CREATIS - UMR CNRS 5220 - Inserm U1206
www.creatis.insa-lyon.fr

Sujet de thèse

Nom du laboratoire : CREATIS
Equipe : Images et Modèles
Directeur de thèse : Carole Lartizien
Comité d'encadrement : Ievgen Redko
Contact : ievgen.redko@creatis.insa-lyon.fr, carole.lartizien@creatis.insa-lyon.fr
Thématique de la thèse : Traitement du signal et de l'Image

Provably accurate metric learning for heterogeneous medical imaging: application to multi view learning and domain adaptation

Domaine et contexte scientifiques, mots-clefs:

In recent years, machine learning algorithms have been widely and successfully applied in various medical imaging domains including segmentation, image registration and functional brain mapping. The CREATIS team 'Images and Models' has developed knowledge in machine learning for medical imaging over the past few years and expertise to prototype novel computer-aided diagnosis (CAD) systems for cancer screening (Niaf, Flamary et al. 2014) and neurological disease imaging (El Azami, Hammers et al. 2016) based on multimodality (MRI, PET, CT) imaging.

CAD has become a major research subject due to its capacity in assisting radiologists during their diagnostic task by providing information on the location and characterization (malignancy score) of suspicious regions of interest. The algorithms used to build CAD systems learn a multi-class (mostly binary) decision model in a multidimensional feature space based on training samples from the different classes of interest. Diagnostic performance of such decision support systems is highly impacted by the quality of the training database that should contain a large number of correctly annotated and homogeneous (i.e., acquired with similar imaging protocols) cases of all classes. Such a condition, however, is not easily met in clinical practice and thus needs to be addressed with new methodological and algorithmic advances. The proposed thesis proposal focuses on developing new efficient machine learning algorithms for medical imaging with heterogeneous data.

Mots-clefs: Machine learning, heterogeneous data, medical imaging

Verrous scientifiques

The current success of machine learning methods in real-worlds applications, especially the impressive results obtained using deep learning, largely depends on the size of the annotated sample available for learning. Indeed, it is a known fact that one cannot hope to capture the general patterns and peculiarities in the data using training samples of a very limited size because they are not representative enough. Unfortunately, **in medical imaging, the problem of acquiring sufficiently big data sets is widely present** and is often due to ethical issues as well as time consuming processes required to build annotated databases. On the other hand, **real-world data are often heterogeneous**: for instance, a given person can be described in multiple ways including its physical appearance, social relationships or financial background. In the context of medical imaging, the heterogeneous nature of the data available for each patient through medical texts, images of different modalities and clinician notes becomes even more important as it can allow to adapt the best treatment based on the different angles of view. Therefore, finding solutions for these two problems presents a major challenge in medical imaging.

Objectif

The main scientific goal of this thesis proposal is to improve the performance of CAD systems by learning from multi-modal and multi-source data for various kinds of medical imaging data. We believe that this goal answers to two major problems widely presented in medical imaging:

(1) how to fusion heterogeneous data from different clinical sites acquired using different imaging protocols?

(2) how to avoid time-consuming manual labelling of samples by using already available annotated images?

While both questions may seem similar from a distance, they are treated in a different manner from the machine learning point of view.

Contributions originales attendues

We expect to produce new efficient algorithms for multi-view learning and domain adaptation with strong theoretical guarantees based on the ideas from metric learning.

Programme de recherche et démarche scientifique proposée:

This thesis proposal is structured around two different axes which are **multi-view learning** and **domain adaptation.**

Multi-view learning

Multi-view learning deals with learning tasks where the same data points are being represented in different feature spaces [Xu et al., 2013]. As an illustrative example presented in Figure 1A, one may consider the realistic situation where we possess images related to different modalities (e.g. PET and MRI) or different types of sequences of the same modality (e. g. MRI FLAIR and MRI T2 sequences). In this case, every imaging type cited above can be considered as a different view of the same object with clearly different statistical properties. While most of the existing applications of machine learning in medical imaging are limited to learning from a data sample acquired on a certain type of imaging modality (e.g., MRI, CT, PET etc.) and coming from a single data source, we propose to derive new algorithms that will be able to benefit from different data representations when learning a classifier. We hypothesize that taking into account different data representation of the same studied objects may increase the diversity of the model and thus compensate the need for more data.

A natural way of incorporating the knowledge provided by different data representations is to learn based on the similarity matrix obtained by combining available data views. In this thesis proposal, we would like to develop new multi-view learning algorithms built upon the notion of (epsilon, gamma, tau) good classifiers (Balcan, Blum et al. 2008). More formally, we will extend the paper of (Bellet, Habrard et al. 2012) to the case of learning with multiple data representation by building one bilinear similarity function for each view and then combining them using an appropriate data regularisation term.

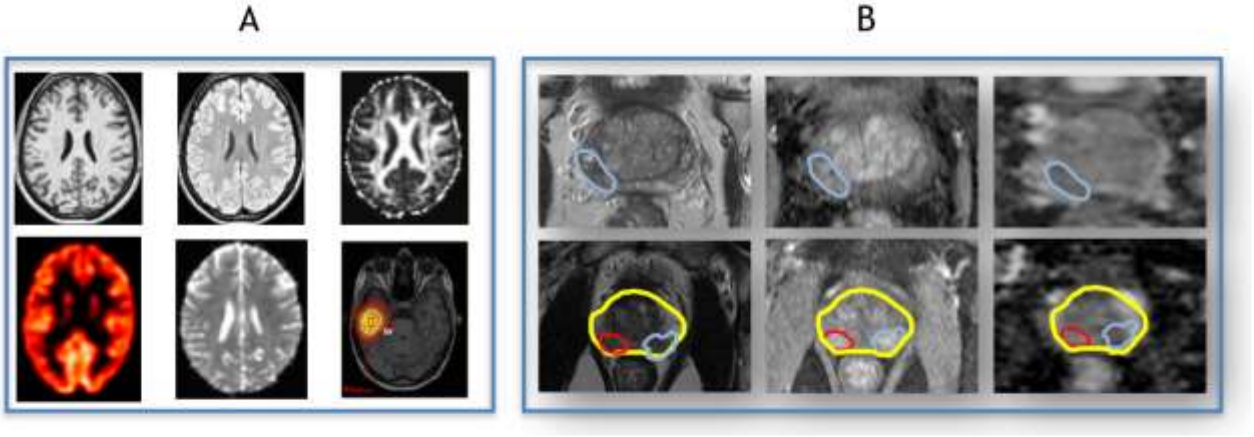

Figure 1. Examples of multi view learning (A) and domain adaptation (B) in medical imaging

The possible application of this algorithm will be the automated detection of epileptogenic lesions based on multiparametric MRI and PET acquisitions. The PhD student will indeed have access to a data set of about 100 normal and 20 pathological exams performed by our clinical partner from Centre de Recherche en Neurosciences de Lyon (CRNL).

Multi-source domain adaptation

Second objective of this proposal is to work on domain adaptation, the particular case occurring when learning is performed based on the data samples having different nature and thus following different probability distributions. An illustrative example for this learning scenario is given in Figure 1B where top and bottom rows are multiparametric MR images (T2, DCE, ADC) of two different patients with prostate cancer. Here one can consider learning on the annotated images of first patient in order to improve the performance of the diagnostics for the second patient. A striking usefulness that domain adaptation may have for medical imaging lies in its capacity to treat the data acquired on different clinical sites by adapting them. This may lead to the creation of new labeled benchmark data sets that will be further used in large-scale learning.

In the multi-source domain adaptation problem, we consider each patient as a source domain and try to induce the annotation of a previously unseen patient (target domain) using the labeled samples in order to avoid, partially or completely, manual labeling (Pan and Yang 2010, Patel, Gopalan et al. 2015). According to the theory of domain adaptation proposed in (Ben-David, Blitzer et al. 2010), the successful adaptation between a set of data sources from different probability distributions becomes possible if the divergence between their distributions is minimized for the observed samples. In this thesis proposal, we will explore the idea of learning similarity measures for both source and target patients with respect to the same sets of the reasonable points. The intuition behind this procedure consists in building classifiers that are provably good for source data points but transferrable to target ones due to the shared space of landmarks used to classify data.

This work will also include an important investigation of theoretical properties of the produced algorithms as well as its application to two use-cases : (1) brain decoding based on fMRI data and (2) prostate cancer mapping based on multiparametric MR imaging. These two use-cases will be derived from datasets collected as part of two projects developed in our team in collaboration with clinical partners from Institut des Neurosciences de la Timone (INT) and Hospices Civils de Lyon (HCL), respectively.

Encadrement scientifique :

PhD student will be supervised by Carole Lartizien (CR1, CNRS) and Ievgen Redko (MCU, INSA de Lyon). The participation of Carole Lartizien will be at around 25% while that of Ievgen Redko about 75%. The PhD student will benefit from the strong expertise of Carole Lartizien in developing CAD systems based on machine learning techniques as well as from the in-depth knowledge of theoretical and algorithmically aspects of statistical learning of Ievgen Redko.

Profil du candidat recherché :
Successful candidate will have strong knowledge in at least two of the following fields:

- Image processing
- Statistical learning (machine learning)
- Applied mathematics

Objectifs de valorisation des travaux de recherche :

The results of this project will be presented at top machine learning and medical imaging venues as well as at the dedicated workshops concerning their intersections. The extended experimental evaluations will be submitted to top peer machine learning and medical imaging journals. On the other hand, the developed algorithms will be implemented and made available through the VIP platform (https://www.creatis.insa-lyon.fr/vip/) [ANR-09-COSI-03] in order to enable the users to perform empirical studies using the algorithms on the collected data sets. We expect these new methods to have a strong impact on the work of the implicated scientists from medical imaging community that heavily use multivariate machine learning models to analyze different kinds of MRI data they routinely acquire and that is clearly in demand for multi-subjects multivariate techniques.

Compétences développées au cours de la thèse et perspective professionnelle :

The successful candidate will develop strong skills in statistical machine learning, medical imaging analysis and their applications. Due to the theoretical and applicational aspects of this proposal, the successful PhD student will be able to join research departments in both industry and academia.

Références bibliographiques sur le sujet :

Balcan, M., et al. (2008). Improved guarantees for learning via similarity functions. COLT**:** 287–298.

Bellet, A., et al. (2012). Similarity learning for provably accurate sparse linear classification. International conference on machine learning (ICML).

Ben-David, S., et al. (2010). "A theory of learning from different domains." Machine Learning **79**((1-2)): 151-175.

El Azami, M., et al. (2016). "Detection of Lesions Underlying Intractable Epilepsy on T1-Weighted MRI as an Outlier Detection Problem." PLoS One **11**(9): e0161498.

Niaf, E., et al. (2014). "Kernel-Based Learning From Both Qualitative and Quantitative Labels: Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging." IEEE Transactions on Image Processing **23**(3): 979-991.

Pan, S. J. and Q. Yang (2010). "A survey on transfer learning." IEEE Transactions on Knowledge and Data Engineering **22**: 1345-1359.

Patel, V. M., et al. (2015). "Visual domain adaptation : A survey of recent advances." IEEE Signal Processing Magazine **32**(3): 53-69.