Proposition de thèse de doctorat

Sélection de variables en grande dimension et en présence de variables de nuisance : application à l'identification de biomarqueurs

Laboratoire d'accueil : Laboratoire ICube (UMR 7357), Université de Strasbourg, CNRS.

Encadrement : Sylvain Faisan, ICube, équipe MIV (faisan@unistra.fr)

Collaborations :

- Frédéric Blanc, ICube, équipe IMIS. Analyse des résultats obtenus concernant le diagnostic différentiel entre la Maladie d'Alzheimer et la Démence à Corps de Lewy.
- Entreprise HypnoVR. Autre application des méthodes développées : identification des biomarqueurs de l'état d'hypnose.

Financement : contrat doctoral Université de Strasbourg

Compétences souhaitées : cette thèse s'adresse à un(e) étudiant(e) titulaire d'un master 2 (ou d'un diplôme d'ingénieur) avec un bon niveau en mathématiques appliquées (statistique, analyse de données, apprentissage) et en programmation (C++ ou Python).

Contact : envoyer CV, lettre de motivation, résultats et classements de Master ou école d'ingénieur à Sylvain Faisan (faisan@unistra.fr)

Mots-clés : diagnostic assisté par ordinateur, classification supervisée, importance des variables.

Sujet

La Démence à Corps de Lewy (DCL) est la seconde démence la plus répandue après la Maladie d'Alzheimer (MA). Les critères diagnostiques actuels de la DCL, bien que spécifiques, manquent de sensibilité, ce qui conduit à ne pas diagnostiquer cette pathologie chez un grand nombre de sujets, retardant ainsi leur prise en charge thérapeutique. Il est donc un enjeu majeur de proposer des outils pour l'aide au diagnostic différentiel entre MA et DCL, notamment pour les patients présentant les premiers signes annonciateurs de ces pathologies (stade prodromal). L'objectif de la thèse est de proposer des méthodes automatiques, basées sur des techniques d'apprentissage supervisé (forêts aléatoires **[1]**, LASSO **[2]**, ...), permettant d'identifier des biomarqueurs pertinents en imagerie par résonance magnétique (IRM). Ce projet s'appuiera sur une cohorte de 250 individus[1], composée de sujets sains, de patients DCL et MA.

1https://clinicaltrials.gov/ct2/show/record/NCT01876459

La problématique abordée recouvre principalement deux difficultés. D'une part, le nombre de sujets dans la cohorte est relativement faible par rapport aux nombres de caractéristiques si bien qu'une attention toute particulière devra être apportée pour s'assurer que les résultats soient généralisables. Dans ce cadre, il est primordial de mettre en place des stratégies de manière à ne pas surestimer l'importance prédictive des variables : on pourra notamment mettre en œuvre des approches de validation croisée et utiliser des tests de permutation permettant de déterminer la significativité des résultats des modèles prédictifs **[3]**. D'autre part, les caractéristiques extraites sont corrompues par des variables de nuisance. Ces dernières induisent une variabilité sur la mesure du biomarqueur, qui peut entraîner une réduction drastique de ses capacités de prédiction. Par exemple, le vieillissement normal induit une atrophie cérébrale. Ainsi, de manière à ce que les caractéristiques associées aux volumes des structures cérébrales puissent permettre une bonne discrimination entre les classes, il faut que l'âge soit pris en compte. Trois stratégies sont généralement utilisées. La première possibilité est de diviser les données en sous-groupes homogènes pour les variables de nuisance (analyse stratifiée), limitant ainsi leurs effets. Cependant, cette stratégie nécessite d'avoir de nombreuses observations à disposition. La seconde possibilité est de considérer les variables de nuisance comme des variables « ordinaires » en espérant que le modèle représente correctement les « liens » qui existent entre les différentes variables. Cependant, cette manière de faire n'est pas sans risque. Par exemple, dans le cadre d'un problème de classification à deux classes, si une variable de nuisance permet de discriminer relativement bien les deux groupes (en raison d'un « mauvais » échantillonnage des populations), on va estimer à tort que cette variable de nuisance est d'intérêt. La dernière possibilité consiste à enlever l'influence des variables de nuisance (le modèle linéaire général est souvent utilisé). Dans un contexte de classification, les auteurs de **[4]** ont observé qu'utiliser des modèles de régression différents pour chaque groupe réduit l'effet lié aux groupes. De plus, la régression des caractéristiques d'un individu nécessite de connaître son groupe d'appartenance (afin de choisir le bon modèle de régression), qui n'est a priori pas connu avant l'étape de classification. Ainsi, les données sont en général régressées suivant un unique modèle de normalité ce qui n'est pas véritablement satisfaisant. Différentes stratégies pourront être utilisées pour définir une stratégie pertinente : par exemple, restaurer la normalité d'un individu pathologique en utilisant un modèle de régression lié à une maladie peut être un moyen de définir un diagnostic.

Les approches développées seront suffisamment génériques pour être utilisées dans d'autres applications telles que l'identification des biomarqueurs de l'état d'hypnose.

[1] Genuer R, Poggi J.-M., Tuleau-Malot C. (2010) *Variable selection using Random Forests*. Pattern Recognition Letters.

[2] Zou H., Hastie T (2005) *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society. Series B: Statistical Methodology

[3] Golland P., Liang F., Mukherjee S., Panchenko D. (2005) *Permutation Tests for Classification*. Lecture Notes in Computer Science, vol 3559. Springer, Berlin, Heidelberg.

[4] Dukart, J., Schroeter M. L., Mueller K. , A. D. N. Initiative et collab. (2011), *Age correction in dementia–matching to a healthy brain*, PloS one