Proposition de stage recherche en laboratoire 2015-2016

Titre : Modèle de Markov Caché par Apprentissage semi-contraint.

Fiche - Résumé :

Le sujet de stage est orienté sur la partie apprentissage dynamique d'un système de prédiction d'événements extrêmes ou récurrents basés sur des systèmes multi-capteurs.

Le but du stage est d'estimer automatiquement les paramètres d'un modèle de Markov Caché par une technique de classification semi-supervisée. La partie apprentissage non supervisé du modèle de Markov Caché a déjà été développé (en langage R).

La continuité du sujet de M2 comme un sujet de thèse est possible.

Compétences souhaitées :

- Bonnes connaissances en Machine Learning.
- Programmation R

Encadrement :

- Emilie CAILLAULT, André BIGAND.

Lieu et rémunération :

- lieu : Laboratoire LISIC – ULCO – Calais.
- rémunération : @ 500€/mois
- durée de stage : 4 à 6 mois

Contact :

- caillault@lisic.univ-littoral.fr

Mos clés : Modèle de Markov Caché, classification semi-supervisée, apprentissage dynamique.

Détails du stage.

Contexte et objectif :

Rousseeuw et al. [1] ont montré théoriquement et expérimentalement qu'il est possible de générer un ***Modèle de Markov Caché (MMC)*** par apprentissage non supervisé. L'utilisation de la théorie associée à la classification spectrale a permis de mettre en avant une structure et caractérisation du MMC cohérente vis-à-vis des observations sans aucune intervention de paramétrage, sans aucun *a priori* sur la structure. Cette technique a été appliquée pour la première fois pour modéliser la dynamique des efflorescences phytoplanctoniques en Manche orientale et détecter des états particuliers à partir d'une base de données dite « haute résolution » dans le domaine de l'observation marine côtière (fréquence : 20 min.).

La prise en compte des connaissances même à des échelles différentes permettrait d'améliorer le système de modélisation et prédiction des états. Récemment, ***l'apprentissage semi-supervisé*** a reçu une attention toute particulière dans les approches discriminantes tels que les classifieurs spectraux ou machines à vecteurs supports. Cet apprentissage semi-supervisé permet de faire intervenir des connaissances a priori dans un processus de ***décision à partir de contraintes sur les données*** (association de points appartenant à la même classe/état ou exclusion) ou de ***labellisation réduite*** (quelques états sont connus). Un ensemble de travaux [2,3] ont montré la puissance de ces approches pour réaliser des partitionnements entre données, cohérents à la fois vis-à-vis de la structure géométrique des données et la connaissance existante. Il est alors intéressant d'étendre cela à la modélisation de données spatio-temporelles.

L'apprentissage d'un modèle de Markov Caché par apprentissage non supervisé ouvre une porte considérable pour traiter des applications réelles où la taille des séries temporelles collectées est tel qu'il n'est plus possible de demander à ces échelles, vu la multiplicité et variété des capteurs/ mesures, un étiquetage à dire d'expert de chaque évènement. Un apprentissage supervisé est donc inimaginable et source d'erreurs importantes. L'apprentissage semi-supervisé reste la piste la plus cohérente puisque nous pouvons disposer de quelques connaissances a priori et intégrer par classification spectrale la géométrie des données.

L'objectif de ce stage est ainsi d'une part d'évaluer l'applicabilité des techniques d'apprentissage semi-supervisé, notamment de ***classification spectrale contrainte*** afin de déterminer un MMC fidèle à la connaissance existante à la fois en terme de ***structure et de temporalité*** et, d'autre part, de proposer un système robuste à des événements/états extrêmes ou non appris. Ces états nécessitent la définition d'un ***rejet*** lors de la prédiction d'une nouvelle donnée éloignée des observations/données existantes et la mise en place d'un ***apprentissage dynamique*** de ces nouveaux états.

Cadre de ce projet dans l'équipe IMAP et les collaborations LISIC

Ce projet de stage s'inscrit dans une thématique forte de l'équipe IMAP, ***l'apprentissage automatique*** et un projet phare de l'université : l'***Environnement*** au travers

de collaborations fortes passées et récentes avec l'IFREMER-LER BL/LISIC (convention d'accueil de Emilie Poisson Caillault de sept. 2014 à août 2016), convention de collaboration ULCO/LISIC-Ifremer/LER BL- Agence de l'Eau Artois Picardie signée le 13 octobre 2015. Ce projet s'inscrit dans la continuité de la dynamique de recherche engagée dans le cadre des activités du **Groupement d'Intérêt Scientifique (GIS) « Campus International de la Mer et de l'Environnement Littoral »**.

D'autre part, ce projet s'inscrit aussi dans le **projet ARCUS 2**, déposé à l'ULNF avec le Maroc, la Palestine et le Liban. H. Hijazi [4,5] a mis au point des outils d'interprétation et de visualisation augmentée de données, par l'intermédiaire de méthodes de réduction de la dimension pour l'analyse exploratoire de données multidimensionnelles. Ces méthodes sont étendues à l'apprentissage semi-supervisé (thèse de co-tutelle de M. A. Darwich démarrée en 2014) et devraient être appliquées au modèle de Markov Caché par Apprentissage semi-contraint sans problème par le biais de ce projet.

Ce projet bénéficiera également de la dynamique générée par des projets complémentaires, comme le **CPER-MARCO** « Recherches marines et littorales en Côte d'Opale : des milieux aux ressources, aux usages et à la qualité des produits aquatiques qui se veut être un projet structurant multi-laboratoires, multi-organismes associant la mise en place d'instruments et d'outils (enquêtes, indicateurs) pour une approche globale de l'étude du milieu marin, de la ressource et de la qualité des produits aquatiques » et le projet **H2020 JERICO-Next** (New European eXpertise for coastal observatories, http://www.jerico-fp7.eu/) avec la collaboration étroite de Alain Lefebvre, responsable du Laboratoire IFREMER Environnement Ressources de Boulogne-sur-mer.

Bibliographie associée.

[1] Rousseeuw, K., Poisson-Caillault, E., Lefebvre, A. and Hamad, D. "Hybrid Hidden Markov Model for Marine Environment Monitoring", in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, doi 10.1109/JSTARS.2014.2341219. 20 août 2014. (IF: 2.86)

[2] Wacquet, G., Poisson Caillault E., Hébert P.A., "Semi-supervised K-Way SpectralClustering with Determination of Clusters », in *Computational intelligence, Series Studies in Computational Intelligence*, Springer, vol 465, pp. 317-330, isbn={978-3-642-35637}, 2013.

[3] Wacquet, G., Poisson Caillault E., Hamad, D., Hébert P.A., 'Constrained Spectral Embedding for K-Way Data Clustering », in Pattern Recognition Letters, d*oi:10.1016/j.patrec.2013.02.003,* v.34 n.9, p.1009-1017, July, 2013 *2013.*

[4] H.Hijazi, O.Bazzi, A.Bigand : " A new nonlinear discriminant analysis algorithm using a combined version of LDA and LLE", Congrès "ICPV2011 ", pp. 106 à 109, 18-23 juillet 2011, Las Vegas, USA.

[5] H.Hijazi, O.Bazzi, A.Bigand : " Out of samples extensions for SC-LLE, new non-linear dimensionality reduction algorithm", Congrès "ICCIT2013 ", 19-21 juin 2013, Beyrouth, Liban.