

Fiche Stage

Intitulé du Stage

Combinaison de classifieurs

Mission:

Ces dernières années, les recherches dans le domaine de l'apprentissage automatique ont montré que l'utilisation d'un ensemble de modèles permettait d'améliorer de manière significative les performances obtenues en classification par rapport à l'utilisation d'un classifieur unique. Intuitivement, un ensemble de classifieurs permet de réduire le nombre d'exemples qui seraient mal classés par un prédicteur unique si ses membres commettent des erreurs différentes. La comparaison entre les algorithmes basés sur le Bagging et le Boosting en est l'illustration la plus évidente.

Même s'il semble que fusionner les décisions provenant de plusieurs classifieurs spécialisés induit généralement une réduction de l'erreur plus importante que celle obtenue avec un ensemble de modèles créés séparément (donc aléatoirement), au cours de ce stage on s'intéressera uniquement aux méthodes où il est possible d'apprendre chacun des membres de l'ensemble de manière individuelle (au contraire donc des méthodes du type Adaboost). L'idée étant de répartir leur apprentissage dans un système HDFS.

Le but du stage sera de tester certaines des méthodes de l'état de l'art.

Documentation :

- Le cours de Laurent Huette : <http://www.univ-rouen.fr/psi/heutte>
- Les articles sur ce sujet de Gavin Brown (<http://www.cs.man.ac.uk/~gbrown/>) ou Lundvina Kuncheva (<http://pages.bangor.ac.uk/~mas00a/>)
- La thèse de David Gacquer de 2008 «Sur l'utilisation active de la diversité dans la construction d'ensembles de classifieurs. Application à la détection de fumées nocives sur site industriel »

Profil:

- Le profil souhaité est BAC + 5, Master Industriel informatique et/ou statistiques ou école d'ingénieur.
- Intérêt pour les aspects applicatifs et théoriques du sujet.

Compétences

- Le profil souhaité est BAC + 5, **Master Industriel informatique** et/ou statistiques ou école d'ingénieur.
Intérêt pour les aspects applicatifs et théoriques du sujet.
- Le stagiaire devrait avoir une « bonne » connaissance des systèmes **HDFS**.
- Les connaissances en Matlab (ou équivalent) et Python seraient les bienvenues.
- Des connaissances en **statistiques**, mathématiques et/ou apprentissage statistique sont un plus.

Modalités

5 mois, printemps-été 2012, Lannion (Bretagne)

Le plus de l'offre

Proche de la mer, vous serez dans l'équipe de traitement des données d'Orange Labs directement en lien avec des problématiques opérationnelles du groupe Orange sur le CRM et l'Audience. Le stagiaire évoluera dans un contexte très recherche sur un sujet très porteur. Il sera intégré dans l'URD au sein d'une équipe recherche. Rémunération de l'ordre de 1000€ net.

Contacts

Nicolas Voisine – 02 96 05 28 16 – nicolas.voisine@orange.com
Vincent Lemaire - 02 96 05 31 07 - vincent.lemaire@orange.com