

TRAVAUX DE GROUPE AIA 2009-2010

Les travaux seront, selon le sujet, réalisés par deux ou trois étudiants travaillant en groupe. Ils devraient représenter environ 20-30h de travail personnel par étudiant, en ce compris la préparation d'un bref rapport individuel et d'une présentation des résultats obtenus par groupe.

Nous avons imaginé trois types de sujets/travaux:

1. Approfondissement de certains développements théoriques récents ou moins récents en apprentissage automatique (lecture d'articles, analyses critiques)
2. Développement de certains algorithmes récemment publiés dans la littérature et validation empirique.
3. Applications originales de l'apprentissage automatique (image, bioinformatique, etc).

Règles du jeu:

Les étudiants se constituent en groupes de deux ou trois et nous contactent (L.Wehenkel@ulg.ac.be ; P.Geurts@ulg.ac.be) avant le 24 novembre 2008 pour nous signaler le sujet de travail qu'ils auront choisi, en indiquant la composition du groupe et en précisant brièvement comment ils comptent se répartir le travail (mettre comme sujet du mail "TRAVAIL DE GROUPE AIA").

Chaque groupe réalise le travail en commun et prépare une présentation de 15 minutes (avec transparents). En cas de questions, il peut contacter par email l'auteur du sujet du travail en nous mettant en copie (mettre comme sujet du mail "TRAVAIL DE GROUPE AIA").

Chaque étudiant fera aussi un bref rapport (limité strictement à 4 pages) précisant le travail qu'il a réalisé au sein du groupe, les sources qu'il a utilisées à cette fin (articles, logiciels, wikipedia, notes de cours, etc.) et ses principales conclusions à l'issue du travail.

Si plusieurs groupes choisissent le même sujet, ils peuvent éventuellement communiquer entre eux sur le sujet, s'ils le souhaitent. Si c'est le cas, nous leur demandons de le signaler clairement dans les rapports individuels et lors de la présentation orale, en indiquant sur quels points il y a eu collaboration et avec quel autre groupe.

Le but de ces travaux est de vous aider à vous "approprier" la matière vue au cours en vous poussant à aller au delà d'une approche purement scolaire "cours oral - syllabus - examen". Tenez aussi compte du fait que, lors de la présentation orale, vous vous adresserez autant aux autres étudiants qui suivent le cours qu'aux profs.

La date de remise des travaux sera déterminées lors du cours du 23 novembre. Rapports, codes source, et transparents pour la présentation devront être envoyés par mail. Tous les lundi PM, de 14 à 16 heures, nous sommes à votre disposition au local usuel pour répondre à vos questions et vous aider à progresser. On vous demande cependant d'envoyer un mail avant le dimanche soir minuit qui précède pour signaler votre présence à cette séance de questions.

Ci-dessous, vous trouverez une liste de sujets que nous proposons pour cette année. Si vous avez vos propres idées sur un sujet qui vous intéresse plus particulièrement et qui n'est pas repris, n'hésitez pas à nous consulter (un mail, avec le bon sujet, adressé à PG et LW fera l'affaire, mais il ne faudrait pas traîner).

1. Recherche de règles d'association par arbres de décision

Le but d'une analyse par règles d'association est de trouver des combinaisons de valeurs d'attributs qui apparaissent fréquemment dans une base de données. Ces méthodes peuvent par exemple être utilisées pour trouver les articles achetés fréquemment simultanément par les clients d'un super-marché (dans le but par exemple de prendre en compte cette information pour réarranger les rayons du magasin).

L'objectif de ce travail sera de valider un système de règles d'association basé sur des arbres de décision tel que décrit dans le chapitre 14 du livre de référence du cours (14.2.5 et 14.2.6). Le travail commencera par une lecture de la section 14.2 du livre. Ensuite, un logiciel sera développé pour générer une version permutée d'une base de données. Des tests seront ensuite réalisés sur une ou plusieurs bases de données (fournies éventuellement). Pour les arbres de décision, les étudiants pourront utiliser le logiciel PEPITo.

Contact: p.geurts@ulg.ac.be

2. Apprentissage sur expression de gènes par comparaison de paires d'attributs

D. Geman et ses collègues ont proposé une méthode relativement simple de classification basée sur des règles comparant des paires d'attributs entre eux [1,2]. Cette méthode semble donner de bons résultats sur des données d'expression de gènes.

Le but de ce travail sera de faire une implémentation simple de cette méthode et de la comparer aux méthodes d'apprentissage disponibles dans pepito (SVM, ensemble d'arbres, etc.) sur un ou plusieurs problèmes tests.

[1] Geman et al. Classifying gene expression profiles from pairwise mRNA comparisons. Statistical applications in genetics and molecular biology (2004) vol. 3 pp. Article 19

http://cis.jhu.edu/publications/papers_in_database/GEMAN/SAGMB_04.pdf

[2] Tan et al.. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics (2005) vol. 21 (20) pp. 3896-3904

<http://bioinformatics.oxfordjournals.org/cgi/reprint/21/20/3896>

Contact: p.geurts@ulg.ac.be

3. Implémentation et validation de la méthode RELIEF

L'algorithme RELIEF est une méthode très populaire de sélection de variables pour l'apprentissage supervisé. Le but de ce travail est d'implémenter différentes variantes de cette méthode et d'évaluer leur comportement sur quelques bases de données. La méthode sera comparée et combinée aux méthodes d'ensemble d'arbres de décision.

http://www.tsi.enst.fr/~campedel/Biblio/FeatureSelection/robnik03-mlj_RReliefF.pdf

Contact: p.geurts@ulg.ac.be

4. Problèmes de déséquilibre de classes

Dans beaucoup de problèmes de classification réel, la base de données est déséquilibrée: les différentes classes ne sont pas représentées de manière équitable dans l'ensemble d'apprentissage. Un déséquilibre trop important affecte généralement négativement la précision des algorithmes d'apprentissage (qui ont tendance à favoriser la classe majoritaire) et différentes méthodes ont été proposées dans la littérature pour pallier à ce problème (voir [1,2]).

La première partie du travail consistera à comparer la robustesse des différentes méthodes d'apprentissage au déséquilibre de classe sur différents jeux de données artificiels. Ensuite, une méthode simple de correction du déséquilibre sera implémentée (par exemple, celle proposée dans [1]) et son effet sera évalué sur différentes méthodes.

[1] SMOTE: Synthetic Minority Over-sampling Technique, N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Journal of Artificial Intelligence Research, 16 (2002), 321-357.

<http://www.jair.org/media/953/live-953-2037-jair.pdf>

[2] A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. Gustavo E. A. P. A. Batista, Ronaldo C. Prati, Maria Carolina Monard.

<http://www.sigkdd.org/explorations/issues/6-1-2004-06/batista.pdf>

Contact: p.geurts@ulg.ac.be

5. Classement automatique de fichiers PDF

Le but de ce travail est de mettre au point un système de classement automatique de fichiers PDF basé sur les algorithmes de clustering. Un programme sera mis au point pour transformer chaque document pdf en un vecteur d'attributs (correspondant par exemple chacun à la fréquence d'apparition d'un mot particulier dans le document). Ensuite, un ou plusieurs algorithmes de clustering seront utilisés pour regrouper les documents en différentes classes. L'analyse en composantes principales (ou tout autre méthode) sera ensuite utilisée pour visualiser les groupes mis en évidence. Le choix du corpus de documents est laissé libre.

La librairie "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering" disponible sur <http://www.cs.cmu.edu/~mccallum/bow/> offre un ensemble d'outils (écrits en C) qui pourraient être utiles à ce travail.

Contact: p.geurts@ulg.ac.be

6. Application du LASSO à la classification de données biologique

Le LASSO est une méthode de régression linéaire qui a la particularité de rechercher un modèle économe ("sparse") dans les attributs qu'il utilise réellement. Cette méthode est donc souvent utilisée pour faire de la sélection de variables en apprentissage supervisé. Le but de ce travail est d'appliquer cette méthode à des données biologiques (microarray, protéomique, ou SNP). Le travail pourra soit se focaliser sur l'implémentation d'une

variante de cette méthode, soit sur l'application d'une implémentation existante à des données réelles.

Références:

- <http://www.cs.ubc.ca/~schmidtm/Software/lasso.html>
- http://www-stat.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf

Contact: p.geurts@ulg.ac.be

7. Classification de documents par Bag-of-Words

SVM-light est une implémentation open-source de référence pour les machines à support vectoriel. Elle est disponible sur <http://svmlight.joachims.org/> avec documentation, exemples, et références bibliographiques.

Une première partie du travail (25%) consistera à présenter, à partir de l'exemple de la classification de texte disponible sur la page web (à la rubrique "Inductive SVM") la classification par machine à support vectoriel, les sorties fournies par svm-light, la façon de les exploiter pour la classification, et les options d'estimation de performance disponibles sur svm-light.

Une seconde partie du travail (75%) consistera à présenter les bases de données "Reuters-21578" et "RCV1-v2" disponibles sur <http://www.daviddlewis.com/resources/testcollections/reuters21578/> http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

et à résoudre, à l'aide de SVM-light, un problème de classification de texte sur ces bases de données, par exemple l'apprentissage d'un classificateur de documents appartenant à la catégorie dite "CCAT".

Contact: Boris Defourny (bdf@montefiore.ulg.ac.be)

8. Classification automatique d'images par boosting

Évaluation et comparaison des performances d'algorithmes de Boosting sur une représentation par "sac-de-mots" des images (les images sont représentées par des vecteurs qui décrivent des morceaux de l'image, aussi appelés des patches) et identification des images "difficiles".

Suggestion: évaluer les algorithmes et outils de visualisation disponibles dans JBoost: <http://jboost.sourceforge.net/index.html>

Note: Images et représentation seront fournies.

Contact: raphael.maree@ulg.ac.be

9. "One-class SVM" pour la classification d'images

Dans certaines applications, on dispose d'exemples d'une seule classe et l'apprentissage consiste à apprendre une fonction dont la valeur est positive pour les objets de cette classe et négative partout ailleurs. La méthode présentée dans [1] permet de faire ça à la manière des machines à vecteurs de support. Ces approches sont particulièrement intéressante lorsqu'il s'agit de résoudre un problème de classification où le nombre de classes est très important et éventuellement pas connu à l'avance: un modèle "one-class" est appris pour chacune des classes séparément et on fait ensuite voter ces différents modèles lorsqu'il s'agit de faire de classer un nouvel objet.

Ce travail consistera d'abord à s'informer sur la méthode [1]. Ensuite, cette méthode sera appliquée pour faire de la classification multi-classe sur une base de données de caractères manuscrit et comparée à une approche SVM multi-classe plus classique. Le choix du logiciel à utiliser est libre mais une implémentation efficace de la méthode est fournie dans la librairie libsvm [2].

[1] Estimating the support of a high-dimensional distribution. Scholkopf, Platt, Shawe-Taylor, Smola, and Williamson. Neural Computation 13, 1443-1471, 2001
<http://axiom.anu.edu.au/~williams/papers/P132.pdf>

[2] <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Contact: p.geurts@ulg.ac.be, raphael.maree@ulg.ac.be

10. Etude bibliographique et comparaison des méthodes de réduction de dimension

La première partie de ce travail consistera à lire l'article [1] qui fait une revue de l'état de l'art dans le domaine des méthodes de réduction de dimension. La seconde partie du travail visera à comparer empiriquement et de manière critique au moins 3 de ces méthodes sur une base de données dont le choix est laissé libre. Pour cette comparaison, les étudiants pourront utiliser la toolbox matlab développée par les auteurs de l'article [2].

[1] van der Maaten et al. Dimensionality reduction: a comparative review. (2008)
http://www.iai.uni-bonn.de/~jz/dimensionality_reduction_a_comparative_review.pdf

[2]
http://ict.ewi.tudelft.nl/~lvandermaaten/Matlab_Toolbox_for_Dimensionality_Reduction.html

Contact: p.geurts@ulg.ac.be

11. Théorie de l'apprentissage et bornes en classification

Ce travail théorique consistera en une lecture approfondie et critique de l'article suivant:

Tutorial on Practical Prediction Theory for Classification: John Langford; 6(Mar):273--306, 2005.

<http://www.jmlr.org/papers/volume6/langford05a/langford05a.pdf>

Cet article fait un survol des travaux qui cherchent à prédire les performances des algorithmes de classification en fournissant par exemple une borne supérieure sur le taux d'erreur d'un classificateur.

Contact: p.geurts@ulg.ac.be

12. Etude théorique du Boosting

Ce travail théorique consistera en une lecture approfondie et critique des articles suivants:

- Schapire. The boosting approach to machine learning an overview. Nonlinear Estimation and Classification (2003) pp. 23
- Freund et al. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences (1997)
- Bühlmann et al. Rejoinder: Boosting Algorithms: Regularization, Prediction and Model Fitting. Statist. Sci. (2007) vol. 22 (4) pp. 516-522

Les 2 premiers donnent la vision Schapire-Freund, le troisième la vision des statisticiens.

Il s'agira de lire les articles (et éventuellement les références qu'ils citent, si nécessaire) pour en faire une analyse fine et critique. Un travail de synthèse visera ensuite à comparer les deux points de vue.

Contact: p.geurts@ulg.ac.be