

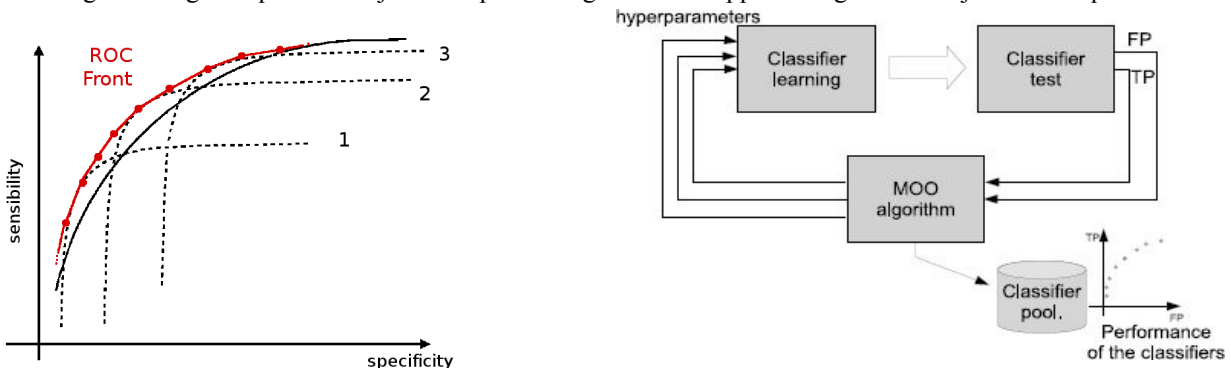
Stage de Master 2 Recherche

Optimisation multiobjectif et apprentissage

Encadrants : Clément Chatelain, Sébastien Adam & Laurent Heutte

Contexte : Dans de nombreux domaines, les performances d'un classifieur sont souvent évaluées par un taux d'erreur qui est également utilisé comme critère d'apprentissage du système. Cependant, lorsque les coûts de mauvaise classification sont déséquilibrés, ce critère n'est pas adapté : par exemple, en médecine, dans le cas de la détection d'une maladie rare, décider que toutes les personnes sont saines produira un taux d'erreur très faible puisque les exemples malades sont très rares. On peut donc s'interroger sur la pertinence d'un tel critère... Il est alors indispensable de distinguer deux types d'erreur pour qualifier les performances du système, et donc de considérer deux critères pour l'apprentissage d'un système : le taux de Fausse Alarme (décider qu'une personne saine est malade) et le taux de Faux Rejet ou de Non Détection (décider qu'une personne malade est saine). L'apprentissage d'un tel système nécessite de se placer dans un contexte dit « multiobjectif » permettant d'obtenir, non plus un classifieur unique, mais un ensemble de classifieurs, chacun de ces classifieurs proposant un compromis « Fausse Alarme / Faux Rejet » optimal.

Lors de nos travaux récents [Chatelain06, Chatelain07], nous avons mis au point une méthode originale permettant d'obtenir efficacement cet ensemble optimal que nous avons baptisé ROC front (cf. figures). Cette méthode est basée sur un algorithme génétique multiobjectif couplé aux algorithmes d'apprentissage mono-objectif classiques.



Maintenant que nous disposons de ce front, la question qui se pose est la suivante : est-il possible de faire encore mieux ? Selon nous, la réponse à cette question pourrait résider dans la combinaison des classifieurs du front ROC. En effet, la mise en oeuvre d'une méthode de combinaison pourrait éventuellement nous permettre d'obtenir localement des classifieurs plus performants.

Etude proposée: Il s'agira d'aborder ce problème de combinaison dans le contexte plus spécifique de la sélection statique de classifieurs: étant donné un ensemble de classifieurs du front ROC, comment sélectionner un sous-ensemble de classifieurs à combiner de telle sorte que la courbe ROC du sous-ensemble soit localement meilleure que le front ROC? Se posent alors les problèmes :

- du choix des classifieurs à combiner: quel critère appliquer pour sélectionner les classifieurs (diversité, performance,...)? Combien en retenir?... On pourra s'inspirer des nombreux travaux de la littérature sur la sélection statique de classifieurs ou la sélection de caractéristiques;
- du choix de l'opérateur de combinaison: quelle est l'influence de l'opérateur sur les performances? Peut-on ou doit-on appliquer un opérateur différent en fonction des classifieurs sélectionnés? Comment combiner un sous-ensemble de classifieurs à partir de leur courbe ROC? On pourra notamment s'inspirer des travaux présentés dans [Caruana04, Coetzer07] pour construire l'opérateur de combinaison.

Pour évaluer la robustesse des différentes solutions retenues, les expériences seront menées sur des problèmes de reconnaissance de chiffres manuscrits et sur plusieurs bases de données de UCI Repository, qui regroupe un échantillon représentatif de différents problèmes de reconnaissance de formes.

Références

- [Chatelain06] C. Chatelain, "Extraction de séquences numériques dans des documents manuscrits quelconques," Thèse de Doctorat, Université de Rouen, 2006.
- [Chatelain07] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet, "Multi-Objective Optimization for SVM Model Selection," in *International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007*
- [Caruana04] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. « Ensemble selection from libraries of models ». In *proceedings of ICML*, 2004.
- [Coetzer07] H. Coetzer and R. Sabourin, « A Human-Centric Off-Line Signature Verification System », in *International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007*