CLASIFICACIÓN DE LA COBERTURA Y DEL USO DEL SUELO URBANO USANDO IMÁGENES DE SATÉLITE Y ALGORITMOS SUPERVISADOS DE INTELIGENCIA ARTIFICIAL

URBAN LAND-COVER AND LAND-USE CLASSIFICATION FROM SATELLITE IMAGES USING ARTIFICIAL INTELLIGENCE ALGORITHMS

Fecha de recepción: 30 de julio de 2008 / **Fecha de aprobación:** 30 de septiembre de 2008

Iván Lizarazo

Resumen

Este artículo presenta una comparación del funcionamiento y de las capacidades de dos algoritmos de Inteligencia Artificial, *retro-propagación* (redes neuronales artificiales) y *árboles de decisión*, que representan métodos alternativos para la clasificación digital de imágenes de sensores remotos frente a los algoritmos estadísticos convencionales. En particular, se muestran las ventajas y limitaciones de las nuevas técnicas, teniendo en cuenta conceptos teóricos al igual que la evaluación de los resultados obtenidos en su aplicación en la clasificación de cobertura y uso del suelo en una zona piloto de la ciudad de Bogotá, Colombia.

Palabras clave: clasificación de imágenes digitales, cobertura del suelo, uso del suelo, redes neuronales artificiales, árboles de decisión.

Abstract

This paper aims to compare the capabilities of two algorithms, *back-propagation* (artificial neural networks) and *decision trees*, as alternative methods for digital classification of remotely sensed images. In particular, advantages and limitations of the new techniques are examined using both theoretical concepts and experiences in their implementation in a land-cover and land-use classification experiment in a study zone located in Bogotá, Colombia.

Key words: Digital Image Classification, Land-Cover, Land-Use, Artificial Neural Networks, Decision Trees.

1. Introducción

El uso de árboles de decisión y de redes neuronales artificiales en la clasificación de datos de sensores remotos ha ganado popularidad en los últimos años. Algunos trabajos han evaluado su capacidad frente a la de los clasificadores estadísticos convencionales (Paola y Schowengerdt, 1995; Fitzgerald y Lees, 1994). Otros han propuesto métodos para mejorar su desempeño y confiabilidad (p.e., German *et al.*, 1999; Kanellopoulos y Wilkinson, 1997; Gahegan *et al.*, 1998). La principal aplicación de los métodos señalados ha sido la clasificación de la cobertura y el uso del suelo en zonas rurales, como lo indican Tso y Mather (2001). Algunos trabajos más recientes reportan su aplicación en zonas urbanas (Pesaresi y Benediktsson, 2000; Schiavon *et al.*, 2003; Del Frate *et al.*, 2004; Lizarazo, 2005; Lizarazo *et al.*, 2006).

La clasificación de la cobertura del suelo urbano usando bandas espectrales es un asunto complejo debido a que las diferentes clases se superponen en el espacio espectral (Lizarazo, 2005). La figura 1 muestra, a manera de ejemplo, la superposición existente entre las clases Agua, Arbustos, Construcciones y Vías en el espacio de atributos correspondiente a las *bandas XS2 (Verde), XS3 (Roja) y XS4 (Infraroja) de una imagen QuickBird correspondiente a la zona urbana de Bogotá,* Colombia. Cada clase está delimitada por una esfera que encierra los píxeles que conforman la muestra representativa de cada una de las coberturas presentes en dicha zona.

En este artículo se discute inicialmente el problema de realizar una clasificación supervisada. Luego se describen de manera breve las redes neuronales y los árboles de decisión, desde el punto de vista de su funcionamiento y de la configuración requerida para realizar las tareas de clasificación supervisada. Enseguida se plantean algunos elementos que deben tenerse en cuenta para obtener resultados apropiados con dichas técnicas de clasificación. Para finalizar, se presentan algunos elementos comparativos entre los dos métodos, a la luz de los resultados obtenidos en la aplicación de sus correspondientes algoritmos en la zona de estudio.

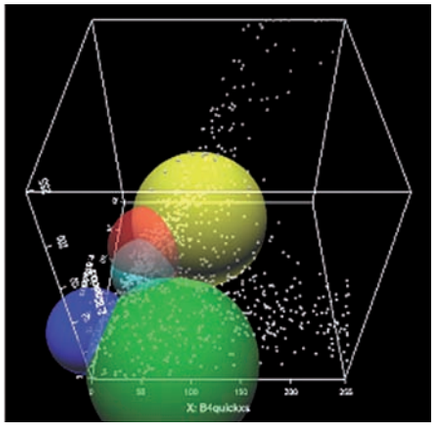

Figura 1. Espacio tridimensional de atributos definido por tres bandas de una imagen QuickBird multiespectral. Las esferas representan las muestras de entrenamiento utilizadas para la clasificación de la cobertura del suelo en el caso de estudio reportado en este artículo. *El agua está representada en color azul intenso, los arbustos en verde, las construcciones en café y rojo y las vías asfaltadas* en azul petróleo. Es evidente que existe traslapo espectral entre las muestras y que, por tanto, existen problemas de separabilidad entre clases.

2. Clasificación de imágenes

La clasificación de una imagen es una tarea que se realiza con el propósito de convertir datos cuantitativos (generalmente los niveles digitales de los píxeles en cada banda espectral) en datos cualitativos (temas o clases que son importantes en un dominio específico del conocimiento) (Richards y Jia, 1999). La motivación principal de una clasificación es la de representar un fenómeno que ocurre sobre la superficie terrestre a partir de la generalización y agrupación de datos obtenidos mediante sensores remotos (Jensen, 2005). Una buena clasificación debe representar de manera exacta la realidad a partir de las características pictórico-morfológicas presentes en las imágenes (ASPRS, 1997).

La función que relaciona las entradas (los atributos que se consideran relevantes) y las salidas (las clases deseadas) se puede establecer de manera analítica. Si el problema es muy complejo para ser resuelto analíticamente, es posible usar aproximaciones heurísticas, como las ofrecidas por las técnicas de Inteligencia Artificial conocidas como aprendizaje de máquina inductivo. La comparación entre las técnicas de redes neuronales y árboles de decisión, las explicaciones sobre la manera de construir el espacio de atributos y las opciones que existen para realizar la partición de dicho espacio, al igual que la descripción de las capacidades de generalización de las dos técnicas se ha realizado tomando como referencia básica el trabajo realizado por Gahegan y West (1998).

Los algoritmos de clasificación de imágenes se pueden dividir en dos grupos, supervisados y no supervisados. Las técnicas supervisadas tienen una fase de entrenamiento en la cual se usan muestras representativas de las clases seleccionadas para establecer un modelo del proceso de clasificación. Las técnicas no supervisadas no requieren ningún entrenamiento y tampoco suponen la definición previa de una clase, ellas se basan únicamente en la agrupación de los datos usualmente utilizando alguna métrica euclidiana. En el numeral siguiente se explica el proceso de clasificación supervisada cuya tarea principal es producir una función general que se aplique a los datos de entrada para obtener las clases deseadas, a partir de la generalización

de las funciones específicas que se obtienen usando una muestra de datos que sirve de entrenamiento.

3. El proceso de clasificación supervisada

El proceso de clasificación supervisada consiste en una secuencia de pasos genéricos que se utilizan independientemente del algoritmo utilizado, ya sea árboles de decisión (DT), redes neuronales artificiales (ANN) o Clasificador de Máxima Probabilidad (MLC). El proceso que se indica enseguida está basado en el propuesto por Richards y Jia (1999).

1. Definición del Problema: las clases objetivo deben ser definidas, ω_i, $i = 1, \ldots R$, al igual que el conjunto de atributos que se utilizarán para identificar las clases x_j, $j = 1, \ldots n$.
2. Selección de las muestras de entrenamiento para cada una de las clases objetivo. Para que la clasificación sea exacta, esas muestras deben ser 'representativas' de cada clase. Es recomendable realizar algún tipo de análisis exploratorio para establecer si las clases se están caracterizando de manera correcta, al igual que entender si existen dificultades para la separación de las clases. Si se descubre algún problema de caracterización, se deben modificar las clases objetivo y/o cambiar los atributos que se utilizarán para diferenciarlas.
3. Construcción del clasificador usando criterios predeterminados En Inteligencia Artificial (IA) este paso se conoce de manera indistinta como *fase de entrenamiento* ó como *aprendizaje inductivo* (Briscoe y Caelli, 1996).
4. Validación de los resultados del entrenamiento. Este paso busca evaluar el desempeño del clasificador usando datos nuevos que no se han utilizado en el entrenamiento. Si los resultados no son satisfactorios, puede ser necesario repetir el proceso de entrenamiento utilizando criterios diferentes.
5. Aplicación del clasificador a todos los datos de la imagen para producir una clasificación de toda el área de interés.

   Es importante tener en cuenta que la selección de un clasificador específico afecta principalmente el paso 3 y que tiene un impacto menor en los demás pasos. Sin embargo, los factores limitantes más grandes en una clasificación tienen que ver con los pasos 1 y 2; específicamente la capacidad de diferenciar las clases depende del cuidado que se tenga en la selección de las clases objetivo y de los atributos que se utilizan para caracterizar esas clases y realizar su discernimiento.

   La clasificación es un proceso de división del espacio de atributos en compartimentos que "engloban" a cada una de las clases objetivo. Las redes neuronales y los árboles de decisión dividen el espacio de atributos mediante una serie de funciones discriminantes –en las redes neuronales se conocen como *hiper-planos* y en los árboles de decisión como *reglas de decisión* o *cortes*. En las dos técnicas mencionadas hay que sumar las funciones que definen las diferentes particiones.
4. Funcionamiento de una red neuronal

   Las redes neuronales artificiales también son conocidas como "aproximaciones de la función universal" por su capacidad para modelar relaciones matemáticas y transformaciones complejas de una manera heurística[1]. Existen muchas maneras de interpretación del comportamiento interno de una red neuronal (p. e., Carpenter, 1989; Kohonen, 1989). En un perceptrón multi-capa (MLP) como el que se muestra en la figura 2, que tiene 5 atributos, 6 nodos ocultos y 3 clases, una interpretación se basa en la relación entre los valores de los atributos y las clases objetivo, aplicando el método de retro-propagación (BP) (p.e. Pao, 1989; Tso y Mather, 2001).

1 Se denomina **heurística** a la capacidad de un sistema para realizar de forma inmediata innovaciones positivas para sus fines. (fuente: WIKIPEDIA. La enciclopedia libre).

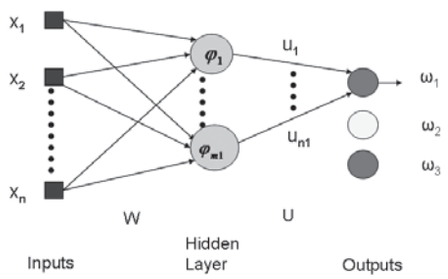

Figura 2. Arquitectura típica de una red neuronal. El diagrama muestra únicamente algunas de las conexiones entre los nodos de entrada, los de la capa oculta y los de salida (tomada de Gahegan, 1998).

La distancia algebraica entre cada muestra en el espacio de atributos y cada hiper-plano es evaluada por la red. En el caso de un espacio 3D esa distancia es:

$$ve = \frac{ax_0 + by_0 + cz_0 + d}{\sqrt{a^2 + b^2 + c^2}}$$

en donde (*a*, *b*, *c*, *d*) son los coeficientes del plano en el espacio (*x*, *y*, *z*) y la posición de cada atributo está definida por el punto P_0 (x_0, y_0, z_0) como se muestra en la figura 3. Los parámetros del hiper-plano (*a*, *b*, *c*, *d*) son los valores del vector de pesos *W* que convierte los valores de los nodos de entrada en valores de los nodos en la capa oculta.

La ecuación de la distancia algebraica es derivada de la ecuación de un plano en el cual los puntos cumplen: $ax + by + cz + d = 0$. (2)

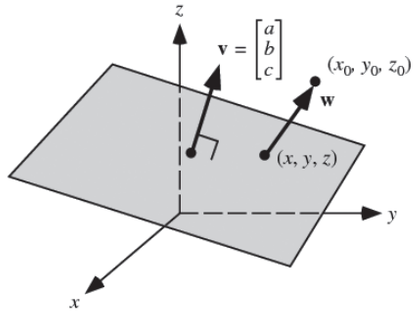

Figura 3. Ilustración de la distancia algebraica entre un punto y un plano.

La distancia euclidiana entre un punto P_0 y un plano está definida por la proyección del vector que une el punto con el plano, definido por [$x-x_0$, $y-y_0$, $x-x_0$], en el vector normal definido por [*a, b, c*]. La distancia algebraica incluye el signo, el cual es positivo si el punto está en el mismo lado del plano que el vector normal. Una medida de la pertenencia de una muestra se puede establecer usando distancias algebraicas. Un método consiste en determinar en cuál lado del hiper-plano está la muestra, lo cual es indicado por el signo de la distancia algebraica, ya sea *+ve* o *-ve*. Esta aproximación es la que se usa en los clasificadores lineales. Otro método es sumar las distancias algebraicas y luego usarlas para definir la clase. Este último es el esquema típico usado en las redes neuronales.

En el caso de dos atributos X_1 y X_2, la figura 4 muestra un conjunto posible de hiper-planos que pueden proporcionar un buen desempeño. Las distancias algebraicas resultantes en los nodos ocultos son combinadas mediante una función no lineal con la matriz *U*, que une la capa oculta con las salidas, para mapear cada clase objetivo. De esta manera, los hiper-planos fragmentan el espacio en un conjunto de regiones parciales de clasificación que pueden ser sumadas y asignadas a la clase apropiada. En resumen, la descripción de cada clase es la combinación ponderada de las funciones discriminantes. Por tanto, si las diferentes regiones que se muestran en la figura 4 se combinan se pueden representar las dos clases de interés (cuadrados y rombos).

Idealmente, para una muestra de la clase ω_1, los atributos X_j para $j = 1, \ldots n$ deberían producir una respuesta inequívoca {1, 0, 0} para ω_i en donde $i = 1, \ldots, 3$. Cualquier desviación de este ideal representa un 'error' de clasificación, que la red debe minimizar en el conjunto completo de los datos.

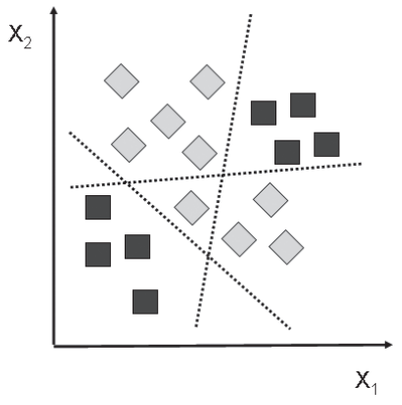

Figura 4. Hiper-planos de separación entre dos clases a partir de los atributos X_1 y X_2.

Los pesos entre las tres capas (matrices *W* y *U* en la figura 2) están determinados para que se obtenga el error mínimo en el mapeo del conjunto de las muestras de entrenamiento. La función de error se puede definir como:

$$\sum_{i=1}^{3}(O_i - A_i)^2 \qquad (3)$$

donde O_i para i = 1, ..., 3 representa la salida *deseada* de la red –por ejemplo {1, 0, 0} para la clase ω_1– y A_i representa la salida *real* –por ejemplo {0.9, 0.2, 0.3}– para una muestra de entrenamiento. El error se usa para modificar los valores de *W* y *U* y reducir la distancia entre cada muestra y el hiper-plano. Usando el criterio de error que se describió antes se busca obtener un mínimo global para todas las muestras y clases. La posibilidad de encontrar ese mínimo global depende de las condiciones iniciales de la búsqueda (los pesos en *W* y *U* se definen usualmente de manera aleatoria) y la complejidad del espacio de error (la presencia de muchos mínimos locales).

El uso de la distancia algebraica en redes neuronales significa que ellas producen una superficie continua que se puede interpretar como una clasificación "blanda" en la cual cada muestra de entrenamiento tiene una probabilidad finita de pertenecer a cada clase, de una manera similar a las probabilidades finales producidas por MLC. La salida debe tener una etiqueta de clase discreta, que usualmente se asigna mediante la selección del valor de salida dominante.

Observe que en esta configuración estándar la red neuronal no es capaz de diferenciar en cuáles casos el valor dominante es correcto y en cuáles no. En consecuencia, la red se penaliza en función de cualquier valor residual (por ejemplo, los valores 0.2 y 0.3 del ejemplo anterior). Teniendo en cuenta que las clases no son completamente separables, ese castigo puede ser poco razonable y, al aplicarse, causa oscilación en el posicionamiento de los hiper-planos a medida que la red trata de ajustar el valor en el nodo de salida que corresponda. Una explicación matemática detallada del funcionamiento del algoritmo de retro-propagación se puede encontrar en Lizarazo, *et al.* (2005).

5. Funcionamiento de un árbol de decisión

Los árboles de decisión descomponen el espacio de atributos en P conjuntos disjuntos, K_r, r = 1, ... P, usando reglas o cortes de decisión que son ortogonales a los ejes de los atributos (ejes paralelos). Por consiguiente, una regla de decisión puede ser descrita como una expresión simple de la forma $X_i \geq T_i$, o $X_i < T_i$, donde T_i es algún umbral, lo cual hace que esta operación sea fácil de entender por el usuario. La figura 5 muestra un árbol de decisión junto con las reglas que ocurren en cada nodo. Usando una muestra de datos que tienen X_i atributos y cuya clase real se conoce, los árboles de decisión seleccionan los atributos más apropiados para realizar la partición del espacio n-dimensional y producir una clasificación cercana a la real. La figura 5 muestra un árbol de decisión 'normal' (Quinlan, 1993) que solamente usa valores de atributos. En la figura 5, cada nodo toma una decisión binaria basada en uno (ó más) valores de atributos. Los atributos pueden ocurrir en cualquier parte del árbol y las hojas (nodos extremos) realizan las decisiones finales para obtener las clases de interés.

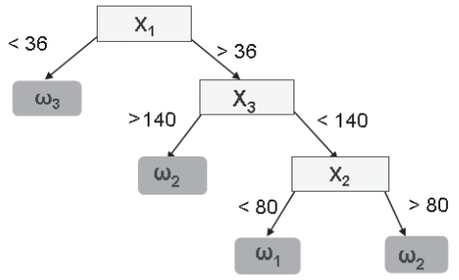

Figura 5. Funcionamiento de un árbol de decisión simple para obtener tres clases en función de tres atributos.

La clasificación de una región en un árbol de decisión está determinada por la trayectoria descrita desde la raíz del árbol hasta la hoja. Por tanto, cada regla refina progresivamente la clasificación de una manera jerárquica. El árbol refleja la complejidad del problema de clasificación: si la separación es problemática hay que usar muchas reglas, si ella es simple unas pocas son suficientes, de manera que la profundidad del árbol depende de la naturaleza del problema. De igual forma, cada regla

solamente considera aquellos atributos que ayudan a resolver el problema de clasificación y los otros son simplemente ignorados. El árbol se construye iniciando desde la raíz y evaluando cada uno de los atributos para determinar cuál de ellos es el que mejor divide los datos en dos conjuntos disjuntos. El mejor atributo es aquel que produce el menor número de muestras mal clasificadas. Luego de este paso, el árbol ya está compuesto de dos ramas y un nodo raíz. El proceso se repite por cada rama, realizando la partición de las muestras de manera tal que se puedan clasificar todos los datos.

Los cortes de decisión se hacen de manera paralela a los ejes, como se muestra en la figura 6, y se realizan en una secuencia que tiene implicaciones para la generalización y la optimización, como se verá más adelante. Algunos árboles de decisión tales como el Clasificador Oblicuo OC1 (Murthy, *et al.*, 1994) pueden usar una regla de partición no ortogonal (oblicua), que realiza modificaciones a los cortes ortogonales en un intento de mejorar el desempeño de la regla.

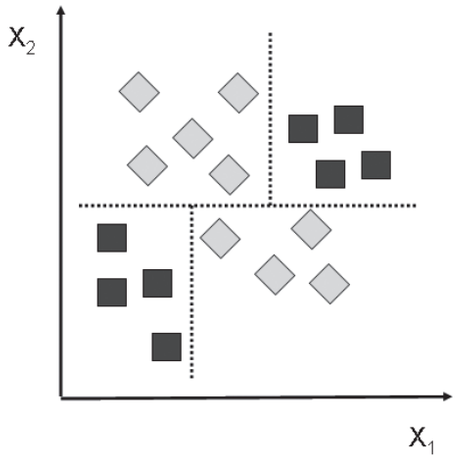

Figura 6. Participación del espacio de atributos usando un árbol de decisión que usa cortes paralelos a los ejes.

6. Espacio de atributos y supuestos estadísticos

El espacio de atributos requerido por un árbol de decisión, una red neuronal o un clasificador MLC es simplemente un espacio de dimensionalidad *n*, en el cual *n* es el número de atributos que se están utilizando. Cada atributo suministra una dimensión única. Las dos técnicas de IA que se analizan en este artículo pueden ser consideradas como "independientes de la distribución de los datos", en el sentido en que ellas no parten de ningún supuesto estadístico sobre la forma de distribución de los datos de entrenamiento en el espacio de atributos. De hecho, las suposiciones estadísticas son reducidas al mínimo y la única que se hace es que la distancia euclidiana en el espacio de atributos representa el grado de similaridad entre las clases: las distancias pequeñas indican una gran semejanza y las distancias grandes una gran diferencia.

Las muestras de entrenamiento se consideran independientes entre sí. Cuando se usan redes neuronales es usual que se escalen (normalicen) los atributos, de manera que el rango de valores de cada dominio sea consistente, usualmente esto significa que $0 \leq a \leq 1$. Esto es necesario, puesto que los atributos se usan de manera combinada y por tanto deben ser comparables numéricamente. Igualmente, ello ayuda a evitar errores de redondeo. Ese problema no ocurre con los árboles de decisión ya que cada regla de partición se formula para un solo atributo. La única comparación requerida está relacionada con el desempeño de las diferentes decisiones que podrían hacerse para cada tipo de atributo.

Cuando se construye inicialmente el espacio de atributos no se establece ninguna dependencia que se supone puede existir de manera natural (p. e., la correlación entre altura y temperatura o entre el aspecto del terreno y la precipitación). El peso de la evidencia (es decir, la contribución relativa de cada atributo para identificar cada clase) se aprende durante la fase de entrenamiento. Los árboles de decisión asignan pesos de manera implícita a la evidencia ordenando las reglas de acuerdo con el mejoramiento que la partición asociada le brinde al desempeño general de la clasificación. Los atributos que tienen una capacidad baja de clasificación simplemente son ignorados o son relegados a la parte más profunda del árbol.

A diferencia de los clasificadores paramétricos, las redes neuronales y los árboles de decisión pueden funcionar en cualquier tipo de dato estadístico, es decir que los atributos individuales pueden representar una mezcla de tipos de datos nomina-

les, ordinales y cuantitativos (intervalo y proporción). Sin embargo, para ello se requiere que los atributos nominales sean enumerados de manera que puedan ser "mapeados" en un eje y ello puede ocasionar algunos problemas de ordenamiento que se muestran en la figura 7. Considere el caso en el cual algunos valores nominales {A, B, C, D, E, F} representan alguna condición como cobertura de suelo, tres de los cuales {A, B, C} están asociados con un uso del suelo específico, representado por los cuadrados de la figura 7. Dependiendo del ordenamiento de dichos valores, el número de fronteras de decisión requeridas puede variar entre uno (1) y tres (3).

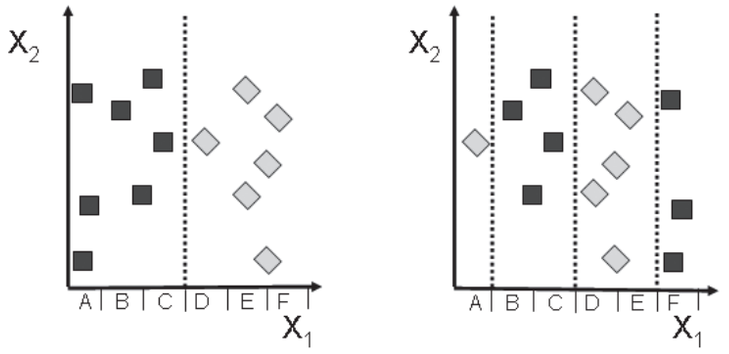

Figura 7. Dos ordenamientos diferentes para un conjunto de atributos nominales {A, B, C, D, E, F, G} en un problema de dos clases (tomada de Gahegan, 1998).

Cuando se usan redes neuronales, este problema de asignación se complica debido a que se debe determinar cuál es el número correcto de hiperplanos. Con los árboles de decisión el problema se manifiesta de manera más sutil; si se separan las diferentes coberturas, como se indica en la parte derecha de la figura 7, es probable que su aporte a la clasificación sea reducido o incluso insignificante. Ello significa que si los datos nominales muestran una agrupación consistente para todas las clases objetivo, ellos deberían ser unidos u ordenados de manera apropiada, para simplificar la tarea de clasificación.

7. Entrenamiento y generalización

La ubicación de las fronteras de los "compartimentos" de cada clase que son apropiadas constituye un asunto problemático que se puede comprender mejor usando como referencia los conceptos de *máxima* generalización y de *mínima* generalización. En la figura 8(a-b) se muestran los dos casos mencionados para dos atributos y dos clases. La máxima generalización toma su decisión con base en el lado de la partición en el cual cae la muestra. Por ejemplo: *si el valor de X es menor que T, entonces pertenece a la clase w_1, de lo contrario pertenece a la clase w_2*. En otras palabras, los atributos no tienen fronteras definidas. La mínima generalización usa los valores de atributos que localizan las muestras en un paralelepípedo en el espacio de atributos e ignora los espacios entre los paralelepípedos. Por ejemplo: *si el valor de X está dentro de T_1 y T_2, entonces pertenece a la clase w_1, de lo contrario su clase es desconocida.* En otras palabras, los atributos tienen fronteras definidas.

La generalización máxima es apropiada si existe una gran confianza en que los datos de entrenamiento son representativos y se espera que, en el futuro, todos los datos caigan en alguna de las clases definidas. Por el contrario, la generalización mínima es mejor en situaciones en las cuales los datos de entrenamiento representan de manera parcial el mundo real. Por ejemplo, en la figura 8(b) se muestra un ejemplo de generalización mínima introduciendo el concepto de una clase desconocida o de fondo que existe en zonas del espacio de atributos en las cuales no existen muestras de entrenamiento.

En cualquiera de las dos formas de generalización puede ocurrir que algunas zonas del espacio de atributos no contengan datos de entrenamiento

y que los datos nuevos que caen en ellas sean asignados a una clase específica. Por ejemplo, considere una muestra que está a cierta distancia de las clases identificadas como el triángulo de la figura 8.

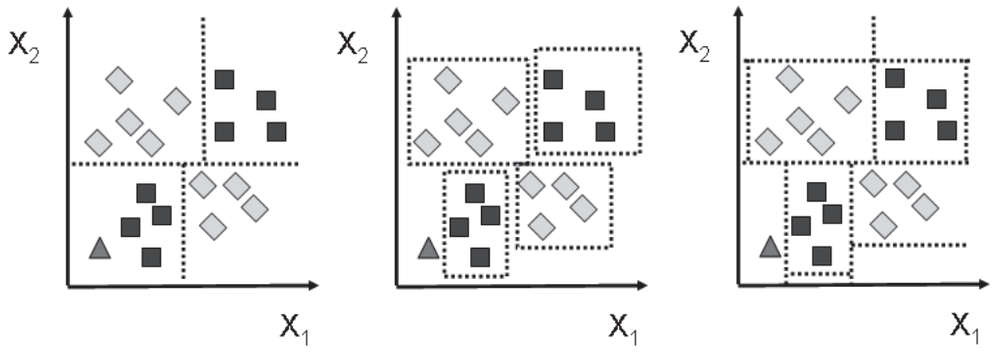

Figura 8. Generalización máxima (a), generalización mínima (b) y generalización híbrida (c). El triángulo representa una muestra no conocida que tiene que ser clasificada. En el caso (a) ella se asigna a la clase 'cuadrado' debido a que cae dentro de la frontera de decisión, pero no es clasificada en los casos (b) y (c) porque es diferente de cualquier cuadrado encontrado durante el entrenamiento.

Las redes neuronales posicionan los hiperplanos de una manera libre dentro de un espacio de atributos continuo, pero la posición final está determinada por la búsqueda de la minimización del error global mínimo como se indicó en el numeral 4. Usualmente, ellas utilizan la generalización máxima debido a que los hiper-planos se posicionan para maximizar la separación entre las clases. Puede ocurrir sobre-entrenamiento cuando la red se enfoca en la minimización de errores que no se deben a tendencias significativas en los datos sino a desviaciones individuales de algunos atributos, lo cual significa que se tiende a la generalización mínima. Este problema es ocasionado por la introducción de muchos hiper-planos con la esperanza de reducir el error. El sobre-entrenamiento solamente se detecta mediante la evaluación el desempeño del clasificador en un conjunto de validación.

En contraste, los árboles de decisión pueden ser configurados para seguir alguna de las dos aproximaciones de generalización. Dado un conjunto finito y ordenado de muestras de entrenamiento $\{v_1, v_2, \ldots v_m\}$, una frontera de decisión (v_1) puede ser colocada en cualquier parte entre v_i y v_{i+1}. Las fronteras de decisión cercanas o coincidentes con v_i causarán generalización mínima de la clase a la cual la división v_1 ayuda a definir.

La generalización mínima en los árboles de decisión usualmente emplea los hiper-rectángulos de menor perímetro como medio para definir las fronteras y por tanto está expuesta a los mismos problemas de sobre-entrenamiento indicados atrás.

En ocasiones, la mejor solución es un compromiso entre la generalización mínima y máxima que ofrece un balance entre las características de las dos opciones. Existe una variedad de técnicas estadísticas alternativas que pueden emplear para ello (Devijver y Kittler, 1982). Una posibilidad, que se muestra en la figura 8(c), es usar la generalización máxima para construir un árbol de decisión preliminar y luego examinar los valores de los atributos de cada hoja e introducir las fronteras basadas en los datos de entrenamiento.

Es clara la importancia de ir ajustando las habilidades de generalización del clasificador a los patrones observados dentro de los datos. Como no existe un mecanismo fijo que permita lograr ello, existe la necesidad de experimentación y validación antes del paso final de la clasificación.

8. Aspectos prácticos

En esta sección se mencionan elementos operacionales que hay que tener en cuenta cuando se usan estos clasificadores. Ellos se basan en las recomendaciones de Gahegan y West (1998), al igual

que en la experiencia ganada por el autor en su aplicación en un proyecto de investigación realizado recientemente. Teniendo en cuenta la complejidad de configuración de los dos clasificadores no se ha realizado una comparación directa entre los clasificadores. Los lectores interesados pueden revisar los parámetros de configuración de los dos métodos en otros documentos (Lizarazo, 2005; Lizarazo *et al.*, 2006).

8.1 Niveles de clasificación

Las dos técnicas señaladas permiten tratar de manera fácil la idea de subclases (una clase compuesta de dos o más regiones disjuntas dentro del espacio de atributos). Cuando se usan árboles de decisión, la existencia de subclases no constituye ningún problema debido a que dicha técnica usa tantas reglas como se requieran para realizar la partición del espacio de atributos. En redes neuronales la situación es más compleja debido a que el aislamiento entre subclases requiere la disponibilidad de hiper-planos adicionales con los cuales describir las regiones separadas al igual que la capacidad de posicionar los hiper-planos alrededor de las áreas problemáticas en el espacio de atributos.

Las clasificaciones binarias, por ejemplo *vegetación* y *no vegetación,* a partir de datos de bandas espectrales, producen generalmente un gran número de subclases, una por cada unidad de cobertura asociada a la clase "restante" definida como *no vegetación.* La selección del método de generalización se convierte en un asunto importante debido a que la exactitud final depende de la delineación de cada clase individual. Algunos investigadores han sugerido que los árboles de decisión tienen un mejor desempeño que las redes neuronales en este tipo de problemas (Evans, *et al.*, 1998). Esto puede deberse a dos razones: (i) los árboles de decisión permiten que el usuario seleccione el tipo de generalización, y (ii) es difícil seleccionar el número correcto de hiper-planos cuando se usan redes neuronales, especialmente si los datos de entrenamiento no ofrecen pistas sobre el número de clases que deben agruparse en el conjunto "restante".

8.2 Arquitectura

Dentro de una red neuronal existe un gran número de decisiones arquitectónicas que afectan de manera dramática su desempeño. La selección del número de hiper-planos (es decir el número de nodos en la capa oculta) es crítico para el éxito. Pocos hiper-planos no permiten la separación completa de las clases, mientras que demasiados hiper-planos ocasionan que la red se sobre-entrene (ver sección 7), con el efecto de una generalización muy pobre. El problema que enfrenta el usuario es que no conoce, con anterioridad a su aplicación, cuántos hiper-planos son requeridos. Eso implica que es necesario probar de manera exhaustiva varias configuraciones de red para encontrar aquella que tenga el mejor desempeño durante la validación. Esto, desde luego, implica una carga computacional muy fuerte y representa una carga de trabajo adicional para el usuario.

Los trabajos más recientes intentan resolver esta debilidad por dos medios principales: (i) proponiendo arquitecturas que se adapten mejor a cada problema, (ii) identificando, por cada par de clases a separar, cuántos nodos son necesarios y cómo se pueden usar en el momento en que se requiera (German, *et al.*, 1998) y (iii) utilizando clasificación de lógica difusa (Tso y Mather, 2001).

Los árboles de decisión también sufren de ese mismo problema aunque sus efectos son menos severos. Los criterios utilizados para determinar cuando una clase está adecuadamente descrita han sido objeto de varias investigaciones. Los umbrales utilizados para cada decisión se pueden seleccionar usando medidas de entropía mínima o de error mínimo. El método de entropía mínimo fue propuesto originalmente por Hunt (1966) y usado por Quinlan (1993) en C4.5. Ese método utiliza el número mínimo de bits para describir cada decisión en un nodo del árbol, tomando como base la frecuencia de cada clase en el nodo de interés. De manera alternativa, se podría utilizar una función de error basada en estadística o en distancia algebraica. Con la entropía mínima, se requiere un criterio de parada que está basado en la cantidad de información ganada por una regla (la proporción de ganancia). Este umbral es definido por el usuario, usualmente mediante la experimentación. La ventaja de usar árboles de decisión es que este refinamiento es simplemente la continuación del proceso de clasificación anterior y no el inicio de

un proceso nuevo como ocurre en el caso de las redes neuronales.

8.3 Búsqueda del error mínimo

Las redes neuronales examinan al mismo tiempo todas las dimensiones en su búsqueda de una solución óptima y para ello usa rutinas sofisticadas que examinan la tendencia del error e intentan encontrar el punto donde ocurre el menor error total, tal como se describió en la sección 4. Estas rutinas pueden ser muy sensibles al gradiente y a la dirección de cambio (Richards y Jia, 1999). Sin embargo, no es posible garantizar que el mínimo global se ha encontrado a menos que la superficie se haya investigado de manera exhaustiva utilizando un método analítico.

En contraste, los árboles de decisión se mueven hacia una solución de una manera incremental, tomando decisiones independientes en cada dimensión de los datos y en cada iteración. El orden en el cual se generan las reglas de decisión determina cómo es la búsqueda de la superficie de error y ella está definida únicamente por los datos. Por ello, los árboles de decisión no usan el concepto de global mínimo, aunque las implementaciones más avanzadas realizan la planeación de las iteraciones siguientes.

Es claro que ninguna de las dos técnicas asegura una solución óptima, ni siquiera en condiciones ideales en las cuales se conoce el número correcto de hiper-planos o los criterios de parada. Por consiguiente, el desempeño del clasificador solamente se puede evaluar de manera confiable usando datos de validación que sean independientes estadísticamente. Cualquier intento para predecir el desempeño basado en la capacidad para clasificar las muestras de entrenamiento conduce a sobre-estimar la exactitud, pues ignora el riesgo del sobre-entrenamiento (sección 7).

8.4 Uso de recursos computacionales

Puesto que los dos métodos se usan en conjuntos de datos de imágenes, en los cuales el número de capas de entrada y/o las clases de salida es alto, es pertinente examinar cómo es el desempeño en términos de tiempo de computador a medida que crecen los atributos y el espacio de clases.

Los árboles de decisión deben determinar cuál es el atributo más importante y cuáles son los umbrales a usar para realizar la partición en cada nodo. Los datos de atributos deben ordenarse y todos los posibles umbrales deben ser evaluados en función de su desempeño. La complejidad final depende de los datos, aunque la experiencia con problemas reales muestra que el ordenamiento es el factor que más incide. Sin embargo, los tiempos de creación de un árbol se miden en términos de segundos y no de minutos. En general, la complejidad del problema aumenta de manera lineal con el número de dimensiones del espacio de atributos, debido a la independencia de las reglas como se describió en la sección anterior.

En las redes neuronales, la complejidad es mayor. El mecanismo de búsqueda es global y la búsqueda se realiza en todos los atributos y clases de manera simultánea. Aunque podría esperarse una relación tipo función cuadrada entre los atributos y la demanda computacional a medida que se incrementa la dimensionalidad, ello no es cierto. Cada dimensión adicional atrae un nodo de entrada extra y por tanto se requieren q conexiones de peso adicionales en la capa oculta (donde q es el número de nodos en la capa oculta). Esto complica la tarea de entrenamiento en términos de la matriz W. Sin embargo, si el número de los hiper-planos requeridos no se incrementa, entonces el incremento en desempeño es nuevamente lineal. Es únicamente cuando los nodos de la capa oculta se incrementan que la complejidad aumenta de forma alarmante, pues ello afecta tanto a W como a U. Pero incluso en ese caso, el uso de procedimientos sofisticados de búsqueda reduce el impacto del cambio, de manera que la complejidad cae en algún punto entre una relación lineal y una función cuadrática relacionada con el número de atributos.

En la práctica, muchos investigadores han encontrado que, en el caso de los árboles de decisión, la escalabilidad no causa problemas. En las redes neuronales se requiere un gran esfuerzo para definir la arquitectura de la red. Esto no se debe tanto a la complejidad computacional sino a la 'saturación' de los nodos de la capa oculta, un problema que ocurre cuando existe un gran número de nodos de entrada que deben conectarse. Entre más pesos existan más difícil es asegurar que únicamente se trabaje con la que es evidencia útil. El

entrenamiento también se vuelve muy demorado, particularmente si la red se inicializa con pesos aleatorios.

9. Conclusiones

Este artículo ha presentado aspectos relacionados con la selección y diseño de dos técnicas de aprendizaje de máquina que son útiles para la clasificación de imágenes digitales. En particular se han revisado temas como la generalización y aspectos de funcionamiento como arquitectura, búsqueda de la mejor solución y evaluación de la complejidad. El propósito central ha sido enfatizar elementos que deben ser tenidos en cuenta al utilizar cualquiera de dichos clasificadores. En resumen, no existe evidencia empírica suficiente para recomendar de manera absoluta el uso de los algoritmos analizados sobre los métodos convencionales. En cada caso particular, se debe realizar la selección tomando como base las características específicas del problema a resolver y el análisis de los datos disponibles.

Los árboles de decisión y las redes neuronales no son una panacea para resolver todos los problemas de los clasificadores paramétricos, ni su uso garantiza por sí mismo un aumento de la exactitud temática. Ellos permiten superar algunas limitaciones de los métodos tradicionales pero a costa de enfrentarse a nuevas dificultades como el desempeño computacional, el comportamiento de la generalización y la configuración de los parámetros adecuados. Las redes neuronales son, hasta cierto punto, cajas negras que realizan clasificaciones buenas sin que el usuario pueda establecer cuál es la contribución de los diferentes atributos. Por ello es que los árboles de decisión pueden ser considerados inicialmente como la mejor alternativa, debido a que son más simples de configurar y a que el usuario puede conocer las reglas de decisión usadas. Sin embargo, es pertinente indicar que los dos algoritmos requieren una intervención cuidadosa de los usuarios, en particular para seleccionar los parámetros operacionales adecuados. Esto significa que hay necesidad de experimentación y validación mediante datos independientes.

La búsqueda global empleada por las redes neuronales puede ayudar a obtener una mejor clasificación, sobre todo en problemas de gran complejidad, a costa de experimentar usando diferentes valores en los parámetros de la red hasta encontrar la configuración adecuada. Como ejemplo, hay que indicar que en la clasificación de la cobertura del suelo de la zona de estudio usando como datos de entrada las bandas originales de una imagen QuickBird multi-espectral complementadas con una imagen de bordes de Sobel las redes neuronales (BP) alcanzaron una exactitud global del 79%, luego de 26 intentos por encontrar una configuración apropiada. Dicha exactitud final estuvo muy cerca a la obtenida mediante MLC (82%) y fue ligeramente superior a la obtenida mediante árboles de decisión (72%) en su primer intento.[2] En la clasificación del uso del suelo realizada en dicho estudio, usando como bandas de entrada las métricas de proporción de cada cobertura y de diversidad de cobertura, el algoritmo BP (ANN) permitió obtener una exactitud temática del 64% mientras que el algoritmo DT alcanzó una exactitud global del 74% y el algoritmo MLC solamente llegó al 42%. Para fines comparativos, las figuras 9, 10 y 11 muestran los resultados gráficos obtenidos por los diferentes métodos.

Por otro lado, es pertinente anotar que un determinado algoritmo puede tener implementación diferente en diferentes programas de procesamiento digital y ello significa que su desempeño –así como las opciones disponibles para los usuarios– pueden variar enormemente. A manera de ejemplo, Lizarazo, *et al.* (2006) reportan que, usando los mismos parámetros de entrenamiento, el algoritmo BP de Idrisi[3] utilizado para clasificar la cobertura del suelo urbano en la zona de estudio necesitó solamente de 3 minutos para producir una clasificación cuya exactitud con los datos de entrenamiento fue del 86% mientras que el algoritmo BP de ENVI[4] requirió de 5 horas, al cabo de las cuales entregó una exactitud de entrenamiento del 79%.

2 Un reporte detallado de los parámetros utilizados y los resultados obtenidos en la zona de estudio puede verse en Lizarazo *et al.* (2006).

3 IDRISI es un software académico producido por Clark Labs (Clark University, MA).

4 ENVI es un software comercial producido por Research Systems Inc.

La dificultad de decidir cuál es el mejor clasificador y cuáles son los parámetros apropiados han conducido a que varios investigadores adopten como solución lo que se conoce como *aproximación a problemas no resueltos* (Dietterich, 1997) (Gahegan, 1999) que puede plantearse así: (1) pruebe inicialmente con algoritmos de aprendizaje convencionales como MLC, (2) use métodos alternativos de clasificación como DT y ANN-BP y evalúe su desempeño, (3) seleccione nuevas muestras de entrenamiento y aplique el algoritmo cuyo desempeño inicial fue mejor y (4) si la exactitud final no cumple los requerimientos, ensaye la combinación de los resultados de diferentes clasificadores como DT y ANN. Esta aproximación empírica muestra que todavía existe mucho terreno por recorrer para encontrar herramientas de clasificación robustas para el procesamiento de imágenes de sensores remotos.

Figura 10. Clasificación de la cobertura del suelo usando el algoritmo de máxima probabilidad (MLC). Este algoritmo permite obtener una delineación aceptable de las coberturas existentes pero es evidente la fragmentación de las construcciones.

Figura 9. Vista parcial de la zona de estudio usando una composición RGB321 de una imagen QuickBird multiespectral (febrero de 2005). En el centro de la imagen se puede observar el parque Simón Bolívar con sus principales clases de cobertura.

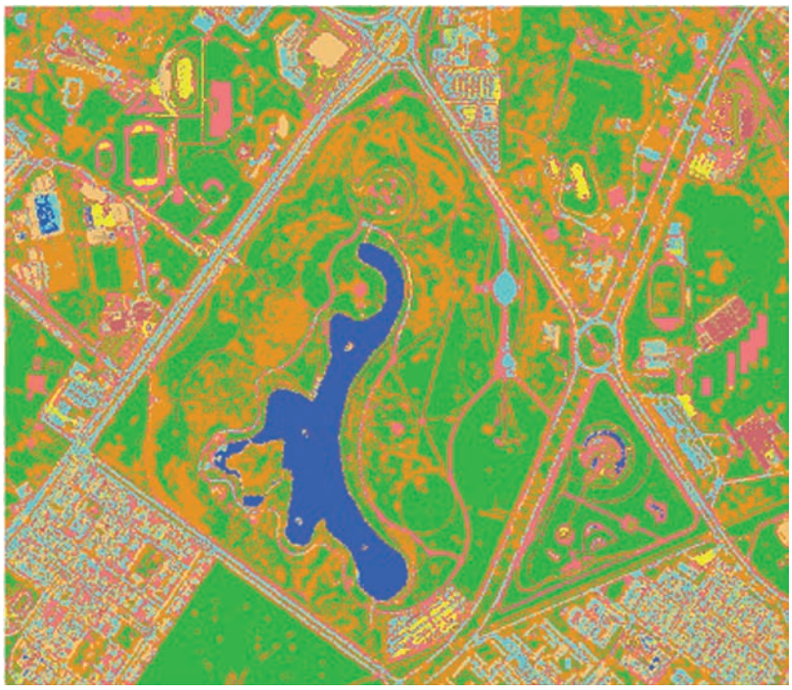

Figura 11. Clasificación de la cobertura del suelo usando el algoritmo de retro-propagación (BP-ANN). Observe que la apariencia visual de las clases es mejor que la obtenida mediante el método MLC.

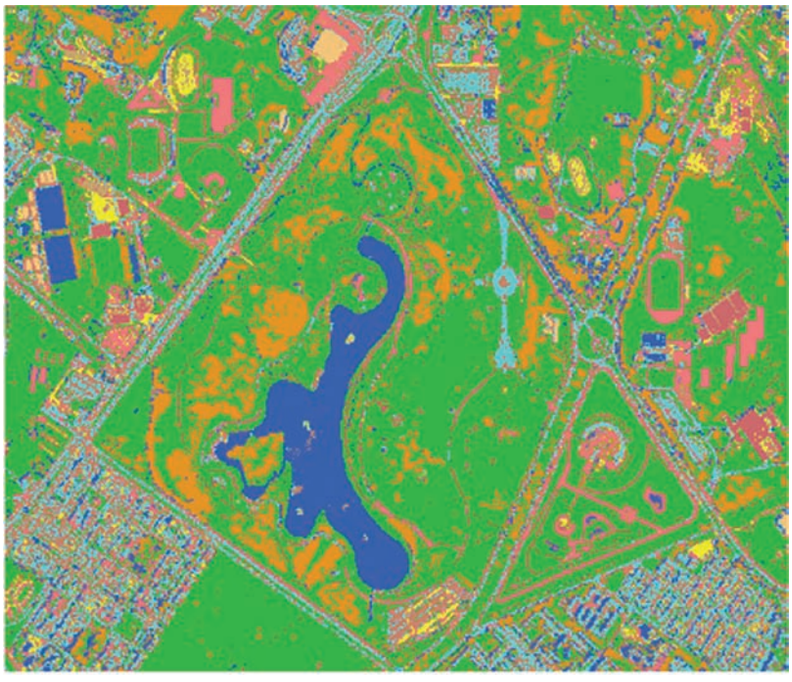

Figura 12. Clasificación de la cobertura del suelo usando el algoritmo de árboles de decisión (DT). Observe que la apariencia visual de las clases es inferior a la obtenida mediante los métodos

Agradecimientos

Esta publicación es uno de los productos del proyecto de investigación "Clasificación digital de la cobertura y uso del suelo urbano usando métodos no convencionales" que fue realizado gracias al apoyo del Centro de Investigaciones y Desarrollo Científico (CIDC) de la Universidad Distrital. Los ingenieros catastrales Ricardo Cuitiva y Samuel Mesa participaron en las pruebas de clasificación mediante redes neuronales artificiales.

Referencias bibliográficas

ASPRS (1997). *Manual of Photographic Interpretation.* American Society for Photogrammetry and Remote Sensing, Bethesda, USA.

Carpenter, G. (1989). Neural network models for pattern recognition. *Neural Networks*, Vol. 2.

Del Frate F., Schiavon, G., Solimini, C. (2004). Application of neural networks algorithms to QuickBird imagery for classification and change detection of urban area. *Proceedings of International Geoscience And Remote Sensing Symposium*. Anchorage, Alaska.

Devijver, P. A. y Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice-Hall International Inc, London.

Evans, F. H., Kiiveri, H. T., West, G. y Gahegan, M. (1998). Mapping salinity using decision trees and conditional probabilistic networks, *Proc. Second IEEE International Conference on Intelligent Processing Systems*, Gold Coast, Queensland, Australia.

Fitzgerald, R. W. y Lees, B. G. (1994). Assessing the classification accuracy of multisource remote sensing data. *Remote Sensing of the Environmen*t, Vol. 47.

Gahegan, M. y West G., 1998. The Classification of Complex Geographic Datasets: An Operational Comparison of Artificial Neural Network and Decision Tree Classifiers. *Geocomputation Conference Proceedings,* University of Bristol, United Kingdom.

Gahegan, M., German, G. y West, G. (1998). Some solutions to neural network configuration problems for the classification of complex geographic datasets. *Geographical Systems,* Vol 15, No. 11.

German, G., Gahegan, M. y West, G. (1999). Statistical and AI Techniques in GIS Classification: A Comparison. *Proc. SIRC Conference*, The University of Otago, New Zelanda.

Hunt, E. B., Marin, J. y Stone, P. J. (1966). *Experiments in Induction*. Academic Press, New York, USA.

Kanellopoulos, I. y Wilkinson, G. (1997). Strategies and best practice for neural network image classification. *International Journal of Remote Sensing*, Vol. 61, No, 4.

Kohonen, T. (1989) *Self organisation and associative memory* (3rd edition). Springer-Verlag, Berlin, Germany.

Jensen J.R. (2005) *Introductory Digital Image Processing – A Remote Sensing Perspective (*3rd Edition). Prentice Hall, New Jersey, USA.

Lizarazo I., Mesa S. y Cuitiva R. (2005). Clasificación de Imágenes Usando Redes Neuronales: Bases Matemáticas. *Revista Científica No. 7.* Centro de Investigaciones y Desarrollo Científico (CIDC), Universidad Distrital.

Lizarazo I., Mesa S. y Cuitiva R. (2006). Clasificación de la Cobertura y el Uso del Suelo Urbano Usando Algoritmos de Redes Neuronales. Presentado al Centro de Investigaciones y Desarrollo Científico (CIDC), Universidad Distrital, para publicación.

Lizarazo I. (2005) *Clasificación Digital del Uso del Suelo Urbano usando índices espaciales e imágenes de alta resolución espacial.* Manuscrito pendiente de publicación.

Murthy, S., Kasif, S. y Salzberg, S. (1994). A system for induction of oblique decision trees, *Journal of Artificial Intelligence Research*, vol. 2.

Pao, Y. H. (1989). *Adaptive pattern recognition and neural networks*. Addison-Wesley, Reading, Mass. USA.

Paola, J. D. y Schowengerdt, R. A. (1995). A detailed comparison of backpropagation neural networks and maximum likelihood classifiers for urban landuse classification. *IEEE transactions on Geoscience and Remote Sensing*, Vol. 33, No. 4, pp. 981-996.

Pesaresi, M. y Benediktsson J.A. (2000). Classification of Urban High-Resolution Satellite imagery Using Morphological and Neural Approaches. *Proc. IGARSS'00*.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California, USA.

Richards, J. A. y Jia X. (1999). Remote Sensing Digital Image Analysis. Springer-Verlag, Berlin, Germany.

Schiavon G., Del Frate F. y Solimini C. (2003). High Resolution Multi-Spectral Analysis of Urban Areas with QuickBird Imagery and Sinergy with ERS Data. *Proc. IGARSS'03*.

Tso B. y Mather P.M. (2001). *Classification Methods for Remotely Sensed Data.* Taylor & Francis, London, UK.

IVÁN LIZARAZO

Ingeniero civil, Universidad Industrial de Santander. Especialista en SIG, Universidad Distrital. MSc en Geographic Information Science, Universidad de Londres. Profesor del Proyecto Curricular Ingeniería Catastral y Geodesia. Integrante del Grupo de Investigación NIDE. Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Correro electrónico: ilizarazo@udistrital.edu.co