
Sujet n°1 : *Induction d'arbres de décision, élagage et boosting*

Résumé

Le but de ce projet est de comparer des méthodes d'induction d'arbres de décision avec élagage et l'emploi du boosting utilisant l'induction d'arbres de décision comme méthode d'apprentissage faible.

1. La première étape du projet consistera à programmer la méthode d'induction d'arbre de décision (en C/C++ ou en Matlab).

On complétera alors cet algorithme par une méthode d'élagage automatique (telle que Reduced Error Pruning ou Pessimistic Pruning) qui sera paramétrable de telle manière que suivant la valeur du paramètre à un extrême on conserve l'arbre d'origine et à l'autre on ne garde que le seul nœud racine.

On fera alors des expériences sur l'une des bases de données disponibles dans le répertoire UCI en comparant à chaque fois pour plusieurs valeurs du paramètre la performance en classification et en généralisation.

2. Dans un deuxième temps, on implémentera une technique automatique de réglage du paramètre d'élagage par validation croisée.
 - i. On testera alors sur la base de données choisie lors de l'étape 1. On fera attention de conserver un ensemble de validation.
 - ii. Après le choix automatique du meilleur paramètre, on mesurera la performance résultante sur les données de l'ensemble de validation.

3. Dans un troisième temps, on implémentera la méthode Adaboost. On l'utilisera alors sur les données sélectionnées lors de l'étape 1 en utilisant comme méthode d'apprentissage faible l'algorithme d'induction d'arbre de décision sans élagage développé dans la première étape.

On comparera alors les résultats obtenus avec ceux obtenus à l'issue de la deuxième étape.

On discutera les résultats.

On pourra recommencer ces expériences avec d'autres bases de données.

On rendra un rapport final rendant compte de la démarche suivie, des expériences réalisées et des leçons à en tirer. Ce rapport fera de 10 à 20 pages.

Une démonstration sur machine aura lieu devant l'enseignant chargé de l'encadrement ainsi qu'une soutenance orale de 15' pour présenter le travail réalisé devant les autres étudiants.