Skyline in Multi-Model Machine Learning

Encadrement

- Dominique H. Li (Maître de conférences HDR), Université de Tours, LIFAT
  - dominique.li@univ-tours.fr
- Nicolas Labroche (Maître de conférences HDR), Université de Tours, LIFAT
  - nicolas.labroche@univ-tours.fr
- Patrick Marcel (Professeur), Université d'Orléans, LIFO
  - patrick.marcel@univ-orleans.fr

Contexte

L'apprentissage avec des modèles multiples est une catégorie de méthodes qui utilise un ensemble de modèles au lieu d'un seul modèle pour la prédiction. Un exemple classique est « Ensemble Learning » [1]. L'avantage de ce type d'approches est de permettre d'obtenir un bon compromis entre le biais du modèle final (lié à un sous-apprentissage) et sa variance (liée à un sur-apprentissage). Dans ce stage, nous proposons une nouvelle méthode pour l'apprentissage multi-modèles basée sur le calcul itératif de skylines [2], plus connues sous le nom d'optimum de Pareto en optimisation multicritères. Cette façon d'appréhender le problème est très différente des approches classiques d'Ensemble Learning.

Les objectifs principaux sont de : (1) entraîner différents modèles avec différentes méthodes, selon le paradigme classique « training + test + cross validation » ; (2) récupérer pour chaque modèle un score attribué à chaque instance dans le jeu de test afin de décider sa classe ; (3) utiliser un calcul de skylines pour filtrer les instances correctement classées selon leurs scores donnés par les différents modèles. Cette proposition repose sur l'hypothèse que le calcul de skyline permettra efficacement de trouver le meilleur ensemble de modèles.

Une première étude sur ce sujet a été faite sur des jeux de données d'anomalie (disponibles publiquement, avec 57 modèles). Les résultats ont montré que les skylines permettent de retourner les instances anormales dans les prédictions de bi-modèles, et les meilleurs bi-modèles ont pu ainsi être déterminés. Le présent projet est la continuité directe de ce stage.

Plus précisément, nous nous intéressons aux deux points ci-dessous :

1. la propriété de dominance de skyline masque la plupart des instances correctement classées, il est donc nécessaire d'avoir un mécanisme pour supprimer les instances dominantes de manière itérative et pour trouver des conditions afin d'arrêter l'itération ;
2. les résultats actuels sont basés sur des bi-modèles, or de plus grands ensembles de modèles sont souvent plus efficaces, on pourra donc étudier comment le calcul de skylines sur des sous-espaces permet de trouver des multi-modèles plus efficaces.

Le cadre applicatif de ce stage est la détection d'anomalies, qui est un sujet important mais difficile dans des plusieurs domaines d'applications. Deux benchmarks publics PyOD (Python Outlier Detection) [3] et ADBench (Anomaly Detection Benchmark) [4] ont été reconnus et diffusés par la communauté. Avec ces benchmarks, notre objectif est de trouver les meilleures combinaisons de modèles pour améliorer la détection d'anomalies et classer les différents types d'anomalies selon les modèles retenus. Note que le problème de la détection d'anomalies est effectivement un problème typique de classification binaire, donc

notre proposition pourra s'appliquer aux problèmes similaires, notamment en Machine Learning.

Résultats attendus

- Une étude théorique sur la condition d'arrêt de notre proposition
- Une chaîne de processus pour la sélection de modèles
- Une implémentation du résultat théorique en Python avec le benchmark PyOD
- Une étude expérimentale pour prouver l'efficacité de notre proposition
- Contribuer à une publication scientifique décrivant l'approche, dans un journal international

Références

[1] Thomas G. Dietterich: Ensemble Methods in Machine Learning. Multiple Classifier Systems 2000: 1-15
[2] Dominique H. Li: Subset Approach to Efficient Skyline Computation. EDBT 2023: 391-403
[3] Yue Zhao, Zain Nasrullah, Zheng Li: PyOD: A Python Toolbox for Scalable Outlier Detection. J. Mach. Learn. Res. 20: 96:1-96:7 (2019)
[4] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, Yue Zhao: ADBench: Anomaly Detection Benchmark. NeurIPS 2022

Dates, durée et lieu

- Début de projet : le 01/02/2024
- Durée : 6 mois
- Laboratoire LIFAT, Université de Tours, site de Blois

Profile attendu des candidat(e)s

- Étudiant(e) stagiaire en Master informatique
- Bonne connaissance sur Python et Machine Learning
- Capacité de travail en autonomie

Candidature

- Une lettre de motivation
- Les relevés de notes depuis L3
- Tous les rapports pédagogiques (projets, stages, etc.) et éventuellement des publications
- Contact : dominique.li@univ-tours.fr

Rémunération

- Environs 600 euros/mois selon le règlement.