Random Forests for Dissimilarity-based learning: application to oncology by jointly using radiomic and genomic

Keywords : Dissimilarity space, Random Forest, Heterogeneous data, Radiomic.

Financement : PhD thesis scholarship granted by the region "Normandie" (from 01/10/2016 au 30/09/2019)

Encadrement :

Laurent HEUTTE (supervisor), laurent.heutte@univ-rouen.fr, (+33) 2 32 95 50 14

Simon BERNARD (co-supervisor), simon.bernard@univ-rouen.fr, (+33) 2 32 95 52 05

Équipe d'accueil :

Learning team, LITIS lab (EA 4108), University of Rouen

(http ://www.litislab.fr/equipe/docapp/)

1 Candidate profil

The candidate must have a Master degree (or equivalent) in Statistics, Computer Science or Computer Engineering, with a major in Data Science and/or Signal and Image Processing. The candidate must have strong skills in Machine Learning and Classification.

2 Research lab

The LITIS lab. (Computer Science, Information and System Processing) is the research unit on information science and technologies in the region "Haute Normandie". It gathers the researchers in Information and Communication Sciences and Technologies (ICST) from the three main universities and engineering school in this region : the University of Rouen, the University of Le Havre and the National Institute of Applied Sciences (INSA) of Rouen. The LITIS lab. is composed of 160 researchers, including 80 PhD students. It is divided into seven research teams, covering a large range of ICST topics, from fundamental research to applied domains, with connections with human sciences.

The PhD student will be part of the "Learning" team made up with around 15 researchers from the University of Rouen and the INSA school of Rouen, and with around 15 PhD students. The research projects of the "Learning" team cope with tools and methods for dealing with diverse data in terms of structure, dimensionality and stationarity, and coming from heterogeneous contexts (signals, images and texts). These research works essentially adopt the machine learning point of view, for pattern recognition problems.

3 The PhD subject

3.1 Scientific context

Radiomic, as introduced by Lambin et al. [LRVL$^+$12], is defined as the extraction and analysis of many quantitative image features, from Computed Tomography (CT) scans, Positron Emission Tomography (PET) scans or Magnetic Resonance Imaging (MRI), but also of clinical and omic data (genomic, proteomic, etc.). These data can efficiently be used for designing descriptive and predictive models that can relate images features to cancer phenotype or genetic fingerprint. The key idea behind radiomic is that descriptive models, learnt from these types of data, can supply precious diagnosis, prognosis and predictive information for an efficient treatment of cancer.

The research work will tackle this problem through the machine learning perspective, by focusing on learning efficient predictive models in these high-dimensional heterogeneous description spaces.

3.2 Contributions

Designing models from these kinds of data obviously face a major issue : the difficulty of handling a large amount of data, very heterogenous by nature. Indeed, clinical data come from demographic information as well as medical information, collected during consultations or extracted from medical reports. Similarly, genomic data can be factual, textual and diverse by nature. At last, images are obtained from different imaging devices and represent different types of medical information. This research work adopts a machine learning point of view for building models that allow to overcome the two following issues : (i) learning and selecting relevant representations, based on the three types of data (image, clinical and genomic) ; (ii) learning predictive models that are able to supply valuable information for cancer treatment (model interpretability).

Both issues, that rise from the data heterogeneity, are planed to be tackled with dissimilarity-based pattern recognition ([PD05]) techniques, that allow to overcome the difficulty of searching for an optimal representation of the data when these data are initialy high-dimensional and diverse (numeric, nominal, sequencial, etc.). Given n raw data samples, initially described with p features, a dissimilarity space is a n-dimensional space, for which the k^{th} dimension describes the dissimilarity of any sample with the k^{th} training sample. This approach is particularly efficient for high-dimensional problems ($p >> n$) since the training samples are projected in a much lower dimensioned space (n), which allows to use a wide variety of machine learning techniques. However, the main issue stems from the design of a suitable metric able to precisely and efficiently measure the data dissimilarity. In particular, when data come from different sources (which is the case in the radiomic context where clinical, genomic, and image-based information are collected for each patient), it is impossible to define a single dissimilarity measure on a heterogeneous pool of features. At best, one measure can be defined per source.

In this research work, it is planed to address the problem of choosing a dissimilarity measure through the metric learning point of view, using Random Forest methods, classification techniques based on ensembles of decision trees for which we already have a strong expertise [BHA12, DBHP13]. In particular, the first goal is to study the way Random Forest approaches can be used to build dissimilarity spaces. Indeed, these approaches seem particularly suitable for this issue since they embed a feature selection mechanism in their learning procedure. They also embed a similarity estimation mechanism, that can be exploited in the radiomic context for computing the dissimilarities between samples. This latter mechanism has already proven to be efficient and flexible in a various range of learning tasks [TPC06, SH06]. Furthermore, Random Forests are widely used in the medical field since they naturaly take into account feature correlations, they can also handle high-dimensions, and they are particularly easy to interpret, understand and analyze. These latter properties often offer to the medical staff, precious information on the relevancy of each feature from the original description space (image, genomic or clinical features) and on their contribution to the diagnosis [SH06].

3.3 Collaborations with foreign research teams

This research work will be led in close collaboration with two other research teams in LITIS lab. : the TIBS team for the clinical and genomic data expertise and the QuantIF team for the medical imaging expertise. An INSERM unit (U918) from the University of Rouen and Henri Becquerel center will also be part of this collaboration. Finally, this research work will be led with the cooperation of Prof. Robert Sabourin from the ETS in Montréal, Canada, and could also be part of a collaboration with the Federal University of Parana, Brazil (Prof. Luiz E.S. Oliveira).

Références

[BHA12] S. Bernard, L. Heutte, and S. Adam. Dynamic Random Forests. *Pattern Recognition Letters*, 33 :1580–1586, 2012.

[DBHP13] C. Désir, S. Bernard, L. Heutte, and C. Petitjean. One-Class Random Forest. *Pattern Recognition*, 46(12) :3490–3506, 2013.

[LRVL+12] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, André Dekker, and Hugo J.W.L. Aerts. Radiomics : Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4) :441 – 446, 2012.

[PD05] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition : Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005.

[SH06] T. Shi and S. Horvath. Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics*, 15 :118–138, 2006.

[TPC06] Alexey Tsymbal, Mykola Pechenizkiy, and Pádraig Cunningham. *Machine Learning : ECML 2006 : 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings*, chapter Dynamic Integration with Random Forests, pages 801–808. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.