DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **20 décembre 2019**

Nom de famille et prénom de l'auteur : **MAKKI Sara**

Titre de la thèse : « *Un Modèle de Classification Efficace pour l'Analyse des Données Déséquilibrées pour Détecter les Fraudes dans le Secteur Financie* ».

Résumé

Différents types de risques existent dans le domaine financier, tels que le financement du terrorisme, le blanchiment d'argent, la fraude de cartes de crédit, la fraude d'assurance, les risques de crédit, *etc.* Tout type de fraude peut entraîner des conséquences catastrophiques pour des entités telles que les banques ou les compagnies d'assurances. Ces risques financiers sont généralement détectés à l'aide des algorithmes de classification.

Dans les problèmes de classification, la distribution asymétrique des classes, également connue sous le nom de déséquilibre de classe (class *imbalance*), est un défi très commun pour la détection des fraudes. Des approches spéciales d'exploration de données sont utilisées avec les algorithmes de classification traditionnels pour résoudre ce problème.

Le problème de classes déséquilibrées se produit lorsque l'une des classes dans les données a beaucoup plus d'observations que l'autre classe. Ce problème est plus vulnérable lorsque l'on considère dans le contexte des données massives (Big Data). Les données qui sont utilisées pour construire les modèles contiennent une très petite partie de groupe minoritaire qu'on considère positifs par rapport à la classe majoritaire connue sous le nom de négatifs. Dans la plupart des cas, il est plus délicat et crucial de classer correctement le groupe minoritaire plutôt que l'autre groupe, comme la détection de la fraude, le diagnostic d'une maladie, etc. Dans ces exemples, la fraude et la maladie sont les groupes minoritaires et il est plus délicat de détecter un cas de fraude en raison de ses conséquences dangereuses qu'une situation normale. Ces proportions de classes dans les données rendent très difficile à l'algorithme d'apprentissage automatique d'apprendre les caractéristiques et les modèles du groupe minoritaire. Ces algorithmes seront biaisés vers le groupe majoritaire en raison de leurs nombreux exemples dans l'ensemble de données et apprendront à les classer beaucoup plus rapidement que l'autre groupe.

Après avoir mené une étude approfondie pour examiner les défis rencontrés dans le cas de déséquilibre des classes, nous avons constaté que nous ne pouvons toujours pas atteindre une sensibilité acceptable (c.à.d. une bonne

classification du groupe minoritaire) sans une diminution significative du taux total de classification correcte. Cela conduit à un autre défi qui est le choix des mesures de performance utilisées pour valider les modèles. Le taux total de classification correcte ou la sensibilité seuls ne sont pas suffisants. Nous utilisons d'autres mesures comme la courbe de Precision-Recall ou le score F1 pour évaluer ce compromis entre précision et sensibilité.

Dans ce travail, nous avons développé deux approches : Une première approche ou classifieur unique basée sur les k plus proches voisins et utilise le cosinus comme mesure de similarité (*Cost Sensitive Cosine Similarity K-Nearest Neighbors* : CoSKNN) et une deuxième approche ou approche hybride qui combine plusieurs classifieurs uniques et fondu sur l'algorithme k-modes (*K-modes Imbalanced Classification Hybrid Approach* : K-MICHA). Dans l'algorithme CoSKNN, notre objectif était de résoudre le problème du déséquilibre en utilisant la mesure de cosinus et en introduisant un score sensible au coût pour la classification basée sur l'algorithme de KNN. Nous avons mené une expérience de validation comparative au cours de laquelle nous avons prouvé l'efficacité de CoSKNN en termes de taux de classification correcte et de détection des fraudes. D'autre part, K-MICHA a pour objectif de regrouper des points de données similaires en termes des résultats de classifieurs. Ensuite, calculez les probabilités de fraude dans les groupes obtenus afin de les utiliser pour détecter les fraudes de nouvelles observations. Cette approche peut être utilisée pour détecter tout type de fraude financière, lorsque des données étiquetées sont disponibles.

La méthode K-MICHA est appliquée dans 3 cas : données concernant la fraude par carte de crédit, paiement mobile et assurance automobile. Dans les trois études de cas, nous comparons K-MICHA au stacking en utilisant le vote, le vote pondéré, la régression logistique et l'algorithme CART. Nous avons également comparé avec Adaboost et la forêt aléatoire. Nous prouvons l'efficacité de K-MICHA sur la base de ces expériences. Nous avons également appliqué K-MICHA dans un cadre Big Data en utilisant H2O et R. Nous avons pu traiter et analyser des ensembles de données plus volumineux en très peu de temps.

Mots-clés : Fraude financière, Déséquilibre de classe, Score F1, Classification sensible aux coûts, Mesure de cosinus, K plus proche voisins, Apprentissage ensembliste, k-modes.