Université Claude Bernard Lyon 1

DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **28 juin 2024**

Nom de famille et prénom de l'auteur : **Monsieur STOCKSIEKER Samuel**

Titre de la thèse : *« Contribution de l'apprentissage automatique à la modélisation des valeurs rares et des données déséquilibrées : applications en assurance »*

Résumé

CONTEXTE : Les données jouent un rôle essentiel dans l'apprentissage automatique, la modélisation statistique et, plus généralement, l'intelligence artificielle. En effet, ces disciplines reposent sur la capacité à extraire des informations significatives à partir de données brutes dans le but de pouvoir expliquer et/ou prédire un phénomène, résoudre des problèmes complexes ou encore pour prendre des décisions de manière autonome. Les données fournissent la matière première nécessaire à la construction des modèles, qu'il s'agisse de réseaux neuronaux profonds, de méthodes d'apprentissage supervisé ou non supervisé, ou de techniques statistiques plus traditionnelles. Plus les données sont riches, variées et représentatives de la réalité, plus les modèles peuvent être précis, généralisables et utiles dans divers contextes. Par conséquent, la qualité des résultats obtenus dépend souvent de la qualité des données utilisées.

L'apprentissage à partir de valeurs rares, extrêmes ou non, et plus généralement des données déséquilibrées reste cependant un défi majeur. Les approches standards ont tendance à négliger ces valeurs, ce qui peut entraîner une qualité de modélisation médiocre pour ce type de données. De plus, les valeurs rares représentent souvent un événement important que les praticiens cherchent à comprendre ou prédire.

PROBLEMATIQUE : Les approches pour aborder le phénomène d'apprentissage sur données déséquilibrées se focalisent principalement sur l'apprentissage supervisé. La majorité des solutions concernent le déséquilibre dans le contexte de la classification, en particulier binaire - contexte pour

lequel une abondance de méthodes existe. En revanche, le cadre de la régression a été relativement peu exploré. Le fait que la variable soit quantitative engendre en effet plusieurs difficultés complexifiant davantage le problème.

Enfin, le déséquilibre peut également concerner une ou plusieurs variables explicatives, par exemple en raison d'un biais de sélection. Il peut aussi être rencontré dans le cadre multi-supervisé où l'influence des variables dans l'apprentissage peut être déséquilibrée et, par conséquent, conduire à négliger les valeurs rares.

SOLUTIONS PROPOSEES : Les travaux sont articulés autour de deux grands axes : l'*Imbalanced Features* et l'*Imbalanced Regression*. Le premier axe aborde la problématique du déséquilibre de caractéristiques c'est-à-dire lorsqu'il concerne les attributs des observations et non la variable à expliquer. La première solution consiste à redresser la distribution d'une covariable par rapport à une distribution cible donnée. Elle propose de combiner un rééchantillonnage pondéré et des générateurs de données synthétiques. Cette stratégie permet notamment de faire face au biais de sélection. Une seconde solution est proposée dans le cadre de l'apprentissage multi-supervisé, en particulier avec les autoencodeurs. Elle s'appuie sur une nouvelle métrique visant à équilibrer l'influence des variables lors de l'apprentissage et est applicable non seulement dans des contextes supervisés et non supervisés, mais également dans des modèles génératifs tels que les *Variational Autoencoder* (VAE).

La seconde partie traite la question de la régression à partir de données déséquilibrées. Différentes solutions de prétraitement, notamment de génération de données synthétiques sont proposées. Dans un premier temps, nous proposons d'explorer l'espace initial des données en introduisant de nouveaux générateurs et une nouvelle méthodologie pour aborder le cas spécifique de la régression. Nous proposons ensuite de plonger les données dans un espace latent dans le but d'offrir un cadre plus propice à la génération de données synthétiques. Nous proposant notamment le *Deep Smoothed Bootstrap* en adaptant le générateur naturel des VAE. Enfin, dans le but de construire un espace latent indépendant, une nouvelle métrique identifiante et mesurant les corrélations non linéaires est proposée.

Mots-clés *: Valeurs Rares ; Données Déséquilibrées ; Génération de données synthétiques ; Corrélation non linéaire ; Autoencodeur variationnel ; rééchantillonnage ; Bootstrap lissé*