L'intérêt pour les algorithmes de classification à catégories multiples ("multi-label classification") s'est grandement accru à la suite des nombreuses avancées dans le domaine de l'apprentissage statistique. En bio-informatique, les chercheurs désirent classer les gènes humains par rapport à leurs différentes fonctionnalités dans l'organisme. En catégorisation de textes, on vise plutôt à classer des articles sous divers sujets. En annotation du contenu en ligne, on souhaite associer des mots-clés à une banque d'images ou de vidéos. Les algorithmes d'apprentissage traditionnels associant une seule et unique catégorie à chaque objet sont nombreux, mais négligent les relations entre les catégories, puisque lorsque ces algorithmes étudient un groupe en particulier, ils ignorent totalement la présence des autres. Il est donc primordial de développer une méthodologie pour incorporer les dépendances entre catégories.

Pour y parvenir, certains préconisent l'utilisation de matrices de corrélation ou préfèrent plutôt une approche hiérarchique. Or, ces techniques comportent de nombreuses lacunes et écartent plusieurs possibilités. Avec une matrice de corrélation, on omet des relations asymétriques. On n'a qu'à penser au terme « pomme » qui est un exemple de « fruit », la relation inverse étant potentiellement fausse. De leur côté, les modèles hiérarchiques imposent que tous les ancêtres dans la structure soient sélectionnés. Or, en réalité, plusieurs termes ont de multiples significations, ce qui cause problème avec ces méthodologies. Il importe alors d'élaborer un algorithme prenant en compte ce type de relations pour ainsi améliorer les méthodes existantes dans la littérature.

De plus, 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. En effet, que ce soit par souci de performance ou de mémoire, il est probable que nous ne voulions pas calculer un nouveau modèle à chaque ajout de données, mais plutôt incorporer l'information au présent modèle. Ce problème se situe dans la théorie de l'apprentissage incrémental ("online learning"), un domaine de recherche ayant pris de l'ampleur avec l'émergence des méthodes d'apprentissage statistique.

Le projet de recherche vise ainsi à développer un algorithme de classification à catégories multiples permettant une modélisation adéquate des dépendances entre les catégories, à travers une modélisation structurelle et stochastique, et pouvant être intégré dans un contexte de mégadonnées. Le développement de cette nouvelle approche pourra conduire à parfaire notre compréhension de l'intelligence artificielle.