INSTITUT DE FRANCE
Académie des sciences

Inria

UNIVERSITÉ CÔTE D'AZUR

VILLE DE NICE

Académie en région
à Nice et Sophia Antipolis

Apprentissage Statistique sur Données Complexes : des réseaux de communication à la médecine computationnelle

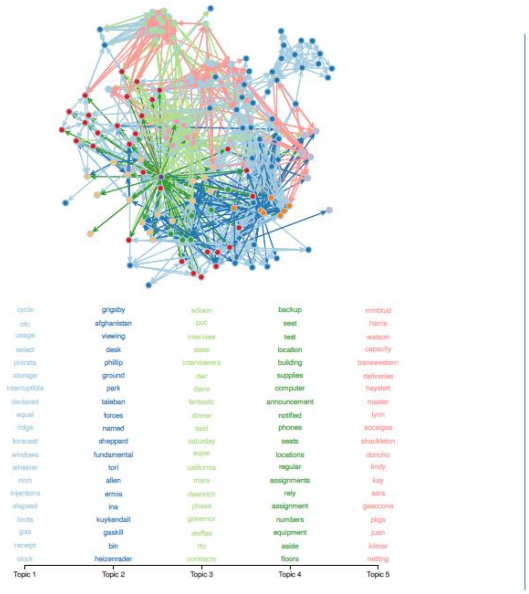
Charles BOUVEYRON
Université Côte d'Azur, Inria

L'apprentissage statistique joue de nos jours un rôle croissant dans de nombreux domaines scientifiques aussi variés que la médecine, l'imagerie, la biologie, l'astronomie ou la défense. Les progrès scientifiques réalisés ces dernières années ont permis d'augmenter sensiblement les capacités de mesure et de calcul, et il est à présent difficile pour un opérateur humain de traiter de façon exhaustive ces données dans un temps raisonnable. En particulier, de nombreuses spécialités médicales, telles que l'imagerie médicale, la radiologie ou la génomique, ont bénéficié dans les dernières décennies d'évolutions importantes de leurs technologies. De même, le secteur des communications a vu, avec l'arrivée d'internet, une augmentation massive des données dites transactionnelles, qu'il est utile d'analyser pour des raisons de sécurité par exemple. Dans ces deux cas, ces évolutions ont amené les spécialistes de ces domaines à devoir repenser leur pratique des données. L'apprentissage statistique, qui doit être vu comme une sous-discipline de ce que l'on appelle aujourd'hui communément l'intelligence artificielle (IA), se propose alors de prendre le relais sur l'expert humain pour modéliser et synthétiser ces données complexes dans le but d'aider les analystes à la prise de décision.

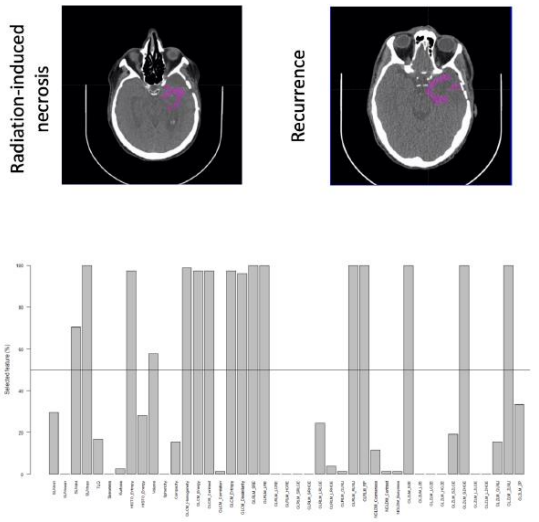
Dans les applications médicales, la classification supervisée (ou analyse discriminante) est probablement la méthode d'apprentissage la plus utilisée pour le diagnostic ou le pronostic lié à des pathologies. Néanmoins, certaines situations pratiques correspondent à des problèmes théoriques qui ne sont pas entièrement résolus. Par exemple, la classification de données de très grande dimension ou la classification de données corrélées sont des problèmes particulièrement présents en analyse d'images et biologie et pour lesquels les solutions actuelles, pourtant déjà très avancées, nécessitent que des recherches soient poursuivies. En particulier, la grande dimension des données (nombre de variables important) pose un ensemble de problèmes à la statistique multivariée classique que l'on résume usuellement par le terme « fléau de la dimension ».

En ce qui concerne le secteur des communications, une problématique centrale de l'analyse des réseaux de communication est la capacité d'identifier des groupes d'individus ayant un comportant homogène dans le réseau. Cette problématique, dite non-supervisée, est commune aux équipes de marketing, de protection du cyber-harcèlement mais aussi de renseignement militaire. Malgré les avancées récentes dans ce domaine, l'analyse conjointe du réseau et des textes portés par les arêtes du réseau restait un problème ouvert.

Il a donc été nécessaire de développer ces dernières années des méthodes capables de pallier les problèmes. Nous avons abordé dans cet exposé deux méthodes récentes d'apprentissage statistique, gsHDDA [1] et Linkage [2], permettant de faire face respectivement à la classification de données de grande dimension pour l'analyse d'images médicales et l'analyse de réseaux de communication. La figure ci-dessous illustre les résultats présentés.



(a) Analyse de réseaux de communication avec Linkage.



(b) gsHDDA pour la médecine computationnelle.

Références :

[1] F. Orlhac, P.-A. Mattei, C. Bouveyron and N. Ayache, *Class-specific Variable Selection in High-Dimensional Discriminant Analysis through Bayesian Sparsity*, Journal of Chemometrics, vol. 32(2), 2019.

[2] C. Bouveyron, P. Latouche and R. Zreik, *The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges*, Statistics and Computing, vol. 28(1), pp. 11-31, 2017.

[3] C. Bouveyron, G. Celeux, B. Murphy and A. Raftery, *Model-based Clustering and Classification for Data Science, with Applications in R*, in Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.