Sexto Congreso Nacional de Riego, Drenaje y Biosistemas
COMEII- 2021 / Hermosillo, Sonora

Artículo: COMEII-21022

Hermosillo, Son., del 9 al 11 de junio de 2021

IMPLEMENTACIÓN DE UNA MÁQUINA DE SOPORTE VECTORIAL PARA CLASIFICAR ZONAS DE INUNDACIÓN A PARTIR DE IMÁGENES DE RADAR

Juan Pablo Ambrosio Ambrosio[1]*, Juan Manuel González Camacho[2]

[1]Colegio de Postgraduados, Montecillo, México 56230

ambrosio.juan@colpos.mx, (+52) 5959574431 (*Autor de correspondencia)

[2]Colegio de Postgraduados, Montecillo, México 56230, jmgc@colpos.mx

Resumen

El uso de imágenes de radar de apertura sintética (SAR) representa una fuente valiosa de información para caracterizar regiones geográficas inundadas debido a que son insensibles a condiciones de nubosidad y obscuridad. El propósito de esta investigación fue identificar cuerpos de agua en una región del sureste de México, con base en el algoritmo máquina de soporte vectorial (SVM) para clasificar cuerpos de agua, infraestructura urbana y cobertura de suelo y/o vegetación a partir de imágenes SAR. La imagen SAR utilizada cubre una zona geográfica proyectada UTM Zona 15 Norte WSG84 localizada en los estados de Tabasco y Chiapas que fue preprocesada para disminuir los errores generados por el modo de adquisición de la imagen. Se definieron tres categorías de clasificación objetivo: A (Agua, áreas inundadas y cuerpos de agua), I (Infraestructura urbana y suelo desnudo) y V (Vegetación). El modelo se desarrolló en lenguaje Python y fueron entrenados y probados en predicción, a partir de una base de datos de 12,000 muestras con valores de amplitud de la imagen. El modelo SVM obtuvo una precisión global de clasificación 97.9 %(+/-0.003), F1 macro de 0.977 y área bajo la curva ROC igual a 1, 0.979, y 0.984 para clasificar las clases A, I y V respectivamente. Estos indicadores muestran el uso potencial del algoritmo de aprendizaje automático supervisado SVM y de las imágenes satelitales SAR para clasificar e identificar los cuerpos de agua; así como, resaltar su importancia en la evaluación de posibles impactos de inundaciones.

Palabras Clave: algoritmos de aprendizaje automático, clasificación multiclase, métricas de evaluación, zonas de inundación.

Introducción

En la actualidad es factible disponer de imágenes de satélite obtenidas mediante sensores remotos de radar de apertura sintética. A diferencia de los sensores ópticos, las imágenes SAR no dependen de la radiación solar reflejada o la radiación térmica emitida por la tierra, sino que emiten su propia radiación electromagnética para realizar sondeos. Por ello, una imagen SAR no es afectada por condiciones meteorológicas o si es de noche (Fernández & Soria, 2015). En el sureste de México se presentan, con frecuencia, intensos periodos de lluvias en los meses de agosto, septiembre, octubre y noviembre, que originan inundaciones y provocan pérdidas económicas, en diferentes sectores de la población y en ocasiones, pérdidas de vidas humanas (Sánchez et al., 2015). Por ello, es de interés, contar con métodos indirectos de identificación confiables, para identificar superficies afectadas por inundaciones; debido a que una identificación in situ, puede resultar costosa y tardada. En estas situaciones, el uso de imágenes de satélite y de modelos aprendizaje automático supervisado para clasificar zonas afectadas por inundaciones es de gran utilidad.

El área de estudio se encuentra delimitado por el polígono rojo de la Figura 1, de referencia al sistema geodésico de coordenadas geográficas World Geodetic System 1984 (WGS84) con una proyección al sistema de coordenadas Universal Transverse Mercator (UTM). La imagen SAR abarca una región entre los Estados de Tabasco y Chiapas. La zona de estudio abarca una superficie de 408,687 ha.

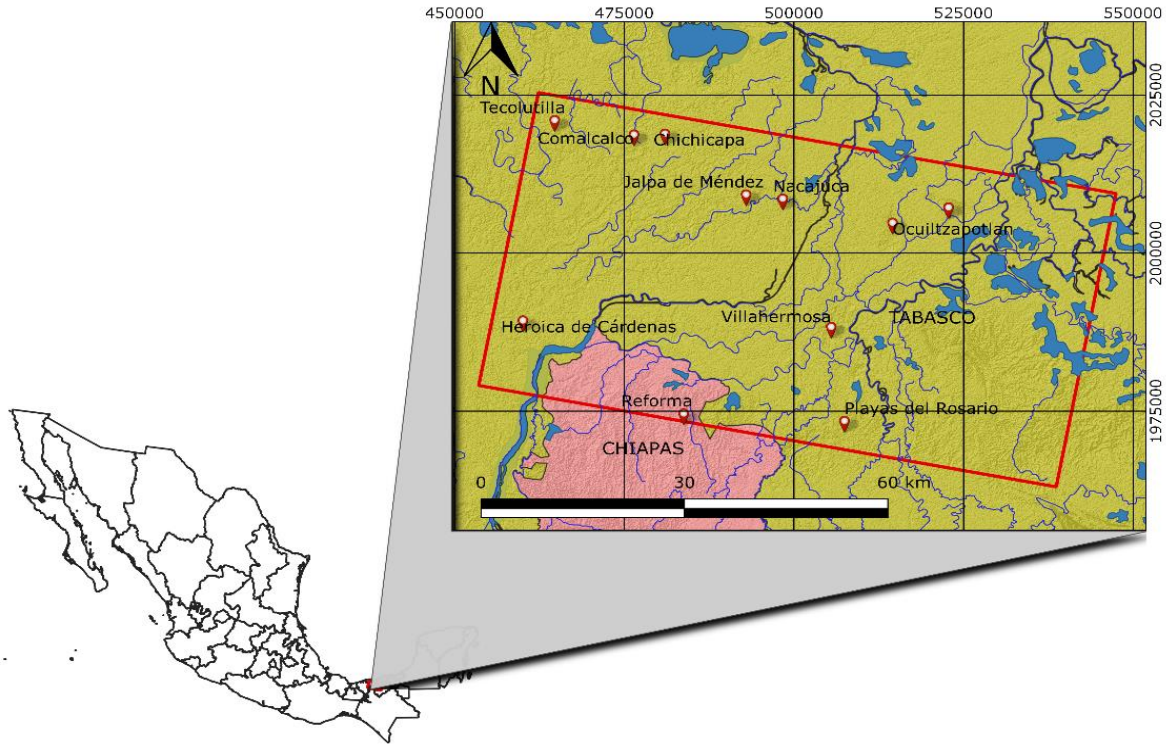

Figura 1. Zona de estudio, Tabasco y Chiapas, SRC UTM Zona 15 Norte WGS84.

Materiales y métodos

Datos de entrada

Para realizar este estudio se obtuvieron imágenes de radar de apertura sintética (SAR, synthetic aperture radar) del satélite Sentinel-1A. La descarga se hizo a través del sitio web del Instituto de Geofísica de la Universidad de Alaska Fairbanks (ASA, 2020). En el Cuadro 1 se describen las características específicas de la imagen SAR.

Cuadro 1. Características de la imagen SAR. Fuente: Copernicus Sentinel data [2017].

Detalles	Descripciones
Producto	S1A_IW_GRDH_1SDV_20171008T
State_Vector_Time	8-OCT-2017 12:00:16.264908
Nivel del producto	1, Producto estándar georeferenciado
Modo de adquisición	IW (Interferometric Wide)
Ancho de banda Azimuth	327 Hz
Tipo de producto	GRD (Grand Range Deteccion)
Polarización	Dual VV+VH
Frequencia	Banda C
Paso	Descendente

Las imágenes se procesaron con el software de uso libre sentinel application platform (SNAP) (ESA, 2019). Las imágenes SAR se transformaron a formato numérico con el software Matlab (MathWorks Inc., 2016). El procesamiento geoespacial de las imágenes SAR se realizó con el software QGIS (QGIS.org, 2020). Los modelos de aprendizaje automático para clasificación se implementaron en lenguaje Python con la biblioteca Scikit-learn (Pedregosa et al., 2011). El análisis de datos se realizó con un sistema de cómputo bajo ambiente Windows 10 de 64 bits, procesador Intel Core i5 @2.50 GHz, memoria RAM instalada de 8 GB.

Máquina de soporte vectorial

Una máquina de soporte vectorial (SVM, support vector machine) es un modelo de aprendizaje automático supervisado utilizado para clasificación o regresión (Raschka y Mirjalili 2017). El entrenamiento de SVM consiste en encontrar un hiperplano que separe las clases objetivo y el margen que existe entre los vectores de soporte sea máximo (Figura 2).

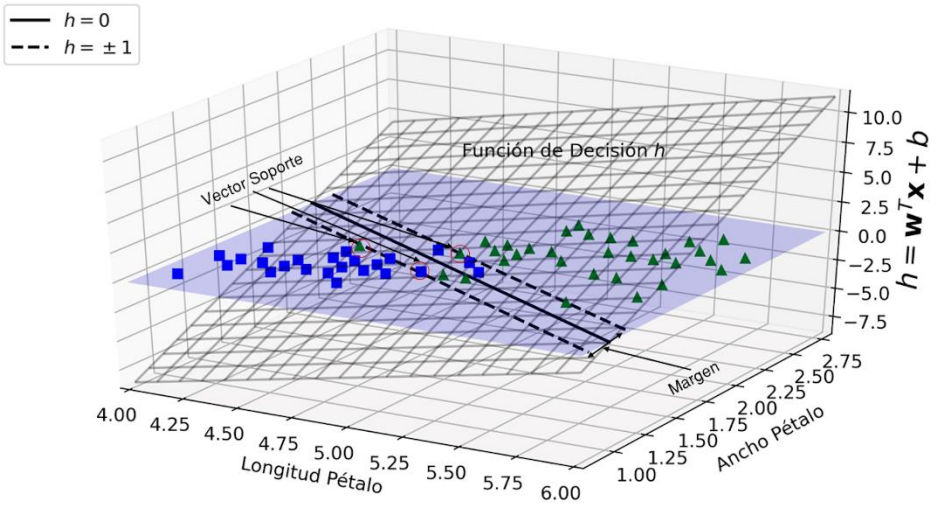

Figura 2. Función de decisión para el conjunto de datos iris con dos características (hiperplano) y límite de decisión la intersección de dos planos representada por la línea solida cuando el hiperplano h es igual a cero. Fuente: Géron (2019).

Smola y Schölkopf (2004) señalan que en un problema bidimensional el hiperplano de separación es una línea, cuya forma se expresa por:

$$f(x) = w^T x + b \tag{1}$$

dónde w es el vector de pesos, x es el vector de las características de entrada y; b es el sesgo. Para encontrar el máximo margen, una alternativa es minimizar la norma de w, esto es, $\|w\|^2 = w^T w$ y tratar como un problema de optimización convexa la expresión siguiente:

$$L = min\ \frac{1}{2}\|w\|^2 \tag{3}$$

$$sujeto\ a \begin{cases} y_i - w^T x_i - b \leq \varepsilon \\ w^T x_i + b - y_i \leq \varepsilon \end{cases}$$

dónde y_i representa la i-ésima respuesta deseada y ε es la máxima desviación permitida de las respuestas deseadas. Vapnik (1995) introduce el concepto de variables de holgura ζ para suavizar las restricciones lineales y alcanzar la convergencia en la optimización de problemas que no son linealmente separables. Con este enfoque Géron (2019) señala que la función de objetivo del modelo se convierte en:

$$L = min_{w,b,\zeta}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}\zeta_i \tag{4}$$

$$sujeto\ a \begin{cases} y_i(w^T x + b) \geq 1 - \zeta_i \\ \zeta_i \geq 0\ para\ i = 1,2,\dots,k \end{cases}$$

dónde y_i es la respuesta objetivo; x_i son las características de entrada; w es el vector de pesos o parámetros; b es el sesgo; ζ_i representa la variable de holgura; k es el número de clases y; C es el hiperparámetro que permite encontrar un balance entre reducir las variables de holgura y maximizar el margen, con una norma de los pesos mínima (Deisenroth et al., 2020)

Procesamiento de la imagen SAR

Las imágenes SAR se preprocesaron, previo a su análisis, para corregir errores en su adquisición. El preprocesamiento consiste en definir un subconjunto que representa el área de estudio y es manejable computacionalmente. Se aplica el filtro Lee (parámetros look=1 y tamaño de ventana 3x3) para eliminar moteados de la imagen que representan áreas sin información. Finalmente se aplica una corrección geométrica para georreferenciar la imagen y habilitar su posterior exportación en formato gráfico tif (Abdurahman et al., 2016; UN-SPIDER, 2020; Podest, 2018).

Selección de muestras

La clase A se integra por cuerpos de agua, tales como: ríos perennes e intermitentes, lagunas, y zonas agrícolas inundadas. Los pixeles de esta clase fueron seleccionados con un muestreo de conveniencia y se empleó la condición lógica AND entre el algoritmo crecimiento por región (disimilitud de 0.035) y la definición manual del umbral. La clase I representa la infraestructura física consistente de carreteras, edificios, casas, comercios, fábricas y suelo con poca o nula vegetación. La definición de esta clase se llevó a cabo mediante un muestreo de conveniencia y se aplicó la definición manual del umbral para la segmentación binaria. La muestra de grandes ciudades mostrada en formato uint8 se segmenta con el umbral 0.35-255. La clase V se integra por pixeles que representan la vegetación: bosque, pastizal, zona agrícola, selva. La selección de pixeles para esta clase se llevó a cabo mediante un muestreo de conveniencia, que consiste en definir pequeñas ventanas que representan en su totalidad un tipo de vegetación y submuestreo se lleva a cabo mediante el algoritmo crecimiento por región con disimilitud igual a 0.5.

La exportación de las bandas sigma0_VH y sigma0_VV se hace únicamente para los pixeles (x, y) con valor 1, resultado de la segmentación junto con sus respectivas coordenadas en formato txt. En la integración de la base de datos se lleva a cabo el análisis de duplicados para depurar y formar una base datos de 12000 muestras estratificadas.

Implementación del modelo

El entrenamiento del modelo consiste en seleccionar los mejores hiperparámetros, para ello se realiza una partición aleatoria, donde 90% de las muestras se emplea para el entrenamiento y 10% para la prueba del modelo. Con el conjunto de entrenamiento se efectúa una validación cruzada k=10, se crean 10 subconjuntos disjuntos estratificados por clase objetivo. Esto produce la creación de 10 modelos y cada uno emplea k-1 subgrupos para su entrenamiento y un subgrupo para su validación. El promedio del desempeño de los 10 modelos permite evaluar la precisión global de clasificación, y de manera adicional permite detectar problemas de sobre ajuste.

Métricas de evaluación

La búsqueda de los mejores híperparámetros se lleva a cabo mediante una búsqueda por cuadricula. Que consiste en explorar un conjunto de combinaciones de los híperparámetros para cada combinación se realiza una validación cruzada y se compara su desempeño mediante la precisión global de clasificación correcta (Acc). Una vez seleccionada el mejor modelo se evalúa el desempeño con diferentes métricas precisión (P), sensibilidad (S), especificidad (E), F1 score (F1s), F1 macro (F1m), curva ROC y su respectiva área bajo la curva (AUC). Todas las anteriores métricas surgen del análisis a detalle de la matriz de confusión del clasificador.

Discusión y resultados

Los hiperparámetros óptimos para el clasificador SVM fueron C = 500, gamma = 60 y keernel Gaussiano. La función de decisión empleada para la clasificación multiclase fue basada bajo el concepto de uno contra el resto. El entrenamiento del modelo con 90% de los datos obtuvo una precisión global de 97.75%.

La validación cruzada con k=10 empleando el modelo SVM con sus mejores hiperparámetros arrojó una precisión global de 97.92%(+/- 0.0005) en entrenamiento y 97.86%(+/-0.003) en la etapa de validación con 100% de los datos. El rendimiento del modelo en la etapa de entrenamiento es muy similar en la validación lo cual nos indica la ausencia de sobre ajuste del modelo al problema de clasificación y nos permite afirmar el buen rendimiento que se había mostrado en la etapa de prueba. También se puede apreciar que el modelo es bastante robusto debido a ello la diferencia en el rendimiento en cada iteración de la validación cruzada son muy similares generando una varianza mínima.

En la Figura 3A se puede apreciar la matriz de confusión para la décima iteración en la validación cruzada y en la Figura 3B sus respectivas curvas ROC. A través del análisis de la matriz de confusión se obtiene que la clase A tiene P=0.9852, S=1, E=0.9925, y F1s=0.9926; la clase I: P=0.9746, S=0.96, E=0.9875, y F1s=0.9673; y para la clase V: P=0.9725, S=0.9725, E=0.9863, y F1s=0.9725. La comparación de los valores de F1s se puede apreciar que el modelo clasifica muy bien la clase A seguido de la clase V y por último la clase I. Esto indica que el modelo clasifica de manera acertada los cuerpos de agua, pero tiene complicaciones para diferenciar entre la infraestructura y la vegetación. Este hecho se confirma con el área bajo la curva ROC de A=1, I=0.979 y V=0.984.

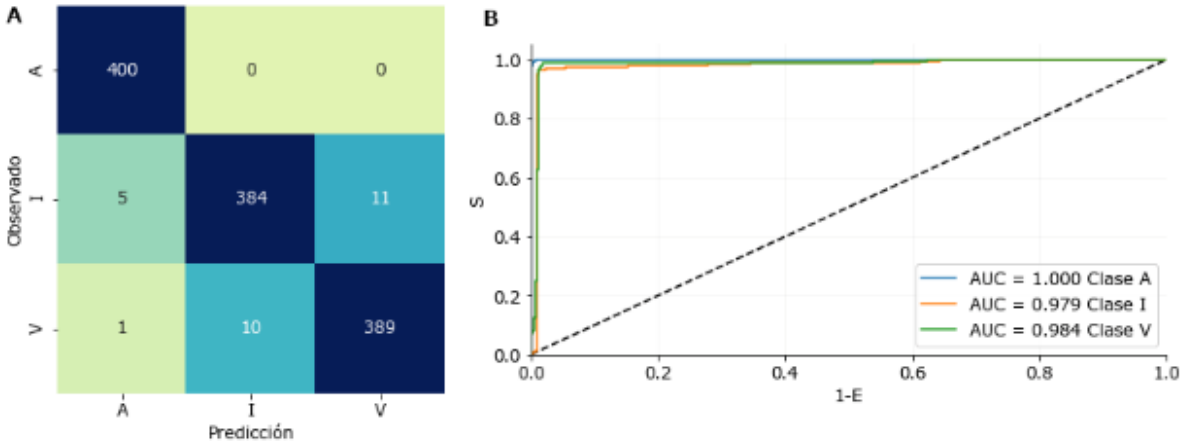

Figura 3. A. Matriz de confusión y B. curvas ROC de la décima iteración para el clasificador máquina de soporte vectorial.

Conclusión

El desarrollo de este trabajo permitió mostrar el procedimiento a emplear sobre las imágenes SAR a su previo análisis. La implementación de la máquina de soporte vectorial para el problema de clasificación multiclase arrojó una precisión global de 97.86% (+/- 0.003). A nivel de clase, la categoría A que representa los cuerpos de agua fue la que mejor se clasificó por encima de I y V. Los resultados obtenidos nos permiten considerar enormemente el uso potencial de estas imágenes para estudios hidrológicos y series de tiempo que permiten la identificación, cuantificación y evolución de fenómenos hídricos.

Literatura Citada

Abdurahman-Bayanudin, A. & Heru Jatmiko, R. (2016). Orthorectification of Sentinel-1 SAR (Synthetic Aperture Radar) Data in Some Parts Of South-eastern Sulawesi Using Sentinel-1 Toolbox. 2nd International Conference of Indonesian Society for Remote Sensing (ICOIRS), IOP Conference Series: Earth and Environmental Science, 47(012007). DOI: 10.1088/1755-1315/47/1/012007.

ASA, Alaska Satellite Facility. (2020). Recuperado de https://asf.alaska.edu/
European Spatial Agency (ESA), Scientific exploitation of operational missions (SEOM). (2019). Sentinel Application Platform (SNAP 7.0). Programa para el análisis y procesamiento de imágenes satelitales [software]. Recuperad de: http://step.esa.int/main/download/snap-download/

Deisenroth, M.P., Faisal, A.A. & Soon, C. (2020). Mathematics For Machine Learnig. Published by Cambridge University Press. 407 p. Disponible en: https://mml-book.com.

Fernández-Ordóñez, Y. & Soria Ruiz, J. (2015). Imágenes De Radar De Apertura Sintética Y Conceptos Básicos De Polarimetría. En: Fernández-Ordóñez, Y., Escalona-Maurice, M.J. & Valdez-Lazalde, J.R. (eds.). Avances Y Perspectivas De Geomática Con Aplicaciones Ambientales, Agrícolas Y Urbanas (pp. 37-66). Montecillo, Texcoco, México: Editorial Colegio de Postgraduados.

Géron, A. (2019). Hands-On Machine Learning With Scikit-Learn, Keras & TensorFlow, Second Edition. O' Reilly Media, Inc. Gravenstein Highway North, Sebastopol, CA 95472. 819 p.

MathWorks Inc. (2016). MATrix LABoratory (MATLAB R2016a). Plataforma matemática sumamente potente en la manipulación de matrices y análisis numérico [software]. Natick, Massachusetts.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Podest, E. (2018, 16 de abril). Imágenes radar (SAR) seminario NASA - Preprocesamiento, clasificación agua, tierra, Sentinel 1 (2/4) [video]. YouTube. Recuperado de: https://youtu.be/pVeh9ChATwA

QGIS.org. (2020). QuantumGis (QGIS versión 3.10.7-A Coruña). Sistema de Información Geográfica, Proyecto de fundación geoespacial de código abierto [software]. Recuperado de: https://qgis.org/es/site/

Raschka, S. & Mirjalili, V. (2017). Python Machine Learning, Second Edition. Published by Packt Publishing Ltd, Birmingham B3 2PB, UK. 850p.

Sánchez, A.J., Salcedo, M.A., Florido, R. & Mendoza, J.D. (2015). Ciclos de inundación y conservación de servicios ambientales en la cuenca baja de los ríos Grijalva-Usumacinta. Contactos, 97.

Smola, A. & B. Schölkopf. (2004). A tutorial on support vector regression. Statistics and Computing 14:199-222.

UN-SPIDER, United Nations Platform for Space-based Information for Disaster Management and Emergency Response. (2020). Step-by-Step: Mudslides and associated flood detection using sentinel-1 data. Recuperado de: https://un-spider.org/advisory-support/recommended-practices/mudslides-flood-sentinel-1/step-by-step

Vapnik, V. (1995). The Nature of Statistical Learning Theory. New York: Springer.