http://revistas.udistrital.edu.co/ojs/index.php/REDES/index

Clasificación de una imagen satelital empleando máquinas de soporte vectorial para cuantificar el área de Pinus Patula en una plantación

Classification of a satellite image using Vector Support Machines to quantify Pinus patula´s area in a plantation

Orlando Riaño Melo[1]
Carlos Daniel Acosta Medina[2]
Robert Orlando Leal Pulido[3]

Resumen

En este artículo se muestran los resultados obtenidos con el método de Máquinas de Soporte Vectorial (SVM), aplicado para clasificar una imagen de una plantación de *Pinus patula* en Antioquía (Colombia). El clasificador SVM empleó como kernel la función de base radial en la solución del problema de clasificar coberturas del suelo. La clasificación se realizó de manera supervisada tomando 1500 puntos de entrenamiento dentro del área de estudio que incluyeron seis clases de cobertura. La evaluación de exactitud temática obtenida se realizó a partir de 3000 puntos de validación. El estudio indica que el índice Kappa para la clasificación con el algoritmo SVM fue de 0,94, que se considera muy bueno y un porcentaje correctamente clasificado (PCC) del 96,26%.

Palabras clave: Función Kernel , hiperplano, Máquinas de Soporte Vectorial, vectores de soporte.

Abstract

This article shows the results obtained with the method Support Vector Machines (SVM), which is applied to classify an image of a *Pinus patula´s* plantation in Antioquía (Colombia). The SVM classifier used the kernel as radial basis function in the solution of the problem of classifying the land covering. The classification was performed under supervision, taking 1500 training points randomly within the area of study, including six cover classes. The thematic accuracy test obtained, was made starting at 3000 validation points. The study indicates that the Kappa index for classification with the SVM algorithm; was 0,94, which is considered as good and a correctly qualified percentage (PPC) of 96,26%.

Keywords: Kernel function, hyperplane, Support Vector Machines (SVM), support vectors.

1 orianom@unal.edu.co
2 cacostam@unal.edu.co
3 rolealp@udistrital.edu.co

Introducción

Los métodos estadísticos de mínima distancia y máxima verosimilitud, ampliamente usados para clasificar imágenes de percepción remota tienen algunas limitaciones, particularmente relacionadas con las hipótesis de distribución normal y restricciones a los datos de entrada [1], [2], [3]. Diversos estudios han demostrado que los sistemas expertos, las Máquinas de Soporte Vectorial (SVM), entre otros, pueden emplearse como métodos alternativos, con buenos desempeños, para resolver problemas de clasificación y de regresión en los que se han aplicado los enfoques tradicionales [4], [2], [3]. En los años recientes, las SVM han venido ganando terreno y reconocimiento [5], [6]. La teoría sobre SVM se basa en la idea de minimización de riesgo estructural propuesto por Vapnik [7].

Las SVM han mostrado mejor desempeño que las redes neuronales, en algunas aplicaciones [8], lo que ha permitido considerarlas como poderosas herramientas para resolución de problemas de clasificación. Una SVM transforma los datos de entrada por medio de una función kernel, que emplea el producto escalar, a un espacio de características de mayor dimensión, donde es posible encontrar el hiperplano óptimo que separe las clases y maximice el margen m entre ellas tal como se muestra en la figura 1.

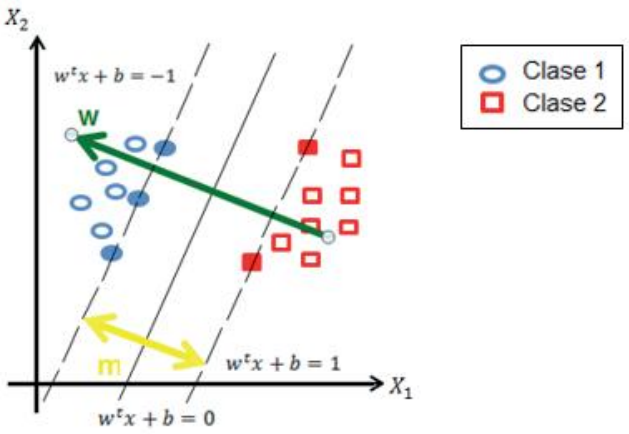

Figura 1. Hiperplano óptimo de separación entre dos clases en el espacio transformado.

Maximizar el margen m es un problema de programación cuadrática (QP) que puede ser resuelto con multiplicadores de Lagrange en su versión dual. El hiperplano óptimo encontrado se define combinando algunos puntos de entrenamiento seleccionados por el algoritmo, llamados vectores de soporte.

La clasificación automática de la cobertura de uso del suelo es importante y básica en las áreas de los Sensores Remotos (SR) y los Sistemas de Información Geográfica (SIG's). El éxito en el desarrollo de los sistemas de clasificación, permite mejorar su exactitud temática generando grandes beneficios medioambientales y económicos. La clasificación de coberturas y uso del suelo a partir de imágenes multiespectrales es un problema complejo debido al traslape en el espacio espectral entre las diferentes clases [9].

El objetivo de este estudio consiste en clasificar y cuantificar la cobertura del suelo en la imagen multiespectral LANDSAT 8-OLI empleando el clasificador SVM, y determinar la exactitud de la clasificación brindada por este método. Se evaluará la hipótesis que la utilización de SVM permite obtener un resultado catalogado como bueno.

Fundamentos

En esta sección se dará una breve introducción teórica a las SVM's, para su empleo en problemas de clasificación.

Máquinas de Soporte Vectorial (SVM's)

Una Máquina de Soporte Vectorial es un sistema de aprendizaje automático que permite resolver problemas de clasificación y regresión de manera eficiente. La SVM se basa en la Teoría de Aprendizaje Estadístico [10]. Su éxito se debe a que posee las siguientes ventajas: (i) una fundamentación matemática sólida, (ii) estar basadas en el concepto de la minimización del riesgo estructural [11], es decir, minimizar la probabilidad de una clasificación errónea sobre nuevos ejemplos, y (iii) disponer de potentes herramientas y algoritmos para encontrar la solución óptima rápida y eficientemente.

2.1.1 Caso de separabilidad lineal entre dos clases

Para resolver un problema de clasificación, la SVM debe definir una superficie de decisión adecuada con base en el conjunto de datos de entrenamiento. Esa superficie de decisión es un hiperplano que separa los patrones de entrenamiento en dos clases, según se hallen a uno u otro lado de él.

En un problema separable linealmente existen muchos hiperplanos que pueden clasificar los datos como se puede observar en la figura 2.

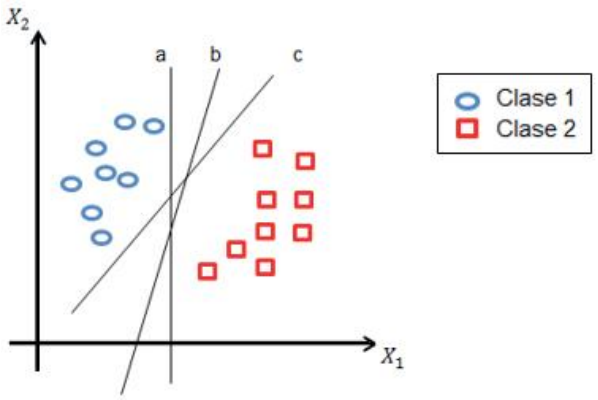

Figura 2. Tres hiperplanos a, b y c que separan las dos clases.

Así, las SVM hallan el único hiperplano que maximiza la distancia (llamada margen) entre él y el dato más cercano de cada clase. Se le llama *hiperplano de separación óptima* y viene dado por:

$$w^t \bullet x + b = 0 \quad \text{donde} \quad w, x \in \Re^n \text{ y } b \in \Re$$

Se debe encontrar el vector de pesos w y el valor de z, que contiene la ponderación de cada atributo, indicando la cantidad de aporte en el proceso de clasificación y el umbral de decisión, respectivamente [12]. Así, se puede separar el punto x_i de acuerdo con la función signo del hiperplano. Esto es:

$$f(x_i) = sign(w^t \bullet x_i + b) = \begin{cases} w^t \bullet x_i + b \geq 1, & y_i = 1 \\ w^t \bullet x_i + b \leq -1, & y_i = -1 \end{cases}$$

donde el signo resultante indicará a cual clase y_i pertenece un dato determinado. Esta expresión genera gran economía computacional, debido a que la función no se realiza para todos los puntos de entrenamiento, sino únicamente sobre los vectores de soporte que por lo general es un pequeño porcentaje de n.

2.1.2 Caso de separabilidad no lineal

Al no existir una superficie de decisión lineal apropiada en el espacio de entrada, se transforma el vector de entrada a un espacio de mayor dimensión $\Re^c$ llamado espacio de características τ, que está dotado de producto escalar. Al elegir el espacio τ apropiado, se realiza la transformación y se busca el hiperplano de separación óptima de la misma manera del caso anterior y que será lineal en $\Re^n$, pero es un hiperplano no lineal en el espacio de entrada $\Re^n$. Al suponer que esta transformación se realiza mediante una función no lineal de la forma:

$$\phi(X): \Re^n \rightarrow \Re^c, c>n, \text{definida por}$$

$$\phi(X)=(\phi_1(X),\phi_2(X),...,\phi_c(X))$$

Se observa una dificultad para hallar el hiperplano de separación óptima en τ debido a que replantea el objetivo de minimización donde los vectores de entrenamiento solo aparecen en la forma de producto escalar, pero definido en el espacio de características. El cálculo de

$$\phi(X_i) \bullet \phi(X_j)$$

es exigente computacionalmente, debido a que $c > n$. Este grave inconveniente se soluciona con las funciones kernel.

2.1.3 Función kernel

Como no se tiene conocimiento de la función de transformación, ϕ(X) el cálculo de la función de decisión es imposible. Sin embargo, la SVM posee la buena propiedad que no es necesario tener ningún conocimiento acerca de ϕ(X). Sólo se necesita una función K, llamada *kernel*, que calcule el producto escalar de los puntos de entrada en el espacio de características [11], es decir:

$$K: \Re^n \text{ x } \Re^n \rightarrow \Re, \text{definida por}$$
$$K(X_i, X_j) = \phi(X_i) \bullet \phi(X_j)$$

Con el empleo de esta función se obtiene el producto escalar en el espacio de características, pero realizando el cálculo en el espacio de entrada cuyo grado de complejidad es menor.

Los tipos de funciones kernel más utilizados son: polinomial-homogéneo, perceptrón, función de base radial y sigmoidal [13].

Datos y métodos

3.1 Datos

El área de estudio es una finca en Angostura (Antioquía-Colombia) que ocupa 824,9 hectáreas en la cual el Pinus patula ocupa la mayor superficie. La localización de la finca se muestra en la figura 3.

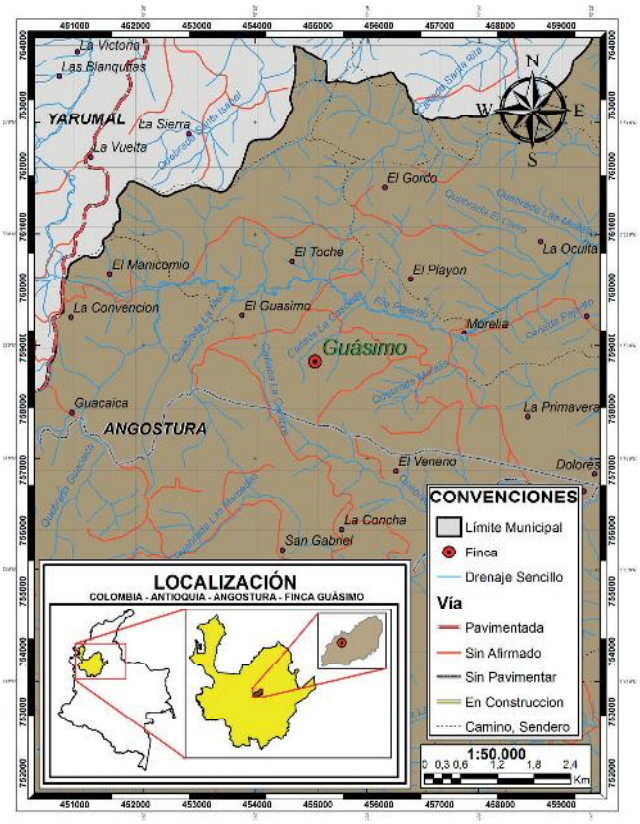

Figura 3. Localización del área de estudio.

En la zona de estudio se identificaron las siguientes clases de cobertura y uso: *Pinus patula* (Pn), suelo descubierto (Sd), nubes (Nb), sombras (Sm), cuerpos de agua (Ag) y vegetación herbácea baja (Vh).

Se empleó una imagen multiespectral de Path 9 y Row 55 del satélite LANDSAT 8-OLI con resolución espacial de 30 m por 30 m en los canales 1 a 7 y 9 y 15m por 15 m en el canal pancromático 8 [14], tomada el 26 de noviembre de 2015 [15].

El procesamiento de los datos se realizó utilizando la implementación del clasificador SVM, incluido en el paquete de software estadístico libre R versión 3.3 [16].

3.2 Método

Los pasos generales seguidos para hallar la clasificación con SVM, se muestra en la figura 5 y se describe a continuación.

3.2.1 Preparación de la imagen y archivo vector con clases

La imagen del satélite LANDSAT 8-OLI de la zona de estudio tiene 30 m por 30 m de resolución espacial en la toma. Sin embargo, se llevó a cabo un remuestreo a 15 m por 15 m. Los polígonos de la figura 7 muestran las zonas de entrenamiento para las seis clases definidas.

3.2.2 Cargue de la imagen

Se procedió a especificar el directorio de trabajo en el que se copió la imagen, se cargaron las librerías *foreign, kernlab, maptools, mda, raster, rgdal, sp y vcd* y se apilaron las bandas en el objeto imagen y se leyó la capa que contiene las diferentes clases tomadas en el terreno (puntos de entrenamiento).

3.2.3 Cálculo de estadísticas y elaboración de la composición falso color estándar

En este paso, se realizaron las estadísticas unibanda y multibanda de la escena. En la figura 6, se ilustra la matriz de correlación entre las distintas bandas de la escena.

Como era de esperarse se obtuvieron valores altos entre las bandas del visible, debido a que están adyacentes espectralmente.

Se realizó la composición falso color estándar OLI543 (RGB), que se puede apreciar en la figura 4. En rojo aparece la vegetación vigorosa y/o herbácea y *Pinus patula*. En marrón se identifica vegetación arbustiva muy variable en función de la densidad y del tono del sustrato. En blanco las áreas sin vegetación pero de alta reflectividad: nubes y suelos desnudos.

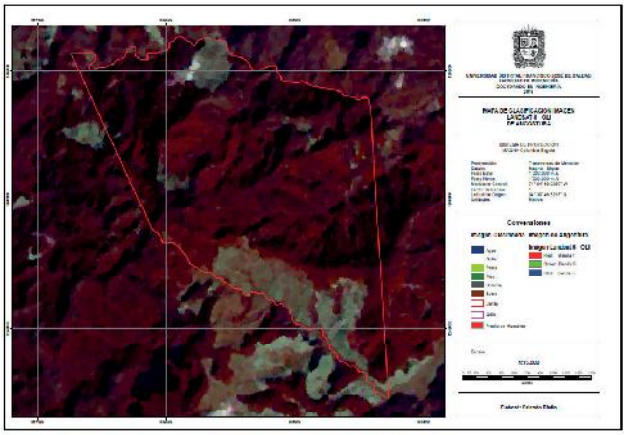

Figura 4. Composición Falso color estándar de la finca.

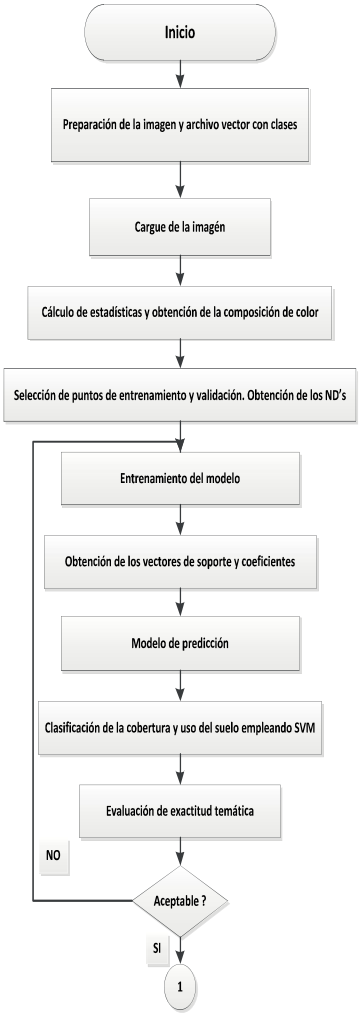

Figura 5. Diagrama de flujo del método SVM.

3.2.4 Selección de puntos de entrena-miento y validación. Obtención de los ND´s

En las primeras pruebas se emplearon 500 puntos de entrenamiento y 1000 de validación escogidos aleatoriamente y sin semilla, lo que no permitió una buena exactitud temática. Se cambió la muestra a 1000 y 2000 puntos respectivamente, lográndose una pequeña mejoría en los indicadores temáticos. Finalmente se obtuvieron mejores resultados con 1500 puntos de entrenamiento (figura 7) y 3000 puntos de validación (figura 8). Se extrajeron los niveles digitales y los valores de clase de todos y cada uno de los puntos muestra.

3.2.5 Entrenamiento del modelo

Para tal fin se empleó el algoritmo ksvm de R, empleando la función kernel de base radial, con valor de 2,5 para sigma y 50 como costo de penalización para que haya un balance entre la maximización del margen y la asignación equivocada de clases.

3.2.6 Obtención de los coeficientes $\alpha's$ y b

El algoritmo determinó vectores de soporte y 15 valores de $\alpha' \sigma \psi \beta$ para la definición de la función objetivo.

3.2.7 Modelo de predicción

Se aplicó la instrucción *predict* de R para usar el modelo SVM para predecir la clase de cobertura y uso de toda la imagen y con el comando *raster* de R se creó una capa con las clases de cobertura obtenidas por el algoritmo SVM.

3.2.8 Clasificación de la cobertura por SVM

Se aplica la predicción a la totalidad de la imagen, lo que permite obtener el resultado que se muestra en la figura 9.

	OLI-1	OLI-2	OLI-3	OLI-4	OLI-5	OLI-6
OLI-1	1.000	0.996	0.996	0.898	0.949	0.975
OLI-2	0.996	1.000	0.998	0.922	0.967	0.985
OLI-3	0.996	0.998	1.000	0.911	0.968	0.988
OLI-4	0.898	0.922	0.911	1.000	0.950	0.932
OLI-5	0.949	0.967	0.968	0.950	1.000	0.992
OLI-6	0.975	0.985	0.988	0.932	0.992	1.000

Figura 6. Matriz de correlación de la imagen.

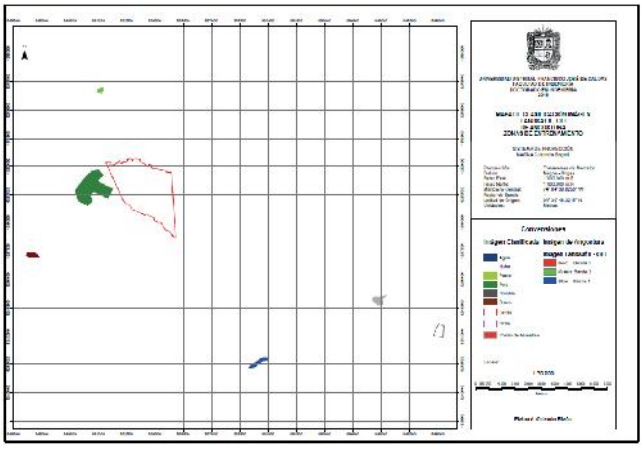

Figura 7. Capa vector con 1500 puntos de entrenamiento.

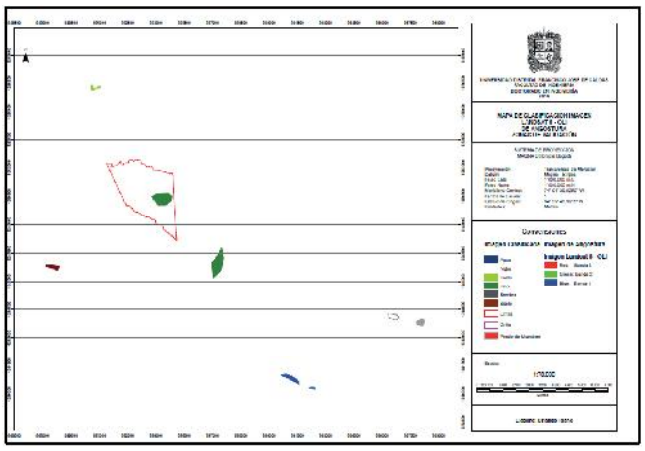

Figura 8. Capa que muestra la distribución de 3000 puntos para validación.

3.2.9 Evaluación de exactitud temática

En los casos en que las clases obtenidas no coinciden completamente con la verdad obtenida en el terreno, se pueden cambiar u optimizar los valores para los parámetros de *C* y *sigma* en el algoritmo para mejorar la clasificación. De todas formas, se hace necesario realizar una evaluación de la exactitud temática de la clasificación obtenida, que se logra evaluando la matriz de confusión, el porcentaje correctamente clasificado (PCC) y el índice Kappa.

Resultados

La figura 9 muestra el resultado obtenido de la clasificación con la aplicación del algoritmo SVM. Las clases obtenidas y su superficie dentro de la finca fueron: con *Pinus patula* 704,32 ha (85,38%), suelo desnudo 50,42 ha (6,11%) y vegetación herbácea 70,15 ha (8,51%).

La figura 10 presenta la matriz de confusión que se obtiene para esta clasificación temática del uso y cobertura del suelo de la plantación con el uso de una SVM.

Esta clasificación produjo un porcentaje correctamente clasificado (PCC) del 96,26% y un índice Kappa de 0,94, que satisface el porcentaje de clasificación correcta y que la hace aceptable [13], además de cumplir adecuadamente con el estándar de ser mayor que 0,8 para ser calificada como muy buena [17].

El intervalo de confianza al 95% de PCC va de 95,13% a 98,33%.

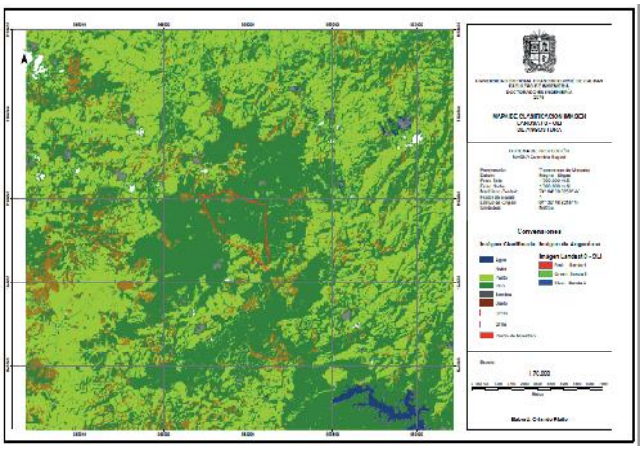

Figura 9. Clasificación con el uso de SVM.

Así, se confirma la hipótesis que el clasificador SVM ofrece un resultado muy bueno.

		Clases en el terreno						
Predictorias		Ag	Nb	Vh	Pn	Sm	Sd	Total
	Ag	**389**	0	0	0	0	0	**389**
	Nb	0	**212**	0	0	0	0	**212**
	Vh	0	0	**104**	3	0	3	**110**
	Pn	0	0	26	**1826**	1	0	**1853**
	Sm	0	0	0	12	**154**	0	**166**
	Sd	0	0	62	5	0	**203**	**270**
	Total	**389**	**212**	**192**	**1846**	**155**	**206**	**3000**

Figura 10. Matriz de confusión para la clasificación obtenida con SVM.

Discusión

Debido a que el valor de error para la matriz de confusión y el valor del índice kappa superan en buena medida los estándares mínimos de aceptación, se puede inferir que el empleo del clasificador SVM es sufícientemente exacto, una vez se determinen adecuadamente los valores de los parámetros: sigma y costo de penalización, lo mismo que el número de puntos de entrenamiento y validación.

Los análisis realizados anteriormente permiten observar que no sólo se obtiene una frontera de separación entre clases, sino la mejor de todas las posibles (el hiperplano de separación óptimo) con el clasificador SVM.

6. Conclusiones

Este trabajo presentó una aplicación del empleo del uso de SVM para la clasificación temática de una imagen multiespectral con muy buenos resultados en términos de exactitud temática.

El proceso matemático no sólo permite obtener una buena frontera de separación sino la óptima, en el sentido de maximizar el margen entre clases con SVM.

Una vez obtenido el modelo para SVM, es fácil su implementacion y muestra alto desempeño en la clasificación de los datos.

El éxito del algoritmo SVM radica en la buena selección de la función kernel, debido a que los cálculos se realizan en el espacio de entrada con esta función.

El uso de una función kernel no hace necesario el conocimiento explícito de la función $f(X)$

En cuanto a la clasificación que se llevó a ca-bo en si, la solución presentada es muy buena con SVM.

Se pudo observar, durante el desarrollo de la clasificación, que a medida que aumentaban el número de puntos de entrenamiento y validación, el resultado iba mejorando de la misma manera.
SVM fue capaz de discriminar el *Pinus patula,* mientras que en una pequeña proporción se presenta confusión entre vegetación herbácea y sombras con *Pinus patula* por su gran parecido espectral en las bandas del visible.

Las SVM presentan la debilidad de requerir eficientes metodologías para definir los parámetros de inicialización.

Referencias bibliográficas

[1] P. M. Mather. *Computer processing of remotely-sensed images: An introduction,* 3ed edition. Chichester. John Wiley and Sons. 2004.

[2] J.F. Mas, J.J. Flores. "The application of artificial neural networks to the analysis of remotely sensed data". *International Journal of Remote Sensing*, 29 (3), 617-663. 2008.

[3] G. Mountrakis, J. Im, C. Ogole. "Support vector machines in remote sensing: A review". *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (3), 247-259. 2011.

[4] M. H.Tseng, S. J. Chen, G. H.Hwang and M. Y. Shen. "A genetic algorithm rule-based approach for landcover classification". *ISPRS J. Photogram. Remote Sensing*, 63: 202-212. 2008

[5] C. Burges, Schölkopf and A. Smola. *Advances in kernel methods: Support vector machines.* Cambridge, MA: MIT Press. 1999.

[6] L. Breiman. "Random Forest". *Machine Learning.* 45(1), 5-32. 2001.

[7] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer_Verlag, New York. 1995.

[8] C. Burges. "A tutorial on support vector machines for pattern recognition". *Data Mining and Knowledge Discovery*, vol. 2, no. 2. 1998.

[9] I. Lizarazo, S. Mesa, R. Cuitiva. "Clasificación de imágenes usando redes neuronales: Bases matemáticas". *Revista Científica no. 7. Centro de Investigaciones y Desarrollo Científico (CIDC),* Universidad Distrital. 2005.

[10] V. Vapnik. *Statistical Learning Theory.* Wiley, New York. 1998.

[11] V. Kecman. *Learning and Soft Computing.* MIT Press, London. 2001.

[12] G. Betancourt. "Las máquinas de soporte vectorial". *Scientia et Technica.* Año X!, No. 27: 67-72. 2005.

[13] B. Tso. and P. Mather. *Classification Methods for Remotely Sensed Data.* Taylor and Francis Corp, London, UK. 2009.

[14] A. Ariza. *Descripción y corrección de datos Landsat 8. LDCM. Versión 1.0.* Instituto Geográfico Agustín Codazzi. Bogotá. 2013.

[15] USGS. "Landsat 8 Product. U.S. Department of the Interior". *U.S. Geological Survey.* 2015.

[16] R version 3.3. *The R Foundation for Statistical Computing.* ISBN 3-900051-07-0. 2016.

[17] R. B. Congalton. "A review of assessing the accuracy of classifications of remotely sensed data". *Remote Sensing of Environmental,* 37: 35-46. 1991.

De los autores

Orlando Riaño Melo: Matemático de la Universidad Nacional de Colombia. Colombia. Ingeniero de sistemas de la Universidad Distrital Francisco José de Caldas. Colombia. Especialista en Sistemas de Información Geográfica y Sensores Remotos. Convenio IGAC- Universidad Distrital. Colombia. Magister en Geomática de la Universidad Nacional de Colombia. Colombia. Docente Titular de la Universidad Distrital Francisco José de Caldas. orianom@unal.edu.co

Carlos Daniel Acosta Medina: Matemático de la Universidad Nacional de Colombia. Colombia. Magister en Matemáticas de la Universidad Nacional de Colombia. Colombia. Doctor en Matemáticas de la Universidad Nacional de Colombia. Colombia Postdoctor en Ingeniería Matemática de la Universidad de Concepción. Chile. Docente de la Universidad nacional de Colombia- Sede Manizales. cacostam@unal.edu.co

Robert Orlando Leal Pulido: Ingeniero Forestal de la Universidad Distrital Francisco José de Caldas. Colombia. Magister en Ingeniería Industrial de la Universidad Distrital Francisco José de Caldas. Docente de la Universidad Distrital Francisco José de Caldas. rolealp@udistrital.edu.co