UFR 6
Université Paul Valéry, Montpellier III

Projet TER

Rendu 7 - Interprétation des résultats

Alvin VEDEL, Florian DUBOIS, Matis BREILLAD

14 Avril 2024

Résumé

Ce rapport a pour but de présenter les résultats obtenus lors de notre projet TER de distribution d'espèces. Il comporte une description du contexte d'évaluation du modèle suivi d'une liste exhaustive des modèles proposés et de leurs résultats.

Table des matières

Résumé **i**

Liste des figures **ii**

Introduction **1**

1 Présentation des résultats **2**
1.1 Méthode d'évaluation 3
1.1.1 Le concours Kaggle 3
1.1.2 Métrique d'évaluation 3

2 Les scores obtenus **4**
2.1 Les modèles de Machine Learning 5
2.1.1 Le KNN 5
2.1.2 Le Random Forest 5
2.2 Les modèles de Deep Learning 5
2.2.1 Réseau Dense 6
2.2.2 CNN bioclimatiques 6
2.2.3 CNN Satellites 6
2.2.4 Modèle ensembliste 6
2.3 Mélanger les 2 7

Conclusion **7**

Table des figures

Introduction

Pour rappel du contexte, notre tâche principale consiste à prédire la présence ou l'absence d'une liste complète d'espèces à des endroits donnés. Pour chacun de ces points nous possédons des données de test qui se contruisent de la même façon que nos données d'entrainement à la différence que nous n'avons pas la connaissance sur les espèces prédites. Il s'agit donc d'une tâche de classification multi-classes et multi-labels relativement complexe. Nous présenterons dans ce rapport les conditions d'évaluation de la tâche ainsi que les résultats obtenus par nos modèles prédictifs.

Chapitre 1

Présentation des résultats

Sommaire

1.1 Méthode d'évaluation 3
1.1.1 Le concours Kaggle 3
1.1.2 Métrique d'évaluation 3

1.1 Méthode d'évaluation

1.1.1 Le concours Kaggle

Le projet s'inscrivant dans le cadre d'un concours Kaggle "GeoLife Clef 2024", un jeu de données de validation et de test non accessible sont disponibles et permettent d'évaluer nos modèles. (Le résultat affiché est celui du jeu de validation, à la fin de la compétition c'est le score sur le jeu de test qui déterminera le gagnant). Etant limités à 3 soumissions, nous avons du optimiser notre temps et nous organiser pour maximiser les tests.

1.1.2 Métrique d'évaluation

La métrique utilisée par le Kaggle est le micro F1-score :

$$F_{1,\text{ micro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + (\text{FP}_i + \text{FN}_i) \times 1/2}$$

C'est une mesure harmonique entre le rappel et la précision qui a tendance à moins pénaliser l'excès d'espèces prédites que l'absence de certaines.

Chapitre 2

Les scores obtenus

Sommaire

2.1 Les modèles de Machine Learning 5

2.1.1 Le KNN 5

2.1.2 Le Random Forest 5

2.2 Les modèles de Deep Learning 5

2.2.1 Réseau Dense 6

2.2.2 CNN bioclimatiques 6

2.2.3 CNN Satellites 6

2.2.4 Modèle ensembliste 6

2.3 Mélanger les 2 7

2.1 Les modèles de Machine Learning

2.1.1 Le KNN

Le modèle de KNN qui ne se base que sur les coordonnées géographiques a été optimisé en sélectionnant le nombre de voisins et le seuil des n présences parmi les k voisins. Les meilleurs résultats ont été obtenus avec un seuil de 0.2 et 24 voisins pour un F1-score de 0.25213. Le détail des submits effectués sont disponibles ci-dessous :

	Nombre de voisins											
Seuil	15	20	21	22	24	25	30	40	50	60	80	100
0.2	0.247	0.247	0.250	0.250	**0.252**	0.247	0.245	0.243	0.241	0.236	0.232	0.225
0.15	X	X	X	X	0.249	X	X	X	X	X	X	X
0.25	X	0.233	X	X	X	X	X	X	X	X	X	X

TABLE 2.1 – Tableau de contingence

2.1.2 Le Random Forest

Les modèles de Random Forest se concentrent sur les données tabulaires soit un ensemble de 50 variables quantitatives. Plusieurs paramètres ont été optimisés à la manière du KNN avec le seuil et le nombre d'arbres. Il est à noté que la compilation d'une prédiction du modèle nécessite de réaliser autant d'arbres que d'espèces à prédire ce qui signifie un temps d'exécution très long et peu de soumissions réalisées sur le challenge Kaggle 2024, davantage avait été proposées en 2023. Cela avait contribué à trouver un bon seuil.

	Nombre d'arbres			
Seuil	5	20	40	100
0.2	X	0.23492	0.24589	**0.25165**
0.15	0.16229	X	X	X

TABLE 2.2 – Tableau de contingence

2.2 Les modèles de Deep Learning

Les modèles qui utilisent les images sont des réseaux convolutifs, ils s'inspirent d'un modèle qui généralisent bien : ResNet18. Cela comprend une succession de convolutions avec des kernels de 3x3 avec de la batch normalisation, des strides de 1 (parfois 2). Différentes fonctions d'activations ont été testées sur le dernier layer Dense du réseau mais des ReLU ont été appliquées partout ailleurs.

2.2.1 Réseau Dense

Des architectures sur les données tabulaires ont été proposées, de nombreuses et très différentes avec des nombres de paramètres et des fonctions variables. Une idée restait dominante : proposer une architecture profonde, reputées pour leurs capacitées de généralisation. Cependant les résultats restent médiocres avec un score maximal de 0.19371 mais la plupart des essais ont montré que les réseaux n'apprenent rien et retournent les espèces les plus présentes dans le jeu d'entraînement avoisinant un score de 0.16.

2.2.2 CNN bioclimatiques

Pour le réseau convolutif sur les données bioclimatiques, des résultats satsifaisants ont été obtenus sans layer après les convolutions (0.25247) mais la nécessité d'extraire des features nous a poussé à ajouter une couche supplémentaires. Cela nous a permis d'améliorer les résultats en montant à 0.25678 avec une tangente hyperbolique, d'autres tests avec une ELU ont été moins concluants : 0.23445.

Procéder à une limite du nombre d'époques à été nécessaire pour ne pas overfiter et avoir de bons résultats sur Kaggle (early stopping) avec 20 époques. En montant à 50, 100 ou 200 les performances ont tendance à diminuer.

Le nombre d'espèces à retourner a été estimé par régression xgboost, prédictions auxquelles on a ajouté +17 espèces (valeur maximisant les scores). Cela prend en compte la métrique du F1-score plus laxiste sur les faux positifs.

2.2.3 CNN Satellites

Le même schéma prédictif a été appliqué aux données des séries temporelles landsat desquelles on a pu extraire plus d'informations. Avec la tangente hyperbolique qui présente largement les meilleurs résultats, on obtient un score de 0.27290.

2.2.4 Modèle ensembliste

A partir des 2 fichiers de csv extraits des CNN, une méthode prédictive consistant à regrouper les espèces a été mise en place. Dans un premier temps une fusion naïve des 2 fichiers qui a permis d'augmenter légèrement le meilleur score, à ce moment là, des réseaux convolutifs et d'obtenir 0.25694. En ne prenant pas les K espèces retournées par "surveyId" mais les 2K/3 premières espèces, le score est monté à 0.27268. Le modèle avec sur les landsats vint ensuite écraser ce score et une méthode plus raffinée consistant à prendre les espèces communes aux deux prédictions puis d'ajouter les espèces les plus probables, préférentiellement de landsat, à la liste commune jusqu'à obtenir les K espèces attendues par la régression + 17 espèces. Ce processus ayant permis de faire grimper le score à 0.29124.

2.3 Mélanger les 2

Le dernier objectif visant à extraire le maximum d'informations des données de Présence Absence consiste à utiliser les features extraites des séries temporelles représentées sous forme d'images et à les combiner à l'aide d'un XGBoost ou un Random Forest aux données tabulaires. Le modèle ainsi proposé serait le plus apte à bien prédire de par leurs bonnes capacité de généralisation (petite dimension VC en comparaison des réseaux de neurones) et la quantité de données maximale. Il est attendu que ce modèle présente les meilleurs résultats mais on peut déjà supposer qu'un tel modèle uniquement sur les features serait plus efficace qu'une prédiction par le réseau de neurones. C'est une solution courante dans ce type de tâches de classification. Les résultats n'ont pas encore été obtenus cependant pour affirmer cette hypothèse.

Conclusion

Les résultats montrés précédement sont assez faibles. Il est difficile de faire confiance à un modèle qui n'obtient qu'un F1-score d'au mieux 0.291 et cela sur un jeu de validation où rien ne garantit d'aussi bons résultats surle jeu de test. Cependant nos modèles nous ont permis de nous hisserà la 3ème place du challenge auquel participent une soixantaine de personnes. Les données utilisées restent partielles et cette tâche nécessiterait une approche plus proche des données en réfléchissant davantage à la nature des données. Les données de Présences Seules sont une source d'information colossale intacte jusqu'ici qui permettrait des améliorations notables, bien qu'elles présentent des biais non négligeables.