FLO TEAM

Résumé du compte-rendu de la team FLO pour le projet GeoLifeClef 2018 (consultable sur <http://ceur-ws.org/Vol-2125/>)

La team FLO a proposée quatre modèles prédictifs pour la prédiction des espèces les plus probables :

* Un modèle de réseau de neurones convolutif n’utilisant que les tenseurs environnementaux, utilisant une architecture spécifique. Ce modèle a donné les meilleurs résultats de Geolifeclef (excluant les ‘fusions’). Il n’a pas utilisé les localisations des points mais a pu apprendre leur information de lui même en trouvant les motifs parmis les données environnementales.
* Un réseau de neurone exploitant les points tombant à la position la plus proche (ressemble à knn ?) c’est-à-dire les co-occurrences
* Un classifieur ‘closest-location’ (knn) n’utilisant que coords. spatiales. (La team ST ayant aussi utilisée un knn sur les géolocalisations + tenseurs environnementaux plus performant)
* Une forêt aléatoire basée seulement sur les coords. spatiales. C’est le meilleur modèle basé seulement sur la géoloc.

Elle a aussi combinée ses modèles par des méthodes de ‘fusions’ basée sur des probabilités prédictives et rangs prédictifs (?). La fusion cnn + forêt à est le meilleur modèle de Geolifeclef 2018 bien qu’il n’est pas significativement meilleure que le Cnn seul statistiquement parlant (MRR +0.0005).

La team FLO s’est inspirée de l’apparition de deep neural networks dans les problèmes de “species distribution modeling”, qui ont montré une efficacité potentiellement meilleure que les modèles conventionnels :

Christophe Botella, Alexis Joly, P.B.P.M., Munoz, F.: A deep learning approach to species distribution modelling. Multimedia Technologies for Environmental & Biodiversity Informatics (2018)

Les modèles conventionnels sont les arbres boostés (BT/ XG-Boost) utilisés abondamment pour les pb. de “species distribution modeling” basés sur données environnementales.

Leur Cnn possède une architecture customisée utilisant le softmax loss (softmax layer et categorical loss – pour prédire l’espèce la plus probable d’être observée à partir des features environnementales apprises)

Ils ont découper les tenseurs environnentaux en séparant des variables qualitatives des quantitatives (Pourquoi?)

Ils ont enfin définie leur propres architecture de couches de convolution détaillée dans la Figure 1 de leur papier.

Informations sur le CNN : Implémentation avec le framework PyTorch, apprentissage sur GPU Nvidia GTX 1080 ti, et descente de gradient stochastique pour l’apprentissage.

Leur second modèle, un réseau de neurones, est basé sur les co-occurrences dont le but est de prédire la vraisemblance de la présence d’une certaine espèce connaissant la présence d’autres espèces, afin de capturer les relations d’interdépendance entre espèces. Leur réseau de neurones *feed-forward* (?) prenant en entrée des vecteurs représentant l’abondance des espèces, et en sortie les labels d’espèces ciblées. Les vecteurs d’abondance ont été construit en aggégeant les occurrences situés les plus à proximitées, et en calculant les fréquences de chaque espèce dans cet ensemble aggrégé.

Ils ont utilisés ensuite un perceptron multicouches avec une couche cachée de 256 neurones totalement connectés (avec fonctions d’activations ReLU et batch normalization (?) pour la couche cachée, et fonction softmax loss comme sortie du modèle).

Implémenté en PyTorch and l'optimiseur Adam.

Pour le modèle de forêt aléatoire, il ont utilisés le modèle implémenté dans Scikit-Learn. Simplement basé sur les coordonnées spatiales en entrée, les sortie sont les labels d’espèces. Pour une occurrence de test, le modèle prédit une liste ordonnée des espèces les plus probables en cette localisation. La forêt aléatoire était composée de 50 arbres de prof. 8).