly by criticality group, and 2) describing hospitalizations by conditions (comparison of new and historical ICUs). A secondary objective is to investigate the possibility of using these models to create discriminant temporal phenotypes, *i.e.* phenotypes that would occur more likely in a group of patients that in the others.

The main steps of the project are the following:

- discovery of the data at hand (MIMIC/APHP Covid-19)
- additional bibliography
- experiments on existing approaches for patient phenotyping (primarily CNTF and TAMM)
- proposal and implementation of a new temporal phenotyping model
- experiment and comparison on synthetic/MIMIC datasets (accuracy, efficiency)
- consolidated implementation
- experiments on APHP Covid-19 dataset
- report writing

Candidate profile

- You are student in a Master 2 in computer science, data science or statistics, or student in a engineering school.
- You are enthusiastic about research, you love to understand in depth the problems and to find them elegant solutions.
- You have an strong background in math and computer science (Python for machine learning environment).
- You are interested in artificial intelligence and, more precisely, in machine learning, optimization techniques, data analysis, ...
- You have interest in the field of health and to contribute to the development of solutions that may help clinicians or epidemiologists.
- You speak and write English and/or French.

Practical information

- Location: Lyon (or possibly Paris).
  The intern will be hosted at Inria Lyon located on the Doua scientific campus, at Villeurbanne. Some meeting will be organized in Paris.
- Contacts: (see people above)
- Data access:
  - Covid-19 dataset is the primary dataset for this internship. Depending on the project progress, additional datasets may be investigated.
  - Access to the data is subject to a specific request to APHP and can only be done in the working environment provided by APHP
  - To facilitate the work, open sources datasets (e.g. MIMIC) may be used

References

[1] Ardavan Afshar, Ioakeim Perros, Evangelos E. Papalexakis, Elizabeth Searles, Joyce Ho, and Jimeng Sun. COPA: Constrained PARAFAC2 for sparse & large datasets. page 793–802, 2018.

[2] AH Aubert, Romain Tavenard, Rémi Emonet, Alban De Lavenne, Simon Malinowski, Thomas Guyet, René Quiniou, J-M Odobez, Philippe Mérot, and Chantal Gascuel-Odoux. Clustering flood events from water quality time series using latent dirichlet allocation model. *Water Resources Research*, 49(12):8187–8199, 2013.

[3] Mathieu Chambard, Thomas Guyet, Yên-Lan Nguyen, and Etienne Audureau. Temporal phenotyping for characterisation of hospital care pathways of covid19 patients. In *AALTD 2021-The 6th International Workshop on Advanced Analytics and Learning on Temporal Data*, 2021.

[4] A. Dagliati, L. Sacchi, A. Zambelli, V. Tibollo, L. Pavesi, J.H. Holmes, and R. Bellazzi. Temporal electronic phenotyping by mining careflows of breast cancer patients. *Journal of Biomedical Informatics*, 66:136–147, 2017.

[5] Rémi Emonet, Jagannadan Varadarajan, and Jean-Marc Odobez. Temporal analysis of motif mixtures using dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):140–156, 2013.

[6] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[7] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining -ACM SIGKDD)*, pages 375–384, 2017.

[8] Rimma Pivovarov, Adler J. Perotte, Edouard Grave, John Angiolillo, Chris H. Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of Biomedical Informatics*, 58:156–165, 2015.

[9] Kejing Yin, Dong Qian, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1246–1253, 2019.