ETAT DE L’ART

1. OVERVIEW

But du challenge est de prédire une liste d'espèces sachant une localisation sur une carte.

On possède une grande base de données d'espèces pour entraîner nos modèles, et chaque occurrence est accompagné d'un ensemble d’images caractérisant l'environnement.

Cependant il n'est pas possible d'apprendre la distribution des espèces directement des informations de localisation à cause du peu d'occurrences que l'on a pour certaines espèces et du biais d'échantillonnage (peu d’occurences ou pas à certaines localisations)

Ce qui est fait en écologie c'est de partir de la base d'une représentation d'un environnement. Typiquement un vecteur composé de ( température moyenne, précipitation ect ) ainsi que d'autres variables comme le type de sol, la couverture terrestre, la distance de l'eau.

L'originalité de GeoLifeClef est d'utiliser des images pour représenter directement ces variables environnementales.

Ces images sont un modèle en k dimensions, chaque patch représentant une valeur de variable de l'environnement ( température moyenne annuelle, max température du mois le plus chaud , précipitations annuel, l'altitude , la proximité à de l'eau fraîche ainsi que la capacité d'eau disponible.

On se ramène à un problème de machine learning de "multi channel image classification task".

Chaque image représente une channel d'une variable d'environnement dans un carré centré dans la localisation.

DATASET:

Fichier .csv

Chaque participants obtient des données d'apprentissages et de test des occurrences des espèces géolocalisées de variables ponctuelles d'environnement.

Les participants ont reçu une séries d'apprentissage et une série de test d'occurrences géolocalisées d'espèces. Les deux étaient d'abord composés d'un fichier.csv contenant les coordonnées spatiales des occurrences, les valeurs ponctuelles des variables environnementales à l'endroit de l'occurrence et, pour le tableau de l'apprentissage, le nom de l'espèce et son identificateur. Deuxièmement, chaque ligne du tableau (train et test) renvoie à une image à 33 canaux contenant le tenseur environnemental extrait à cet endroit.

Les occurrences ont été extraites de GBIF ( Global Biodiversity Information Facility Platforms.  On doit essayer d’avoir les données de géolocalisation les plus précises possible, mais celle du GBIF sont parfois absentes ou imprécises dues à la nécessité de protéger certaines espèces.

Du coup on a pris de ces sources :

1. Carnet en ligne from Tela Botanica.

2. Cartographie des Leguminosae (Fabaceae) en France from Tela Botanica.

3. Naturgucker dataset.

4. iNaturalist Research-grade Observations.

Et on a supprimé les données imprécises du GBIF.  Seules les données du territoires français ont été gardé car c’est un territoire bien connu et avec une biodiversité riche et variée.

L’ensemble du dataset est composé de 291 392 occurrences. Avec plus de 3336 espèces différentes.

Le label qu’on doit prédire dans le test sont le (field species glc id ) . Les 3336 espèces différentes sont associés à leurs noms scientifique qui sont donnés dans le champ espèce\_retenue\_bdtfx.

Il y a 75 668 différentes localisation avec un maximum de 527 points dans une géolocalisation.

ENVIRONNEMENTAL DATA

Chaque occurrence pour un environnement est caractérisé par 33 images de 64x64 pixels. Ces variables d’environnement ont été construite à partir de sources diverses.  On a donc pour chaque occurrences des tenseurs de taille 64\*64\*33 pixels.

TRAIN AND TEST:

Sur les 291 392 occurrences qui ont été séparé de façon aléatoire en un ensemble d’apprentissage ( 218 543 ) et un ensemble de test ( 72 849 ).

Contraintes :

Pour chaque espèce présentes dans le test il y a au moins une observation dans l’apprentissage.

Une observation d’ espèce dans le test est distante d’au moins 100 m des toutes observations de cette espèce dans l’ensemble d'entraînement pour éviter que ce soit trop facile à prédire.

DESCRIPTION DE LA TÂCHE

Pour chaque occurrences du test les participants doivent classer par top 100 sans ex aequo.  L’évaluation métrique qu’on utilise s’appelle la MMR

Avec Q le nombre total d’occurrence xq dans l’ensemble de test et rankq est le rang correct de l’emplacement de l’espèce y(xq) dans la liste des espèces évalué par la méthode pour l’occurrence xq.

La MMR est une mesure statistique permettant d’évaluer les processus qui renvoie une liste de valeurs possibles en réponse à une requête et ordonnés par la plus grande probabilité de vraisemblance.

PARTICIPANTS ET MÉTHODES.

Il y a 22 groupes de recherches qui se sont inscrit et seulement 3 team ont réussi à soumettre un code fonctionnelle à temps.

FLO a développé 4 modèles de prédictions:

(i)

Un réseau de neurone convolutionnel entraîné sur les tenseurs environnemental. (FLO3)

Le CNN a été implémenté selon une architecture propre.

Il a la particularité d’avoir traité les caractéristiques de la couverture du sol indépendamment des variables quantitatives des premières couches ( de neurones ).

L’activation des deux variables ont été fusionné dans des couches plus profonde.

(ii) Un réseau de neurone (FLO 2) entrainé sur les occurrences d’espèces tombant à la location la plus proche, et deux autres modèles uniquement basé sur les occurrences de location des espèces.

(iii) Un classifieur pour la localisation la plus proche (FLO 1)

(iv) Une random forest ( plusieurs arbres de décisions ) modéliser/ adapter aux coordonnées spatiales. (FLO 4).

Les autres exécutions correspondent aux fusions des modèles de base. (FLO 5,FLO 6,FLO 7,FLO 8)

Ou bien classé selon la méthode de Borda. (FLO 9,FLO 10).

ST team, Germany, 16 runs, [14]:

L’équipe ST a experimenté deux types de modèles. Les réseaux de neurones convolutionnel sur les tenseurs environnementaux avec différentes transformation des données comme la rotation et le flip d’image ( ? ) (ST 1, ST 3, ST 11, ST 14, ST 15, ST 18, ST 19)

Et les arbres boostés (XGBoost) sur les vecteurs environnementaux concaténé avec leur position spatiale.(ST 6, ST 9, ST 10, ST 12, ST 13, ST 16, ST 17).

Ils ont aussi proposé un classifieur type plus proche voisin basé sur les occurences de variables d’environnement. (ST 5)

et un modèle de cluster deux espèces (? ). Ou les groupes d’espèces sont constitué par similarité dans les variables d’environnement ou elles apparaissent. (ST 17,ST 8).

Pour raison d’analyse ST2 correspond a un predicteur aléatoire et ST7 correspond à prédicteur constant qui retourne toujours les 100 plus fréquentes espèces classé par ordre décroissant de leur fréquence dans l’ensemble de test.

SSN, India, 4 runs, [10]:

SSN a essayé d’apprendre un CNN-LSTM ( ? ) modèle hybride basé sur une ResNext Architecture (? ) étendu avec une couche LSTM avec pour objectif de prédire des catégories de plantes à 5 différents niveaux de taxonomy ( classe, ordre, famille, gêne et famille de l’espèce). Les 4 executions sont dérivées de ce modèle.

RESULTATS

Les réseaux de neurones convolutionnel surclasse les boosted trees, bien que les BT soient utilisés souvent pour les problèmes environnementaux de ce type.  D’après l’évaluation des différents modèles proposées ils ont montrés que les CNN était plus performant.  Le meilleur BT a atteint une MMR de 0.035 tandis que le meilleur CNN est à 0.043.  Néanmoins les modèles CNN entraîné par les team ST et SSN n’ont pas obtenu des bons scores du tout. Voir pire que le classifieur constant basé sur les fréquences des espèces dans tout le dataset.

Ce qui pourrait être du à un décalage des ID des espèces dans le traitement de données fait par la team SSN.  Pour la team ST cela est peut être du à un échec du modèle d’entrainement.  Cela montre la difficulté de mettre en place un réseau de neurones profond en application sur un nouveau problème sans avoir de références et d’articles sur ce sujet précis.  Les méthodes classique de classification d’image tel que VGG , DenseNet, ResNEXT et LSTM n’ont pas était performant. Il y a eu une grosse différence entre les méthodes CNN conçues spécialement pour ce problème tel que l’a fait FLO team , comparé à ST. Ce qui souligne l’importance du choix de l’architecture.

Les modèles purement spatiales ne sont pas si mauvais

La random forest de FLO team a atteint une MMR de 0.0329 ce qui est assez proche de ce qu’a fait la team ST avec des Boosted Tree.  Usuellement on utilise pas de modèle purement spatiale pour la prédiction d’une distribution d’espèce cela est due à l’hétérogénéité de la densité des observations à travers les différentes régions. En effet la distribution spatiale des distributions des espèces observées est souvent plus corrélé à la préférence de la région de l’observateur qui effectue les mesures plutôt qu’à l’abondance des espèces dans une région.  Cependant le rôle de GeoLifeClef est de prédire la plus grande vraisemblance d’espèce observés étant donné la réelle présence d’une plante.  Donc ce problème d’hétérogénéité devrait posé moins problème et moins de biais qu’une étude écologique classique.   Les CNN capturent déjà les informations spatiales.  La meilleure execution, (FLO 6 ) est le résultat d’une combinaison entre le meilleur modèle environnementale ( CNN FLO 3 ) et le meilleur modèle spatial ( Random Forest FLO 4 ).  Cependant on peut noter que l’amélioration de l’execution de FLO\_6 est très petit par rapport à FLO 3, + 0.0005 ce qui, statistiquement parlant, n’est pas significatif comme devancement.  On a donc déjà un modèle CNN qui capture déjà toutes les informations du modèle spatial ( en tout cas c’est tout comme ). De plus le CNN utilise tout le tenseur environnemental comme entrée ce qui est meilleur que des méthodes XGB Boost qui utilisent seulement la moyenne de chaque matrice d’environnement comme entrée. C’est pour cela que CNN capture plus d’info.

L’apprentissage de communauté d’espèce a du potentiel.

Les points communs entre espèces peuvent aider à mieux prédire la liste d’espèces la plus plausible dans un environnement, et c’est bien plus efficace que de prédire simplement une liste d’espèce par global abondance.( ST 7). En effet si on essaie de prédire par rapport à la similarité entre les espèces comme en faisant une méthode des plus proches voisins (FLO 1 ) ou sur l’environnement (ST 5 ) on peut s’attendre à de bons résultats. Les deuxième méthodes autorisent les interraction entre les abondances des espèces soit en construisant une prédiction sur un groupe d’espèce qui ont les même préférences environnementales soit apprendre l’association entre les espèces qui occurrent dans un espace assez proche ( FLO 2 ). On note tout de même que ces méthodes sont limités et qu’il n’y a pas eu de résultats prodigieux dans leurs executions.

Une bonne marge de progrès est à faire mais on a déjà des résultats assez promettant. Les meilleurs scores MRR paraissent très faible à première vue mais en relativisant par rapport à la nature de la tâche et du fait que beaucoup d’espèces ( dizaine ou centaines ) vivent a peu près au même endroit empêche le fait qu’on ait de gros scores MRR. Un gros score MRR est impossible pour cette tâche mais c’est utile pour comparer les méthodes entre elles. Le score parfait en MMR est de 0.56 ( FLO 3 qui a eu 0.043 est encore loin ) mais c’est bien mieux qu’une distribution aléatoire à priori basé sur MMR. On a les bonnes espèces dans le top 10 dans 25% des cas, et dans le top 100 dans 49% des cas. ( avec 3336 espèces dans l’ensemble d’apprentissage).  Cela veut dire que ce n’est pas si mauvais pour prédire un ensemble d’espèce qui vont être observés selon une location.

Perfect Score pour GeoLifeClef MMR

ANALYSE COMPLÉMENTAIRE

Performance d’un modèle sur une hétérogénéité spatiale. On a calculé des MRR restreint aux occurrences qui tombent dans un carré d’espace de taille 10x10 km sur tout le territoire français. Paris est l’endroit ou l’on prédit le mieux les espèces après on retrouve la région méditerranéenne et les Alpes.  L’analyse complémentaire montre que les régions ont l’on a eu beaucoup d’échantillon ont une MMR plus stable mais pas plus haut en moyenne. Donc pour améliorer des modèles de prédictions on devrait essayer de trouver les raisons de la variance de performances des différentes régions.

Les espèces rares ne sont pas imprévisible. Pour chaque espèce et méthode ils ont calculé le MMR sur les occurrences des espèces dans cet ensemble. Ils ont ordonnés les espèces par ordre décroissant d’occurences comptés dans l’ensemble de test et comparé aux performances de chaque méthodes selon le gradient de l’espèce la plus commune à l’espèce la plus rare.

importantly for rare species, as there are very few occurrences. Thus, we operate a smoothing along the scarcity gradient. For each species we took the median of the MRR over the 40 species of closest rank on this scarcity gradient. Figures 5 and 6 show the result for FLO 3 (environmental CNN), ST 16 (XGBoost), FLO 4 (spatial Random Forest) and ST 7 (Global frequency of species). One can see that ST 7 early cancels along the scarcity gradient. This is because more than 50% of the species over which the median is calculated have a null MRR, which correctly represents the tendancy we want to observe. First, it seems that non-common species have marked spatial preferences because FLO 4 is much better when getting scarcer than ST 7. Second, the progression of predictions of FLO 3 and ST 16 compared to FLO 4 for rare species (in the long tail) suggests that those species mainly have marked environmental preferences that is not easy to capture with a spatial model which doesn’t have access to this information. The CNN is very good at predicting non-common species, which may be a bit surprising as (i) its predictions should be smooth in space according to the width of some environmental images (64x64km for climatic and pedological variables) and the chosen architecture and (ii) rare species often have a restricted niche.

CONCLUSION.

Les modèles CNN s’avèrent être plus performant que les Boosted Trees, néanmoins ils restent difficile à mettre en place (FLO: architecture customisée). Les associations de méthodes basé sur l’environnement et sur les espèces ont montré de bons résultats.  L’intégration de variables d’environnement (? ) semble être bénéfique pour la prédiction d’espèces non commune.  La tâche pour trouver l’exact localisation d’espèce sachant une localisation est difficile étant donné le nombre d’espèces vivant très proches les unes des autres.

II. LES ÉQUIPES

a.) 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).

b.) ST TEAM

Méthode d’extraction des données et traitement des données.

préprocessing :

Les images ont été donné sous deux formats, un format image et un fichier .csv contenant les valeurs singulières

Ils ont donc voulu extraire les valeurs singulières pour chacune des 33 images des occurences, mais ils ont remarqué que certaines valeurs singulière étaient manquantes, ce qui rendait l’image inutilisable.   
Ils ont donc décidé de construire leurs propres valeurs singulières pour les images mais ils ont également vu que ce n’était pas possible de reproduire.   
Ils ont calculé les valeurs singulière en faisant une moyenne des 4 pixels centraux ( peut être trop peu ? )

En raison de ces différences ils concluent que c’est peut être à cause de cela qu’il y a une différence entre le test et l’apprentissage.

Ensuite ils ont compté les valeurs singulières les plus probables pour chaque classe ( les classes représentant les milieux environnementaux ? ) et les ont compté pour pouvoir avoir une identification pour chaque classe, identification qui sera la valeur singulière qui occure le plus dans la classe.

ST team suppose que les images des 33 channels n’ont pas tous la même influence sur la classification de l’environnement, ils ont donc essayé de déterminer quel sont les images avec le plus de “poids”.

Pour chaque classe d’environnement ils ont créé une map de chaleur qui contiennent les valeurs moyenne pour chaque image de cette classe.

Ils ont utilisé 10% de test et 90% en apprentissage.

Les différents Modèles qu’ils ont soumis :

* Un modèle aléatoire : Résultat 0.0016
* Modèle probabiliste: Compter les occurrences d’espèces et mettre la plus fréquente en tête de liste pour chaque classe d’environnement : Résultat 0.0134
* Modèle vectoriel : Ils ont utilisé deux vecteurs pour chaque occurence, une occurrence du test et une occurrence de l’apprentissage et ils ont trié selon la plus petite différence: Résultat 0.0271
* Convolutional Neural Network :   
  Ils ont implémenté plusieurs type de CNN : VGG likes model (single , multi )   
   DenseNet

Meilleur résultat : 0.0153

* XGBoost: Meilleur résultat : 0.0942

Leurs codes sources <https://github.com/stefantaubert/lifeclef-geo-2018>

Ils ont fait des moins bons scores que la team FLO ils pensent que c’est surtout un problème d’implémentation du CNN. Ce qui confirme l’importance de bien implémenter le modèle.

Les SOURCES en plus de ce qui est fourni par LifeClef

CONVOLUTIONAL NEURAL NETWORK

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6210443> Multi-Channel Convolutional Neural Networks Architecture Feeding for Effective EEG Mental Tasks Classification

<https://fr.wikipedia.org/wiki/R%C3%A9seau_neuronal_convolutif>

<http://deeplizard.com/learn/video/YRhxdVk_sIs> Blog sur CNN.

* Quelques liens de vidéos de vulgarisation sur les convolutional neural network :
* <https://www.youtube.com/watch?v=YRhxdVk_sIs>
* <https://www.youtube.com/watch?v=zG_5OtgxfAg> science4all

Cnn pour les modèles de distribution d’espèces :

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)

IMAGE CLASSIFICATIONS

<https://reader.elsevier.com/reader/sd/pii/S1877705817341486?token=BDAD98AC2B0849CC74E37E745FDC283DE8715B5E4E203FC9202DF0C334CC3DFB5E1A55F8E3E9B3ADD77DDE7A396E061C> Supervised multichannel image classification algorithm using

hierarchical histogram representation

<https://mirc.uzleuven.be/download/public/MIC/publications/1635/ESATtech9903.pdf>

Automated model-based tissue classification

of MR images of the brain

BOOSTED TREES:

<https://statweb.stanford.edu/~jhf/ftp/trebst.pdf>   
https://web.archive.org/web/20091110212529/http://www-stat.stanford.edu/~tibs/ElemStatLearn/