# Recherche proposée - Optimisation de systèmes de classification à catégories multiples avec modélisation efficace des dépendances

## a) Problématique

L'intérêt pour les algorithmes de classification à catégories multiples (multi-label classification) s'est grandement accru à la suite de récentes avancées dans le domaine de l'apprentissage statistique [1, 2]. Une de ces applications porte sur l'annotation de contenu en ligne, comme des images, des articles ou des vidéos. Par exemple, on désire classer différents articles avec plusieurs sujets, ou encore associer des mots-clés à une banque d'images. Avec l'émergence des réseaux sociaux et des données numériques, nous sommes confrontés à une quantité volumineuse de données empêchant un travail manuel pour effectuer cette tâche, d'où la nécessité de développer des algorithmes d'intelligence artificielle.

Les algorithmes d'apprentissage traditionnels associant une seule et unique catégorie à chaque objet sont nombreux, mais négligent les relations entre les catégories. En effet, lorsque ces algorithmes étudient un groupe en particulier, ils ignorent totalement la présence des autres. Pour incorporer l'information des dépendances entre les catégories, certains préconisent l'utilisation de la corrélation [3]. Or, ces techniques écartent la possibilité que des relations ne soient pas symétriques. On n'a qu'à penser au terme « pomme » qui est un exemple de « fruit », la relation inverse étant potentiellement fausse. D'autres préfèrent plutôt des approches hiérarchiques [4]. Par contre, une telle structure impose qu'une catégorie sélectionnée implique aussi tous ces ancêtres dans la hiérarchie, alors qu'en réalité, plusieurs termes ont de multiples significations, ce qui cause problème. Il est donc primordial de développer une méthodologie pour prendre en compte ce type de relations et ainsi pallier aux méthodes existantes dans la littérature.

De plus, certains algorithmes performent mieux que les techniques prisées dans des contextes précis, par exemple où le nombre de données est relativement faible. Or, dans la grande majorité des cas auxquels nous faisons face, la quantité de données disponibles est gigantesque. Effectivement, des plateformes comme YouTube et Facebook jonglent avec une quantité astronomique de données, en l'occurrence des images et des vidéos, qui ne cesse de croître. Pour classer des vidéos ou des images, ces algorithmes sont inapplicables en raison des performances médiocres qu'ils procurent. Ainsi, le défi de taille que représente l'immensité des données disponibles nous oblige à revoir l'implémentation de ces algorithmes pour qu'ils performent rapidement et efficacement.

## b) Objectifs de recherche

Développer un algorithme permettant la modélisation structurelle et stochastique des dépendances entre les catégories. Élaborer une méthodologie d'apprentissage incrémental (online learning) pour son intégration dans un contexte de mégadonnées.

## c) Méthodologie et procédure proposée

Tout d'abord, nous développerons un modèle prenant en compte les relations désirées entre les variables latentes. La difficulté majeure se situe dans la représentation de ces dépendances qui, dans d'autres algorithmes, imposent des hypothèses trop strictes. La solution envisagée s'inspire des modèles graphiques probabilistes, pour ses hypothèses flexibles permettant de s'adapter aux différents scénarios possibles. De récents résultats suggèrent qu'une conceptualisation de notre approche selon un cadre bayésien facilitera l'apprentissage incrémental afin que le modèle se mette à jour avec l'ajout de nouvelles données plutôt

qu'un calcul complet du modèle soit nécessaire. La considération de jeux de données réels ainsi que la quantification du gain en performance seront centrales au développement de notre technique. Finalement, une implémentation de l'algorithme sera réalisée pour rendre accessible notre méthodologie, par exemple à travers une contribution à un paquet R.

d) Contribution à l'avancement des connaissances

Ce projet propose le développement d'un nouveau modèle d'apprentissage statistique ayant de nombreuses applications, que ce soit en bio-informatique, pour la classification de gènes par rapport à leurs fonctionnalités dans l'organisme, ou en catégorisation de textes, où une multitude d'articles doivent être classés sous divers sujets. Il s'agit alors d'une contribution ayant des retombées non seulement pour le domaine de la statistique, mais aussi pour plusieurs autres champs de spécialisation scientifiques.

De plus, le développement de nouvelles approches à l'apprentissage incrémental pourra conduire à améliorer la performance d'algorithmes actuels dans des domaines jonglant avec une quantité massive de données, comme la vision par ordinateur, en plus de parfaire notre compréhension de l'intelligence artificielle.

e) Bibliographie

[1] Zhang, M. L., & Zhou, Z. H. (2014). A Review on Multi-Label Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 8, 1819-1837.

[2] Gibaja, E., & Ventura, S. (2014). Multi-label learning : a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, Vol. 4, No. 6, 411-444.

[3] Zhu, Y., Kwok, J. T., & Zhou, Z. H. (2018). Multi-Label Learning with Global and Local Label Correlation. IEEE Transactions on Knowledge and Data Engineering, Vol. 30, No. 6, 1081-1094.

[4] Cerri, R., Barros, R. C., & De Carvalho, A. C. (2014). Hierarchical Multi-Label Classification Using Local Neural Networks. Journal of Computer and System Sciences, Vol. 80, No. 1, 39-56.