Deep Learning Preprocessing in Process-Mining Methods for Analyzing Hospital Care Pathways

Marie-Hélène METZGER[1] and François BOUE [2]

[1] INSERM U1018, Health Service Research, Hôpital Paul-Brousse, Villejuif, France [2] Assistance Publique- Hôpitaux de Parisn Hôpital Antoine-Béclère, Internal Medicine and Clinical Immunology, Clamart, France

Background.

Process mining is a research discipline that is at the interface between artificial intelligence and data mining, on the one hand, and process modeling and analysis on the other. These methods discover a process that is not explicitly specified in the information system, using "execution traces". In the context of healthcare, the patient trajectory, for example during one of his emergency department visits, then admitted to a hospital ward and up to discharge can be considered as a process. These automated methods include process discovery (i.e., extracting process models) from execution traces, conformance checking (i.e., identification of discrepancies by comparing the theoretical model of care pathway developed from national recommendations to traces of execution in the information system), model extension (improvement of the theoretical model based on the discovery and conformance checking analyses).

Objectives.

The objective of our project is to use recent advances in machine learning and emerging deep learning techniques, to enrich the hospital data used for analyzing hospital care pathways with the unstructured data extracted from electronic medical records (EMR). For this objective, we need collaboration with experts in deep learning methods applied to electronic medical records to solve two important challenges:

- the first challenge is linked to the specificities of unstructured data extracted from electronic medical records (Cheng, et al. 2015): *high-dimensionality* with a large amount of different medical events in medicine; t*emporality*: the temporal labelling of the medical events is very important for the correct medical interpretation of these events; *Irregularity*: the same disease can be described differently from one EMR to the other; *Sparsity*: many data are not systematically entered at each visit. Many data are missing and there are recording mistakes in the EMRs. Very recently, new machine learning approach based on deep learning techniques were proposed to overcome these challenges. This approach increases very rapidly with many recent publications showing the high performance of these methods (Shickel, et al. 2017).

- the second challenge will be to develop user-friendly interfaces for medical researchers to optimize the deep learning parameters for this preprocessing step in order to facilitate the overall process-mining analysis.

Methods.

The patient's EMR is a very rich source of medical information, but because of its unstructured nature in the texts, it is not directly available for process-mining analysis. The reconciliation of a certain number of medical information with the description of the use of care (for example, admission to the emergency department, comorbidity described in a consultation report or in a hospitalization report but not coded in the French DRG) are elements which could be done using deep learning methods.

Study setting: The Assistance Publique - Hôpitaux de Paris (AP-HP) is a group of university hospitals, located in the region “Ile-de-France” (39 hospitals – 20 000 beds). Various types of cohorts (people living with HIV, cancer patients, etc.) can be built. All electronically available data for the care pathway analysis will be extracted from the AP-HP hospital information system which include sociodemographic data, medico-administrative data, inpatient, outpatient and emergency room visits data. The textual corpus will be made up of different types of medical reports (discharge summaries, imaging reports, surgery reports, consultation reports, etc.). The clinical documents used will be drawn from the AP-HP clinical data warehouse based on the i2b2 framework. Different variables available in a structured format will also be retained for the analysis: age, gender, clinical descriptors available in a structured format such as PMSI diagnoses (coded with the ICD-10 classification), medications, lab tests.

Conclusion.

Multidisciplinary work bringing together expertise in process-mining, medicine, computer science, machine learning, epidemiology, public health and hospital management is needed. Overcoming the two challenges described above in order to optimize the use of data available in the hospital information system could contribute to a better analysis of hospital care pathways.