Projet IML

Sujet

Vous avez été fraîchement recruté dans la start-up Carotte 4.0[1] (félicitations), dont l'objectif est l'étude de l'évolution de l'occupation des sols en milieu urbain et rural, pour une meilleure gestion et planification des ressources agricoles dans le respect de la biodiversité.
À ce titre, votre patron est en cours de négociation avec l'ESA pour obtenir un accès libre et illimité sur les données hyperspectrales de la future mission EnMAP (https://www.enmap.org/). Pour ce faire, il a besoin de prouver, lors de la prochaine réunion avec les décideurs de l'ESA et les investisseurs de la start-up (prévue le 19 mai), que la start-up est capable de proposer des solutions innovantes et disruptives pour le traitement d'images hyperspectrales. Et c'est sur vous que c'est retombé.

Votre mission, si vous l'acceptez (de toute façon, vous n'avez pas le choix, la survie de la boîte dépend de vous), est de développer un pipeline de classification d'images hyperspectrales. Vous avez à votre disposition deux cubes de donnés hyperspectrales acquis par le capteur AVIRIS, dont les caractéristiques sont proches du futur capteur embarqué par EnMAP[2] : 224 bandes spectrales entre 0.4 µm (visible/bleu) et 2.5 µm (proche infrarouge), avec une largeur de bande de 10 nm.

La première image a été acquise au dessus du site Indiana Pines, dans l'Indiana (USA), et comprend 145x145 pixels, avec une résolution au sol de l'ordre de 20 m/pixel. Les bandes d'absorption de l'eau ont été retirées des 224 bandes, résultant en un total de 200 bandes spectrales. Une vérité terrain est également disponible, comprenant 16 classes (majoritairement des cultures) malheureusement assez mal équilibrées entres elles.

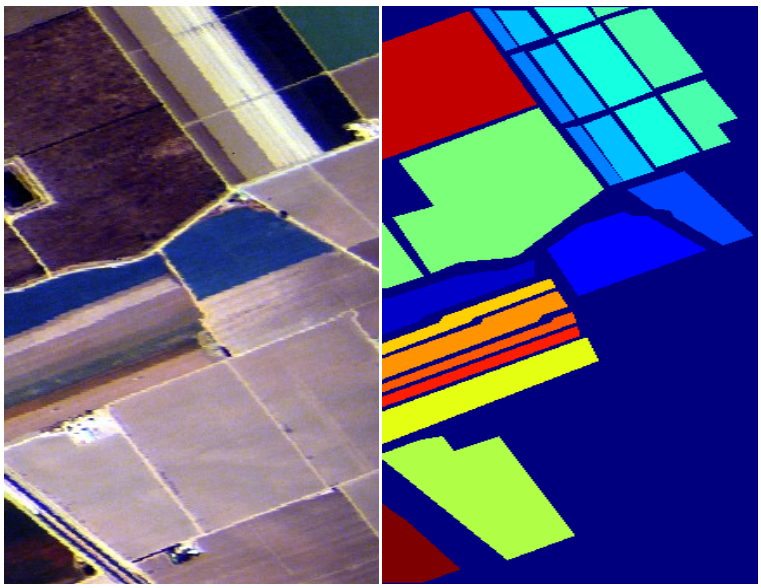
Gauche : composition RGB de l'image hyperspectrale d'Indiana Pines (les bandes ayant servi à la composition RGB sont R=30, G=15, B=2)
Droite : vérité terrain correspondante, les 16 classes sont étalées entre le bleu (1) et rouge (16) (colormap 'jet'). Le fond bleu foncé correspond à la classe 0 (absence d'information).

1 https://www.youtube.com/watch?v=_VB_b4dSmzE&ab_channel=BirahimGUEYE
2 En vrai, pas du tout, mais ça facilite la narration du sujet

#	Class	Samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Répartition des différentes classes de la vérité terrain d'Indiana Pines et nombre d'échantillons par classe.

La deuxième image a été acquise au dessus de la vallée de Salinas, en Californie. Cette image comprend 512x217 pixels, avec une résolution au sol de l'ordre de 3.7 m/pixel. L'avion embarquant le capteur volait en effet plus bas que lors de l'acquisition au dessus d'Indiana Pines, et moins de bandes ont été corrompues par l'absorption de l'eau, résultant en 204 bandes spectrales au total (comparativement aux 200 bandes d'Indiana Pines). La vérité terrain comprend également 16 classes, cette fois-ci mieux réparties.



Gauche : composition RGB de l'image hyperspectrale de Salinas (les bandes ayant servi à la composition RGB sont R=30, G=15, B=2)
Droite : vérité terrain correspondante, les 16 classes sont étalées entre le bleu (1) et rouge (16) (colormap 'jet'). Le fond bleu foncé correspond à la classe 0 (absence d'information).

#	Class	Samples
1	Brocoli_green_weeds_1	2009
2	Brocoli_green_weeds_2	3726
3	Fallow	1976
4	Fallow_rough_plow	1394
5	Fallow_smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes_untrained	11271
9	Soil_vinyard_develop	6203
10	Corn_senesced_green_weeds	3278
11	Lettuce_romaine_4wk	1068
12	Lettuce_romaine_5wk	1927
13	Lettuce_romaine_6wk	916
14	Lettuce_romaine_7wk	1070
15	Vinyard_untrained	7268
16	Vinyard_vertical_trellis	1807

Répartition des différentes classes de la vérité terrain de Salinas et nombre d'échantillons par classe.

Pour développer votre pipeline de classification, vous avez le droit d'utiliser toutes les ressources disponibles dans scikit-learn, mais pas de deep learning (toutes les boîtes en concurrence avec vous pour remporter le contrat de l'ESA vont de toute façon se tourner vers le deep learning, et votre patron a bien insisté : vous devez être disruptif).
Mais attention, votre but est certes de trouver la méthode permettant d'obtenir les meilleures performances de classification, mais il s'agira également d'avoir l'air sérieux le jour de votre présentation le 19 mai : ainsi, il vous faudra montrer et argumenter pourquoi la solution que vous avez retenue est la meilleure, ce qui veut dire benchmarker un minimum les différentes solutions possibles, et les comparer entre elles.

Bon courage, l'avenir de Carotte 4.0 est entre vos mains !