UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

# Techniques d'apprentissage automatique basées sur DCA avec des applications dans la finance et la santé

## THÈSE

présentée et soutenue publiquement le 24 octobre 2023

pour l'obtention du

**Doctorat de l'Université de Lorraine**

**(mention informatique)**

par

PHAM Van Tuan

**Composition du jury**

| | | |
|---|---|---|
| *Rapporteurs :* | Mustapha Lebbah | Professeur, Université Paris-Saclay |
| | Adnan Yassine | Professeur, Université Le Havre Normandie |
| *Examinateurs :* | Hanene Azzag | MCF-HDR, LIPN Université Paris 13 |
| | Yann Guermeur | Directeur de recherche, CNRS |
| *Directrice de thèse :* | Hoai An LE THI | Professeur, Université de Lorraine |
| *Co-directeur de thèse :* | Pascal Damel | Professeur, Université de Lorraine |

**Laboratoire de Génie Informatique, de Production et de Maintenance — EA 3096**

# Résumé

Les techniques d'apprentissage automatique (ML) jouent un rôle de plus en plus central dans le paysage actuel. Le développement de nouvelles techniques de ML est essentiel, étant donné leur rôle crucial dans divers domaines. Cela englobe la résolution des défis posés par les données à grande échelle et de grande dimension, ainsi que par les données déséquilibrées et la rareté des données. Cette thèse se concentre sur le développement de nouvelles techniques de ML pour résoudre des problèmes pressants dans trois domaines principaux : la gestion des données à grande échelle, la gestion des données déséquilibrées dans le domaine financier et l'atténuation de la rareté des données grâce à l'apprentissage par transfert dans le domaine de la santé. Nos méthodologies de ML proposées reposent sur la programmation DC (Différence de fonctions convexes) et les algorithmes DCA (DC), intégrant et améliorant trois techniques de ML fondamentales : SVM, bagging et l'apprentissage en profondeur par transfert.

La thèse comprend cinq chapitres : Le chapitre 1 sert d'introduction, présentant les concepts fondamentaux de ML, la programmation DC et DCA. Dans le chapitre 2, nous examinons la technique de coordonnées par bloc pour traiter les problèmes à grande échelle dans SVM dans le contexte des mégadonnées. En combinant cette technique avec DCA, nous proposons l'algorithme appelé SVM DCA à coordonnées par bloc (BC-DCASVM), qui permet à l'algorithme SVM à noyau de résoudre les problèmes liés à la grande dimensionnalité. Dans le processus d'apprentissage, l'algorithme met à jour un bloc de coordonnées (dimensions) à la fois pour réduire efficacement la valeur de l'objectif tout en maintenant les autres blocs fixes. Le chapitre 3 se concentre sur l'étude de la technique du bagging et de divers problèmes financiers, suivie de la proposition d'une solution pour relever ces défis. Le premier algorithme, appelé Bagging DCA pondéré (BaggingDCA), vise à résoudre les problèmes qui peuvent survenir lors de l'application du bagging à des problèmes d'apprentissage automatique généraux. Le deuxième algorithme intègre BaggingDCA et des techniques de sensibilité au coût dans l'algorithme de bagging, appelé Bagging DCA pondéré sensible au coût (CSB-DCA). Cet algorithme s'attaque directement à l'un des problèmes les plus difficiles en ML, à savoir les données déséquilibrées, qui affectent également de nombreuses tâches de classification financière. En incorporant BaggingDCA et la technique de sensibilité au coût, l'algorithme vise à réduire le biais induit par le déséquilibre et à améliorer les performances prédictives sur des ensembles de données financières biaisés. Nous concevons les algorithmes de bagging pondéré basés sur DCA pour être polyvalents, permettant l'utilisation de différents apprenants de base et de différentes fonctions de perte dans une conception unifiée. Le quatrième chapitre explore divers problèmes de santé, où les données textuelles médicales sont souvent rares en raison de leur sensibilité. En tirant parti des perspectives prometteuses de l'algorithme de DCA stochas-

tique à chaînes de Markov (MCSDCA), un optimiseur basé sur DCA pour l'apprentissage en profondeur, qui a été évalué sur des architectures d'apprentissage en profondeur traditionnelles, nous proposons une nouvelle architecture d'apprentissage en profondeur qui combine CNN et BiLSTM avec l'algorithme MCSDCA pour relever certains défis cruciaux dans le domaine de la santé. De plus, nous utilisons plusieurs modèles de langage pré-entraînés pour relever le défi de la rareté des données, en nous inspirant des principes de l'apprentissage par transfert. En comparant avec des optimiseurs populaires, des architectures d'apprentissage en profondeur et des modèles de langage pré-entraînés, notre méthode démontre sa compétitivité par rapport aux approches existantes. Enfin, le chapitre 5 sert de conclusion de la thèse, fournissant un résumé complet des principales conclusions et des recommandations pour les orientations futures de la recherche.

**Mots-clés:** Programmation DC (Différence de fonctions convexes) et algorithmes DCA (DC), SVM à grande échelle, Bagging pondéré, Apprentissage par transfert

# Abstract

Machine learning (ML) techniques are assuming an ever more pivotal role in today's landscape. The development of new ML techniques is essential, given their crucial role in diverse domains. This encompasses addressing challenges posed by large-scale and high-dimensional data, as well as imbalanced data and scarcity of data. This thesis focuses on the development of new ML techniques to address pressing issues in three main topics: Addressing large-scale data, handling imbalanced data in the financial domain, and mitigating data scarcity using transfer learning in healthcare. Our proposed ML methodologies are based on DC (Difference of Convex functions) programming and DCA (DC Algorithms), integrating and enhancing three fundamental ML techniques: SVM, bagging, and deep transfer learning.

The thesis comprises five chapters: Chapter 1 serves as an introduction, presenting fundamental concepts of ML, DC programming, and DCA. In Chapter 2, we investigate the block-coordinate technique to handle large-scale problems in SVM within the context of big data. By combining this technique with DCA, we propose the algorithm named Block-Coordinate DCA SVM (BC-DCASVM), which enables the kernel-SVM algorithm to address problems arising from high dimensionality. In the training process, the algorithm updates one block of coordinates (dimensions) at a time to effectively decrease the objective value while keeping the other blocks fixed. Chapter 3 focuses on studying the bagging technique and various financial problems, followed by the proposal of a solution to address these challenges. The first algorithm, called DCA weighted Bagging (BaggingDCA), aims to address issues that can arise when applying bagging to general machine learning problems. The second algorithm integrates BaggingDCA and cost-sensitive techniques into the bagging algorithm, which is named Cost-Sensitive weighted Bagging DCA (CSB-DCA). This algorithm directly tackles one of the most difficult issues in ML—imbalanced data—which also plagues many financial classification tasks. By incorporating BaggingDCA and the cost-sensitive technique, the algorithm aims to reduce imbalance-driven bias and improve predictive performance on skewed financial datasets. We design the DCA-based weighted bagging algorithms to be versatile, allowing the use of various base learners and different loss functions within a unified design. The fourth chapter explores various issues in health care, where medical text data is often scarce due to its sensitivity. Leveraging the promising prospects of the Markov-chain stochastic DCA (MCS-DCA) algorithm, an optimizer based on DCA for deep learning, which has been evaluated on traditional deep learning architectures, we propose a new deep learning architecture that combines CNN and BiLSTM with the MCSDCA algorithm to address some crucial challenges in the healthcare field. Moreover, we employ several pre-trained language models to address the data scarcity challenge, drawing inspiration from the principles of transfer learning. Through comparisons with popular optimizers, deep learning architectures, and pre-trained language models, our method demonstrates competitiveness with existing approaches. Lastly, Chapter 5

serves as the concluding chapter of the thesis, providing a comprehensive summary of the key findings and recommendations for future research directions.

**Keywords:** DC (Difference of Convex functions) programming and DCA (DC Algorithms), Large-scale SVM, Weighted bagging, Transfer learning