Master Internship 2020

Exploiting heterogeneous COVID-databases with *limited* manual effort: a domain adaptation approach

SUPERVISORS: Michèle Sebag, Francois Landes, Isabelle Guyon
E-MAIL: Michele.Sebag@lri.fr, Francois.Landes@u-psud.fr, Isabelle.Guyon@inria.fr
LAB: TAU − CNRS − INRIA − U. Paris-Saclay
HEAD LAB: J.-Y. Berthou
DURATION: May-September 2020
LOCATION: Telecommuting.
LEVEL: Master student.

ABSTRACT :
Worldwide efforts are deployed to model the impacts of the COVID virus (to predict its spread, the number of ICU-beds needed in the next days in a given hospital or region, to model economic side impacts and necessary support). Many such models are built using Machine Learning [1] and exploiting (currently often proprietary) databases. Massive manual efforts are deployed to clean such databases and build models in a race against time.

In the next stage of the pandemy, it will be most desirable to upscale the models by leveraging databases from different sources, and possibly different regions/countries, filled using different formats and/or database schemas (e.g. the patient age can be recorded based on their birthdate or their number of years; the co-morbidities can be reported under different names; some information might be missing in some databases).

OBJECTIVE OF THE INTERNSHIP: The challenge is to reconcile as automatically as possible different datasets schemata, based on the knowledge that they relate to the same phenomenon. This goal is known as *data wrangling* in the data mining community [2].

An elementary step in data wrangling is to identify the type (continuous, ordinal) of a variable [3].

The proposed approach is to tackle data wrangling as an adversarial domain adaptation problem [4]. Formally, domain adaptation aims to leverage a source dataset (usually containing a wealth of dataw) to learn from a target dataset (with less or no labels and/or less samples). Adversarial domain adaptation aims to embed both source and target data in a same latent space, while offering some guarantees that: i/ the discriminant information from the source dataset is preserved; ii/ the images of the source and target samples cannot be discriminated:
Find f_{source} and g_{target} such that:

$$f_{source} : X_s \mapsto \mathbf{R}^d, \text{ and } \exists f' \text{ s.t. } f' \circ f_{source}(x) \approx x \text{ for all x in } X_s$$
$$g_{target} : X_t \mapsto \mathbf{R}^d, \text{ and } \exists g' \text{ s.t. } g' \circ g_{target}(x) \approx x \text{ for all x in } X_t$$

(no information is lost)

and the binary classification problem defined from dataset

$$\mathcal{E} = \{(f_{source}(x), 1), x \in X_s\} \bigcup \{(g_{target}(x), 0), x \in X_t\}$$

cannot be solved efficiently.

(one cannot make the difference between the images of the source and target samples).

Adversarial domain adaptation will be adapted to data wrangling, and map two (or several datasets) into a same latent space, in such a way that i/ no information is lost; ii/ the images of the datasets coincide. Some manual verification will be needed to guarantee the quality of the mapping: the principles of active learning (asking a few most informative questions to the expert) will be deployed to guide the manual verification.

The intern will join a dynamic and motivated team of data scientists, including other interns. The research conducted may lead to a publication.

The internship requires excellent theoretical and algorithmic skills (programming environment in Python).

Références

[1] ELLIS conference, April 2: Machine Learning applications in Covid-19 Research. https://www.youtube.com/watch?v=0jg_NNwF7k4

[2] Alex Bogatu, Norman W. Paton, Alvaro A. A. Fernandes, Martin Koehler: Towards Automatic Data Format Transformations: Data Wrangling at Scale. Comput. J. 62(7): 1044-1060 (2019)

[3] Patricio Cerda, Gaël Varoquaux, Balázs Kégl: Similarity encoding for learning with dirty categorical variables. Mach. Learn. 107(8-10): 1477-1494 (2018)

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Victor S. Lempitsky: Domain-Adversarial Training of Neural Networks. J. Mach. Learn. Res. 17: 59:1-59:35 (2016).