Cahier des charges

PIC : Optimisation de sélection des attributs

Réalisé par :
SKANDRANI Yasmine
ERRAZKI Safae
ED-DAHNI Hafsa

—

Sous la supervision de :

M. Arnaud LIEFOOGHE

Contexte

Le projet PIC vise à optimiser la sélection d'attributs dans le domaine de l'apprentissage automatique (ML), en se concentrant sur les méthodes de type *wrapper*, notamment la méthode de Sélection Séquentielle d'Attributs (Sequential Feature Selection - SFS). La sélection d'attributs est cruciale pour améliorer les performances des modèles prédictifs, car elle réduit la dimensionnalité des données et minimise le risque de surapprentissage. Bien que SFS soit une approche efficace, elle est limitée en précision par rapport aux méthodes exhaustives, qui, bien que plus précises, demeurent coûteuses en termes de calcul.

Ce projet se concentrera sur l'évaluation expérimentale et la comparaison de divers algorithmes en fonction des caractéristiques spécifiques du problème, en examinant trois éléments clés : le jeu de données, le modèle d'apprentissage et les scores de performance. Notre objectif est de proposer une amélioration de la méthode SFS en intégrant un algorithme d'optimisation plus performant, capable d'atteindre des résultats proches de ceux d'une recherche exhaustive, tout en réduisant les coûts computationnels. En explorant les relations entre le jeu de données, le modèle d'apprentissage et les scores de performance, le projet vise à fournir une compréhension approfondie des défis d'optimisation et des solutions possibles en sélection d'attributs.

Problématique

La sélection d'attributs est fondamentale pour simplifier et rendre plus interprétables les modèles d'apprentissage automatique, surtout dans le contexte de l'Intelligence Artificielle Explicable (XAI). En réduisant le nombre d'attributs aux plus pertinents, on limite la complexité des modèles tout en optimisant leur performance en termes de calcul et de précision. Ce besoin de simplicité et d'efficacité est d'autant plus important dans des domaines critiques tels que la reconnaissance d'images et le diagnostic médical.

Cependant, la recherche du sous-ensemble optimal d'attributs pour maximiser les performances est coûteuse en ressources. Les méthodes séquentielles, comme la sélection en avant ou en arrière, sont efficaces mais présentent des limites dans leur approche locale et ne garantissent pas toujours le résultat optimal pour l'ensemble de données et le modèle donné. De nouvelles techniques d'optimisation, telles que les algorithmes évolutionnaires et des méthodes d'apprentissage automatique intégrées, peuvent apporter des solutions pour relever ces défis en améliorant la précision et en réduisant les coûts.

Objectifs du projet

1. Optimisation de la sélection d'attributs :

- Développer un modèle de sélection d'attributs optimisé utilisant la méthode **wrapper**, appliqué aux algorithmes de Sélection Séquentielle d'Attributs dans ses deux variantes :
  - **Forward Sequential Feature Selection (FSFS)** : ajout progressif des attributs les plus pertinents.
  - **Backward Sequential Feature Selection (BSFS)** : retrait progressif des attributs les moins pertinents.
- Comparer les performances des méthodes FSFS et BSFS avec celles de la méthode exhaustive, en mesurant l'écart (gap) entre leurs scores de précision respectifs.

2. Réduction des coûts de calcul :

- Concevoir un modèle performant qui maximise la précision tout en étant plus efficace que la méthode exhaustive en termes de calcul, en se rapprochant au plus près des résultats optimaux.

3. Fonction de sélection flexible :

- Créer une fonction de sélection d'attributs prenant en paramètre :
  - Un **score** de classification (selon le choix de l'utilisateur) comme critère de performance.
  - Les **attributs** du dataset (allant de 6 à 18 attributs) comme prédicteurs du modèle.
  - Le **modèle** de machine learning (K-Nearest Neighbors, Naive Bayes, Random Forest, etc.) pour tester le sous-ensemble sélectionné.
- Cette fonction flexible permettra d'adapter le modèle aux spécificités de différents jeux de données et algorithmes de machine learning.

4. Utilisation de la validation croisée :

- Intégrer une validation croisée dans le processus de sélection pour réduire le risque de surapprentissage (overfitting), en particulier pour certains modèles sensibles aux biais.

5. Évaluation et comparaison des algorithmes :

- Comparer les performances des algorithmes FSFS et BSFS avec la méthode exhaustive sur différents types de datasets.
- Analyser les résultats en fonction des caractéristiques des datasets et des modèles de machine learning utilisés.
- Mesurer le gap entre les méthodes FSFS, BSFS et la méthode exhaustive pour évaluer leur capacité à se rapprocher des performances optimales tout en étant moins coûteuses en calcul.