

Clasificación de cultivos con imágenes satelitales

Diciembre de 2020

BATTOCCHIA, Matías; CALDERÓN, Leonardo; ECHAZÚ, Gustavo;
MOYANO, Carlos; MUNAFÓ, M. Victoria

Introducción

El objetivo planteado para la competencia fue la clasificación de cultivos mediante imágenes satelitales y métodos de clasificación supervisada en el departamento de General López, Santa Fe. El mismo surge de la necesidad de facilitar las tareas de relevamiento de los cultivos agrícolas. Estas, normalmente, son realizadas a campo por peritos técnicos que geolocalizan el tipo de cultivo en una determinada campaña. Dada la pandemia de COVID-19, esta tarea no pudo ser realizada durante el corriente año.

Establecimiento de criterios

Selección de la plataforma satelital

Se definió la utilización de los satélites que componen la misión Sentinel-2A y 2B del programa Copérnico llevado adelante por la Agencia Espacial Europea (ESA). Para llegar a esta conclusión se ponderó la alta resolución espacial (10 y 20 m en las bandas de interés) y la alta resolución temporal, ya que permite adquirir imágenes de un mismo sitio cada 5 días. De esta forma es posible conseguir una gran serie temporal que sea sensible a los cambios en la vegetación y su disponibilidad no se vea mayormente afectada por la cobertura nubosa[1].

Selección de las bandas electromagnéticas de interés

De acuerdo a su resolución espectral , se definió[4] la utilización de: las bandas azul, verde, roja, vegetation red edge, infrarrojos cercanos, infrarrojos de onda corta y de los índices calculados, por sus siglas en inglés, NDVI (índice de vegetación de diferencia normalizada), NDBI (índice de diferencia normalizada edificada) y NDWI (índice diferencial normalizado de agua).

$$NDVI = (\rho_{NIR} - \rho_{red}) / (\rho_{NIR} + \rho_{red}) \quad (1)$$

donde ρ_{NIR} es la reflectancia en la banda del infrarrojo cercano y ρ_{red} , la reflectancia en la banda roja[3].

$$NDWI = (\rho_{green} - \rho_{SWIR}) / (\rho_{green} + \rho_{SWIR}) \quad (2)$$

donde ρ_{SWIR} es la reflectancia en el infrarrojo de onda corta y ρ_{green} , es la reflectancia en la banda verde[5].

$$NDBI = (\rho_{SWIR} - \rho_{NIR}) / (\rho_{SWIR} + \rho_{NIR}) \quad (3)$$

donde ρ_{NIR} es la reflectancia en la banda del infrarrojo cercano y ρ_{SWIR} , la reflectancia en el infrarrojo de onda corta[6].

Plataforma de adquisición

Los códigos de identificación de imágenes Sentinel (tile ID) que abarca el departamento de General López son 20HNH, 20HPH, 20HNJ y 20HPJ. Con el fin de extraer la información satelital correspondiente a las verdades de campo y los sitios para los cuales se esperaba la predicción, se recurrió a la API Sentinel-Hub en su versión gratuita.

Se extrajeron escenas de 10×10 píxeles centradas en el punto establecido como verdad de campo.

Porcentaje de cobertura nubosa

Se definió un umbral máximo de cobertura nubosa tolerada en 20 %, debido al efecto de dispersión que producen especialmente para la banda SWIR[2]. Es decir que aquellas con cobertura nubosa por encima del umbral fueron descartadas automáticamente, mientras que para las que estaban por debajo del umbral se aplicó interpolación de los píxeles marcados como nubosos por la máscara de nubosidad.

Periodo de adquisición de las imágenes

El período de adquisición de las imágenes satelitales fue establecido desde el 1ro de noviembre hasta 15 de abril para ambas campañas. De acuerdo a la Bolsa de Cereales Argentina y su delimitación de las "Zonas PAS" (Panorama Agrícola Semanal) el departamento de General López se encuentra en la Zona VII denominada Núcleo Sur. En concordancia con los datos del conjunto de entramiento, los cultivos de maíz y soja son mayoritarios en esta zona y ambos son de desarrollo estival. Con lo cual el período de descarga de las imágenes satelitales fue delimitado en función de los meses de máximo desarrollo de estos cultivos y sus correspondientes períodos críticos (establecidos por la Oficina de Riesgo Agropecuario).

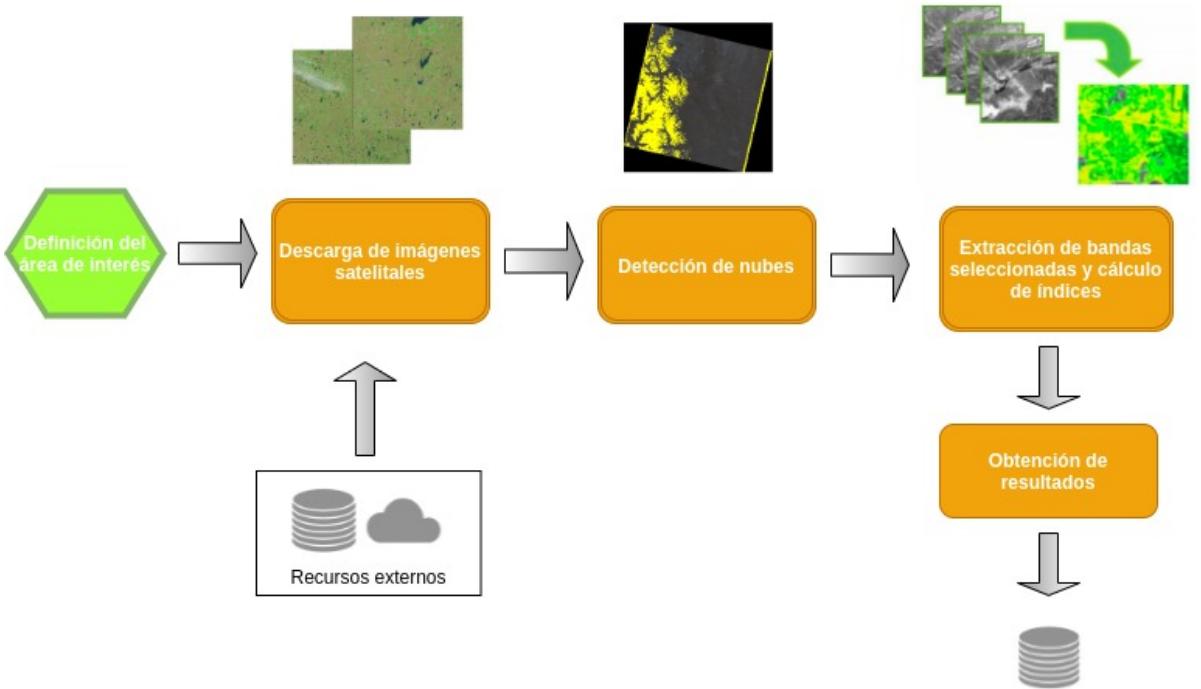


Figura 1: Diagrama de la adquisición de las imágenes mediante la librería eo-learn.

Consideraciones sobre las clases: limpieza del dataset

Durante el análisis del conjunto de entrenamiento se observó que algunas de las clases estaban subrepresentadas en relación a las mayoritarias. En el caso del trigo (T), sólo constaban dos puntos y dado que este cultivo, en esta zona de Santa Fe, se cosecha durante el mes de noviembre (al inicio del período de adquisición) se procedió a la eliminación de esta clase. Lo mismo se hizo con la clase S/M que sólo tenía una observación y desde la organización informaron que fue un error de peritaje.

Con el objeto de evitar datos erróneos en los modelos de entrenamiento se realizó la comprobación y corrección de la ubicación de los puntos (verdades de campo). Para esta tarea se utilizó el Sistema de Información Geográfica QGIS. En primer lugar se visualizaron los puntos correspondientes a los conjuntos de entrenamiento y testeо utilizando un mapa base de Google Satellite de alta resolución espacial. Luego, se reubicaron los puntos ubicados en zonas de alta variabilidad espacial, que pudieran perturbar el correcto funcionamiento del modelo. Tal es el caso de un punto ubicado sobre el margen del campo, próximo

a la calle.



Figura 2: (a) Punto original. (b) Punto curado.

Por otra parte, se seleccionaron las clases con menos de 12 observaciones y se agregaron puntos próximos a los existentes (dentro de los mismos campos en lo posible), con el fin de obtener un mínimo de 12 observaciones por clase.



Figura 3: Puntos adicionales de entrenamiento de clases minoritarias.

Preprocesamiento de datos

De acuerdo a los criterios delimitados con anterioridad, cada punto descargado consta de 11 capas, 8 bandas puras de Sentinel-2 y 3 capas correspondientes a los índices NDVI, NDBI y NDWI, de 10×10 píxeles, con una frecuencia temporal de 5 días para el período seleccionado (noviembre-abril).

Por lo tanto, las dimensiones del tensor para una muestra del dataset son: (cantidad de imágenes disponibles, ancho [px], alto [px], capas) = (34, 10, 10, 11). Debido al filtro de nubosidad, no todos los puntos cuentan con igual cantidad de imágenes disponibles, variando entre 10 y 30 por campaña.

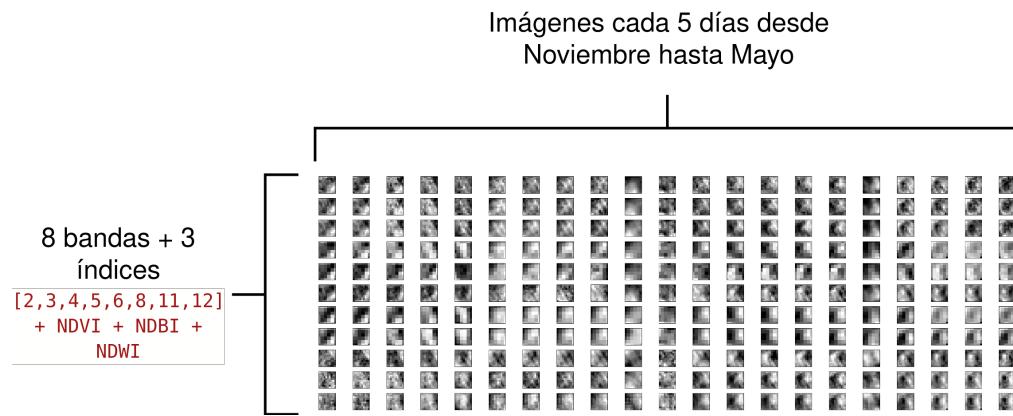


Figura 4: Imágenes correspondientes a un punto de entrenamiento.

Para cada imagen se consideró el valor medio de reflectancia para la banda correspondiente, reduciendo cada tensor a otro de dimensión (cantidad de imágenes disponibles, capas). Con el fin de obtener una serie temporal continua, se realizó una interpolación temporal con resolución diaria (Akima spline de grado 3) y luego fue resampleada (tomando el promedio) a una frecuencia de 5 días.

Para finalizar el preprocesado, se concatenaron las 11 series temporales (una por banda/índice) en un vector de 374 valores (utils.py).

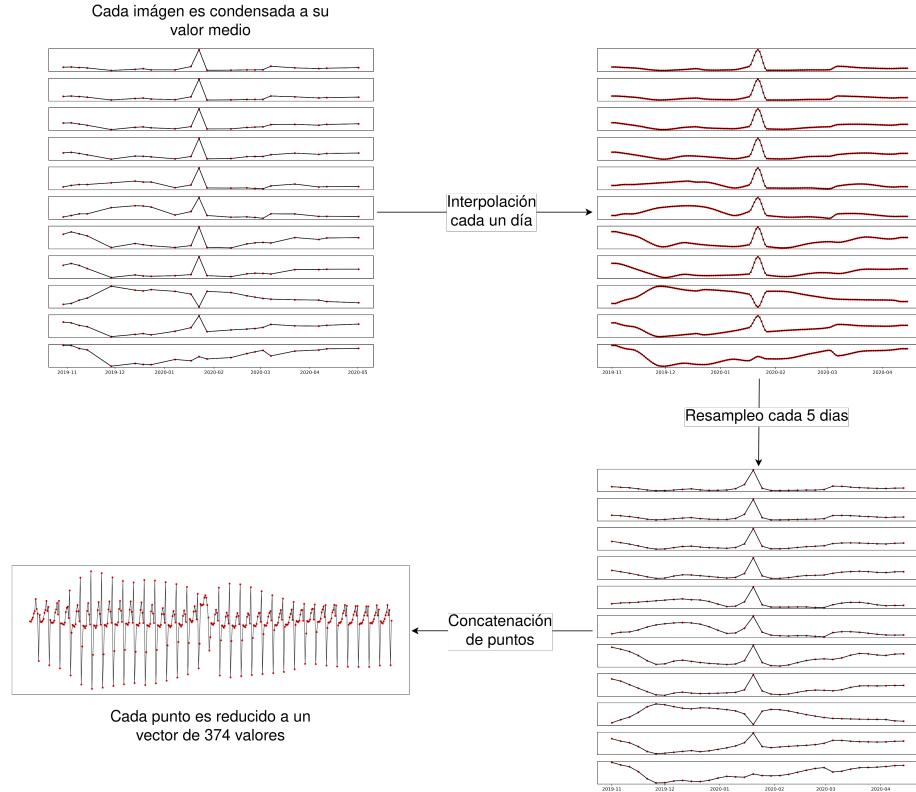


Figura 5: Preprocesado de un punto.

Modelo

Una vez creados los vectores de 374 valores para cada muestra se procedió a generar muestras artificiales con la técnica de over-sampling SMOTE de modo que las clases tuvieran al menos 25 observaciones. Posteriormente, se realizaron 5 muestreos aleatorios conteniendo el 90 % del total de las observaciones, dando como resultado 5 conjuntos de entrenamiento aleatorizados (bagging).

En base a estos conjuntos se entrenaron 5 modelos SVM, utilizando búsqueda de hiperparámetros y validación cruzada (grid_search.py). Con las predicciones del conjunto de modelos se entrenó un bosque aleatorio como meta-modelo para obtener las predicciones finales (main.py).

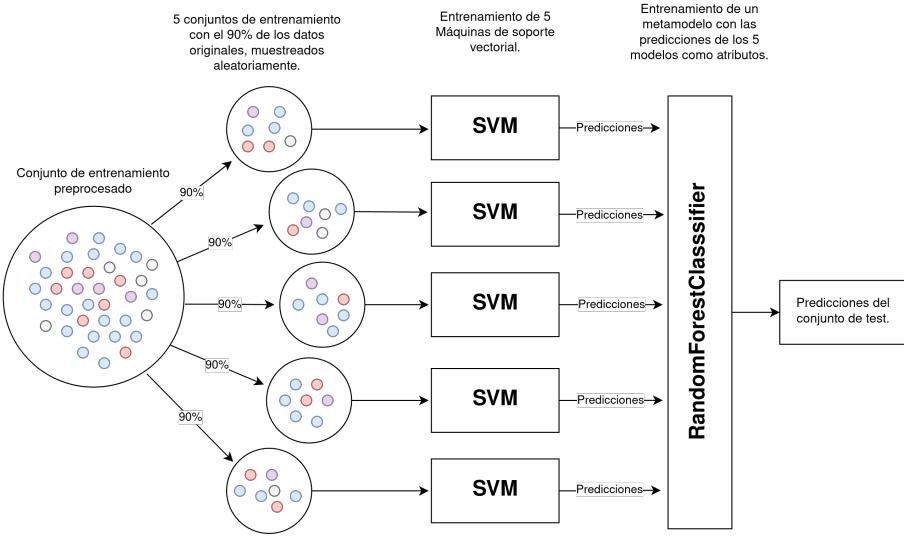


Figura 6: Esquema del ensamble de modelos.

Conclusiones

El mayor desafío de esta competencia fue entrenar un modelo que pueda predecir de manera correcta todas las clases, ya que la métrica de balanced accuracy las ubicó en pie de igualdad, con escasez de datos para las clases minoritarias (conjunto de entrenamiento extremadamente desbalanceado).

En la matriz de confusión para el conjunto de testeo, la clase X (No sabe), fue una de las que el modelo fue incapaz de predecir correctamente. Esto puede deberse a que el modelo quizás pudo predecir la clase verdadera en lugar de la clase X.

Adicionalmente algunos de los cultivos de soja de primera y segunda fueron confundidos entre sí, al igual que los cultivos de maíz temprano y tardío.

		M	N	S	P	X	s	R	U	m	A	B
True	M	131	2	2	3	0	2	0	0	0	0	0
	N	2	29	2	18	1	0	0	0	0	2	0
	S	5	1	211	1	1	8	2	0	0	0	0
	P	1	5	2	23	2	0	0	1	1	1	0
	X	4	2	10	1	0	5	0	0	0	0	0
	s	0	0	9	0	3	45	0	0	0	0	1
	R	0	0	0	0	0	0	3	0	0	0	0
	U	0	0	1	0	0	0	0	6	0	0	0
	m	1	0	0	0	0	0	0	0	1	0	0
	A	0	0	0	0	1	0	0	0	0	0	0
	B	0	0	0	0	0	1	0	0	0	0	2
		M	N	S	P	X	s	R	U	m	A	B

		S	M	N	A	P	R	aa	X	s	G	m	U	B
True	S	329	5	0	0	1	0	0	0	9	0	0	0	0
	M	6	203	1	0	0	0	0	0	0	0	0	0	0
	N	3	2	72	0	5	0	0	0	0	0	0	0	0
	A	0	0	0	12	0	0	0	0	0	0	0	0	0
	P	2	0	3	0	49	0	0	0	0	0	0	0	1
	R	0	0	0	0	0	12	0	0	0	0	0	0	0
	aa	0	0	0	0	0	0	12	0	0	0	0	0	0
	X	11	2	1	0	0	0	0	18	2	0	0	0	0
	s	7	1	0	0	0	0	0	0	81	0	0	0	0
	G	0	0	0	0	0	0	0	0	0	12	0	0	0
	m	0	0	0	0	0	0	0	0	0	0	12	0	0
U	0	0	0	0	1	0	0	0	0	0	0	11	0	
B	0	0	0	0	0	0	0	0	0	0	0	0	12	
		S	M	N	A	P	R	aa	X	s	G	m	U	B

Figura 7: (arriba) Matriz de confusión del conjunto de prueba (etiquetas facilitadas por la organización al finalizar la competencia). (abajo) Matriz de confusión del conjunto de entrenamiento.

El puntaje público de las predicciones fue de 0.768 (1er lugar), mientras que el puntaje privado fue de 0.503 (10mo lugar). Esta caída en el ranking muestra un fuerte overfitting sobre el dataset público. En pos de evitar este inconveniente se respetó la metodología de separar datasets de entrenamiento y de validación mediante cross-validation para el desarrollo del modelo; se hizo data augmentation y upsampling para mitigar la falta de datos, ensamble de modelos mediante bagging para reducir la varianza.

Sin embargo, se hicieron más de 230 presentaciones (lo que se conoce como human-loop en las competencias), no se sacó provecho de técnicas de regularización (parámetro C en los modelos SVM) y utilizamos una cantidad de atributos excesiva (11 capas x 34 instantes temporales = 374 atributos). Estas posiblemente hayan sido las causas del overfitting.

Cuadro 1: Reporte de clasificación del conjunto de prueba.

Class	Precision	Recall	F1-score	Support
A	0.000000	0.000000	0.000000	1
B	0.666667	0.666667	0.666667	3
M	0.909722	0.935714	0.922535	140
N	0.743590	0.537037	0.623656	54
P	0.500000	0.638889	0.560976	36
R	0.600000	1.000000	0.750000	3
S	0.890295	0.921397	0.905579	229
U	0.857143	0.857143	0.857143	7
X	0.000000	0.000000	0.000000	22
m	0.500000	0.500000	0.500000	2
s	0.737705	0.775862	0.756303	58
				555
macro avg	0.582284	0.621155	0.594805	
weighted avg	0.798161	0.812613	0.802740	
accuracy	0.812613			

Cuadro 2: Reporte de clasificación del conjunto de entrenamiento (las muestras artificiales fueron excluidas del reporte, no así del entrenamiento).

Class	Precision	Recall	F1-score	Support
A	1.000000	1.000000	1.000000	12
B	0.923077	1.000000	0.960000	12
G	1.000000	1.000000	1.000000	12
M	0.953052	0.966667	0.959811	210
N	0.935065	0.878049	0.905660	82
P	0.875000	0.890909	0.882883	55
R	1.000000	1.000000	1.000000	12
S	0.918994	0.956395	0.937322	344
U	1.000000	0.916667	0.956522	12
X	1.000000	0.529412	0.692308	34
aa	1.000000	1.000000	1.000000	12
m	1.000000	1.000000	1.000000	12
s	0.880435	0.910112	0.895028	89
				898
macro avg	0.960433	0.926785	0.937656	
weighted avg	0.931527	0.929844	0.927635	
accuracy	0.929844			

Referencias

- [1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.
- [2] R. Richter, X. Wang, M. Bachmann, and D. Schläpfer. Correction of cirrus effects in sentinel-2 type of imagery. *International journal of remote sensing*, 32(10):2931–2941, 2011.
- [3] J. Rouse. Monitoring the vernal advancement of retrogradation of natural vegetation, nasa/gsfg, type iii. *Final Report*, 371, 1974.
- [4] P. Sidike, V. Sagan, M. Maimaitijiang, M. Maimaitiyiming, N. Shakoor, J. Burken, T. Mockler, and F. B. Fritschi. dpen: deep progressively expanded network for mapping heterogeneous agricultural landscape using worldview-3 satellite imagery. *Remote sensing of environment*, 221:756–772, 2019.
- [5] H. Xu. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14):3025–3033, 2006.
- [6] Y. Zha, J. Gao, and S. Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International journal of remote sensing*, 24(3):583–594, 2003.