Minería de datos con información de contexto para la clasificación de imágenes satelitales

Data mining with context information for satellite image classification

Jesús A. González[1]
Leopoldo Altamirano[2]
Juan F. Robles[3]

Resumen

En este artículo se presenta un esquema de clasificación multi-modelos para imágenes satelitales apoyado con información de contexto con el que se mejora la precisión de una pre-clasificación obtenida con algoritmos paramétricos. El nuevo esquema utiliza una red semántica como representación de conocimiento que almacena patrones creados con un ensamble de árboles de decisión (alimentado con características espectrales, de textura y geométricas para describir a las regiones de interés) y por otro lado patrones espaciales creados a partir de una representación basada en grafos (con información de contexto a partir de relaciones espaciales entre las regiones de interés). Los resultados experimentales muestran que el esquema de clasificación propuesto mejora la precisión de la pre-clasificación de los algoritmos paramétricos al utilizar información de contexto.

Palabras-clave: percepción remota; mapas temáticos; minería de datos; clasificación; información de contexto.

Abstract

This paper presents a multi-model classification schema for satellite images supported with context information to enhance the accuracy

1 Instituto Nacional de Astrofísica, Óptica y Electrónica; e-mail: jagonzalez@inaoep.mx

2 Instituto Nacional de Astrofísica, Óptica y Electrónica; e-mail: robles@inaoep.mx

3 Instituto de Investigación y Desarrollo Tecnológico de la Armada de México; e-mail: jfroblesc@semar.gob.mx

of a pre-classification obtained with parametric algorithms. This new scheme uses a semantic network as knowledge representation that stores the patterns created with a decision tree ensemble (fed with spectral, texture and geometric descriptive characteristics to describe the regions of interest) and spatial patterns created with a graph-based representation (with context information obtained from spatial relations among regions of interest). Our experimental results show that the proposed classification scheme enhances the pre-classification accuracy obtained with parametric algorithms when we use context information.

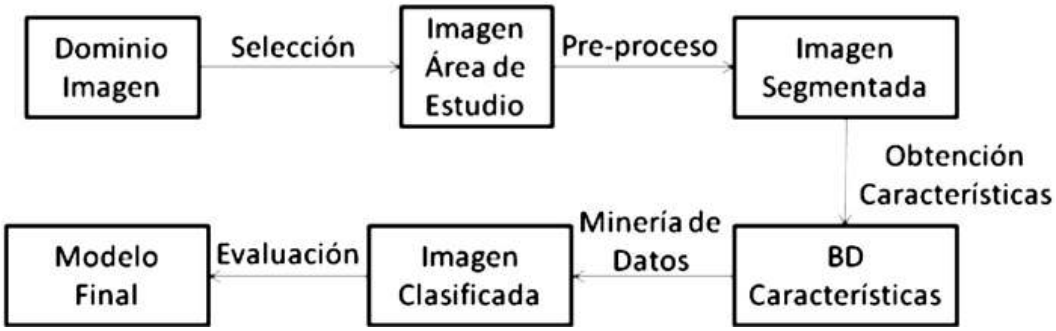
Key words: remote sensing, thematic maps, data mining, classification, context information.

Introducción

Recientemente se ha incrementado el uso de imágenes satelitales para el análisis del territorio en diversas aplicaciones. Tal es el caso de estudios para planear la protección y reforestación de lugares que han perdido gran parte de sus áreas verdes. En este artículo se presenta un método para la creación de mapas temáticos para el análisis de áreas de interés. Para lograr esta meta, en este trabajo se utiliza el proceso de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés), el cual se complementa con información de contexto para mejorar la precisión en la clasificación (HAN y KAMBER 2001). En la figura 1 se presenta el proceso KDD especializado en la clasificación de imágenes, incluyendo los pasos de selección de datos, pre-procesamiento (incluyendo la segmentación de la imagen), obtención de características, minería de datos y evaluación del modelo. Entre los pasos más difíciles al realizar este proceso (al trabajar con imágenes satelitales) se encuentran la segmentación y la clasificación de la imagen y es aquí en donde se ha realizado mucha investigación para mejorar lo más posible las herramientas con las que se realizan.

En algunos trabajos se ha utilizado información de contexto para mejorar

Figura 1. Proceso KDD para la clasificación de imágenes



la clasificación de imágenes satelitales. Por ejemplo, en (KUNZ et al., 1997) se utiliza una base de datos topográfica como apoyo al análisis de la imagen satelital y de esta manera ofrece una representación geométrica y una predicción semántica de los objetos que se pueden encontrar en la imagen de satélite. En (LIEDTKE et al., 1997) se crea el sistema AIDA (A System for the Knowledge Based Interpretation of Remote Sensing Data) en el que se utiliza una red semántica construida con información de un sistema de información geográfica para interpretar imágenes de satélite. En (BÜCKNER et al., 2001 y MULLER et al., 2003) se analizan imágenes satelitales utilizando el sistema geoAIDA, en el que la red semántica contiene información extraída de una base de datos topográfica de los objetos espaciales que se encuentran en el área geográfica que cubre la imagen satelital además de información extraída de esta misma imagen. Esta combinación de información se utiliza para mejorar la precisión de la clasificación de las imágenes satelitales. Sin embargo, una de las desventajas de estos enfoques es que requieren de información geográfica del área a clasificar y no se pueden utilizar (o no funcionan bien) si no se cuenta con dicha información.

El objetivo de este trabajo es crear un enfoque para la clasificación de imágenes satelitales que utilice información de contexto pero sin depender de información apriori de la región a clasificar (como una base de datos topográfica). En específico, mejoramos la precisión de clasificación en un segundo paso en el que se utiliza información espacial obtenida de las regiones de interés (relaciones espaciales entre las regiones de interés).

Materiales y Métodos

Dominio de aplicación

El dominio utilizado está formado por imágenes satelitales SPOT-5 (SPOT 2006) de los Estados de Veracruz y Campeche en México, con una resolución espacial de 10m, una cobertura aproximada de 60 x 60 km y con un nivel de procesamiento 2A con correcciones radiométricas y geométricas (en UTM WGS84 sin considerar puntos de control). Estas imágenes contienen cuatro bandas espectrales en las longitudes de onda "Verde: 0.5 – 0.59 μm", "Rojo: 0.61 – 0.68 μm", "Infrarrojo Cercano: 0.78 – 0.89 μm" e "Infrarrojo Medio: 1.58 – 1.75 μm". Las áreas de estudio tienen aproximadamente 3,600km^2. Las imágenes se dividieron en segmentos de 512 x 512 pixeles (24km^2) y en segmentos de 150 x 150 pixeles para un total de 30 segmentos de imágenes multi-espectrales en 4 bandas. Para poder hacer una evaluación cuantitativa de los experimentos se utilizaron imágenes sintéticas como se describe en la siguiente subsección.

Imágenes sintéticas

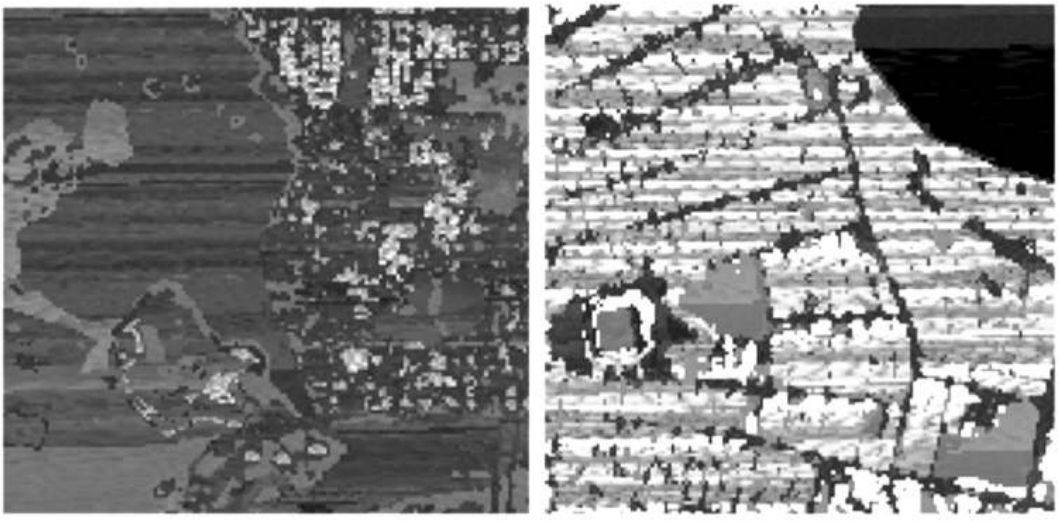
Realizar pruebas de clasificación con imágenes satelitales es un trabajo difícil porque para comprobar el desempeño del clasificador, también se debería realizar un estudio de campo para verificar la calidad y exactitud de las regiones clasificadas. Por lo anterior, en este trabajo se propone el uso de

una Imagen Multiespectral Sintética (IMS) para probar el desempeño de los clasificadores. La IMS se construye tomando las regiones generadas por algún método de segmentación o por algoritmos de clasificación paramétricos y se toman como regiones vacías, sin tomar en cuenta el valor espectral de cada píxel (i,j). En las posiciones (i,j) de cada píxel de la región vacía se asigna un valor espectral conocido (*ground truth*) tomando los datos de prueba en forma aleatoria para las cuatro bandas espectrales. De esta manera, la IMS estará formada por una cantidad conocida de regiones y píxeles con valores espectrales de los que estamos seguros que pertenecen a las diferentes clases de interés. Por otro lado, el utilizar regiones generadas (en cuanto a forma) por métodos de segmentación o algoritmos de clasificación paramétricos, asegura que la complejidad de las regiones en la IMS se asemeje a la distribución de los objetos en el mundo real. Las imágenes sintéticas tienen un tamaño de 150x150 píxeles (para disminuir el tiempo de procesamiento de clasificación al evaluar). La figura 2 muestra un ejemplo de las imágenes multiespectrales sintéticas IMS-1 e IMS-2. Las pruebas cuantitativas de los experimentos se realizaron con estas imágenes.

Clasificación Utilizando Información de Contexto

El proceso para clasificar imágenes satelitales que se utilizó sigue el proceso KDD general descrito en la figura 1. Este proceso adaptado a la aplicación se muestra en la figura 3, inicia con la selección de la imagen satelital multiespectral o selección del área de estudio contenida en dicha imagen. Posteriormente se hace una pre-clasificación utilizando algoritmos paramétricos y así obtener una

Figura 2. Imágenes multiespectrales sintéticas. a) IMS-1, b) IMS-2



a) b)

Figura 3. Proceso de clasificación multi modelos propuesto

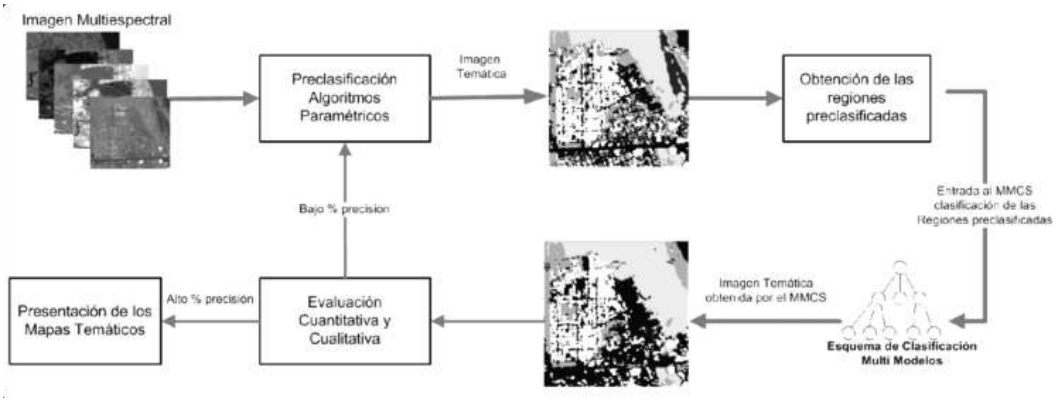


imagen temática. A partir de esa imagen temática se obtienen las regiones de interés clasificadas y que son utilizadas únicamente como esqueleto para generar una imagen sintética como se describe en la sección "Imágenes Sintéticas". Las IMS son entonces preclasificadas por un algoritmo paramétrico y estas regiones se dan como entrada al esquema de clasificación multi-modelos (MMCS), representado con una red semántica, y se obtiene como salida una imagen temática (re-clasificada). Posteriormente se evalúa la precisión de la clasificación obtenida y en caso de que ésta sea baja se regresa al paso de preclasificación con algoritmos paramétricos mientras que si se obtuvo un porcentaje alto de clasificación se presenta el mapa temático. A continuación se describe con más detalle los pasos del esquema de clasificación multi-modelos que consta principalmente de 3 etapas.

En la primera etapa se crea un ensamble de árboles de decisión (QUINLAN, 1996) para obtener reglas que serán utilizadas para clasificar nuevas imágenes considerando las mismas clases de la fase de entrenamiento. Cada árbol de decisión del ensamble se especializa en una de estas clases (puede haber más de un árbol especializado en la misma clase pero debe haber al menos uno para cada clase). La entrada a los árboles de decisión consiste en un conjunto de características descriptivas de las regiones de interés (ROI), en este caso se utilizan características espectrales: media, varianza, desviación estándar, máximo valor, mínimo valor, Índice Diferencial de Vegetación Normalizado ó NDVI (los primeros cinco para cada una de las cuatro bandas de la imagen), características de textura: entropía, anisotropía, energía, correlación, homogeneidad y contraste (para cada una de las cuatro bandas) y características geométricas: área, compatibilidad, excentricidad, convexidad, circularidad y diámetro. Las reglas generadas por los árboles (conocimiento para la clasificación) se almacenan en una red semántica. El segundo paso tiene como objetivo realizar un proceso de minería de datos para encontrar patrones espaciales entre las ROIs, regiones generadas

con algoritmos paramétricos como se describió anteriormente para las imágenes sintéticas. Este proceso se muestra en la figura 4. Como podemos apreciar se inicia con la base de datos que contiene las imágenes satelitales. Posteriormente se realiza el paso de preparación de datos y se obtiene una lista de ROIs (las mismas regiones encontradas por los algoritmos paramétricos) y para cada una de ellas una lista de sus ROIs adyacentes. Posteriormente se transforman los datos a una representación basada en grafos que después se dará como entrada al algoritmo Subdue (COOK and HOLDER, 1994, HOLDER et al., 2002), un algoritmo de minería de datos basado en grafos utilizado para trabajar con dominios estructurados. Es en este momento que se realiza el proceso de minería de datos espacial.

El resultado de esta fase consiste en un conjunto de patrones que asocian ROIs a través de relaciones espaciales topológicas (KOPERSKI et al., 1999, PECH et al., 2003), en este experimento sólo se trabaja con la relación topológica "adyacente", para trabajo futuro se añadirán más relaciones topológicas y también relaciones de distancia, como "cerca" y "lejos" y de dirección, como "norte" y "este". En la figura 5 se

Figura 4. Proceso de descubrimiento de conocimiento en bases de datos utilizando datos espaciales

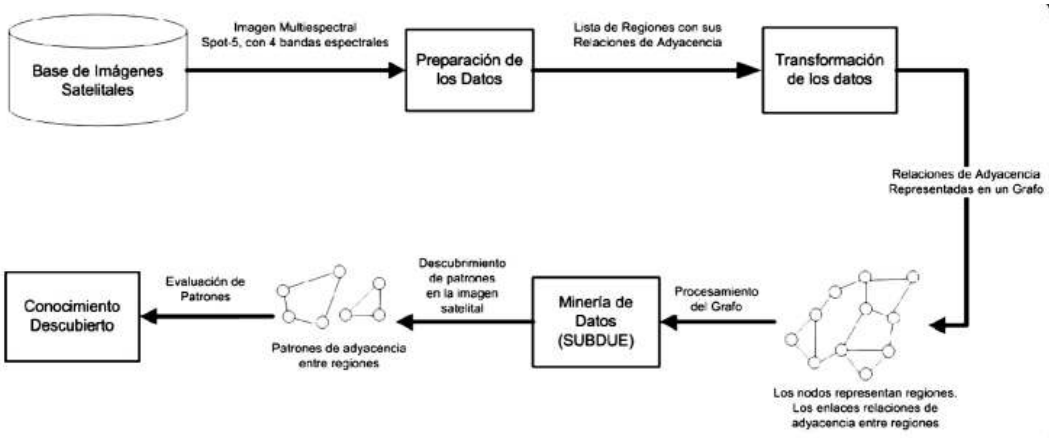


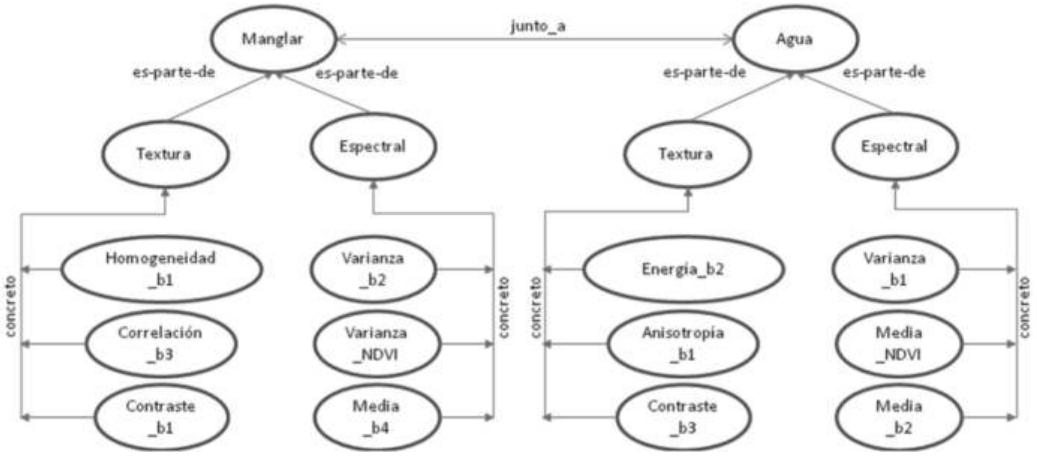
Figura 5. Representación basada en grafos de los patrones contextuales de las ROIs

muestran dos ejemplos de los patrones espaciales encontrados con Subdue para las clases “Manglar” y “Carretera”. El primer patrón indica que generalmente una ROI que pertenece a la clase manglar es adyacente a una región de “Suelo-Descubierto”, a otra de “Vegetación” y a otra de “Agua”.

En la tercera etapa de este proceso se crea una Red Semántica (SAGERER and NIEMANN, 1997) en la que se concentran los resultados de los primeros dos pasos y se utiliza para re-clasificar imágenes previamente clasificadas con algoritmos paramétricos y así mejorar la precisión de la clasificación. La figura 6 muestra una parte de la red semántica que describe las clases “manglar” y “agua”. Cada clase se describe con un patrón de los árboles de decisión con características espectrales de textura y geométricas (los nodos que llegan al nodo etiquetado con el nombre de la clase a través de un arco etiquetado como “es_parte_de”). Cada nodo de características tiene asociado un grupo de características (que varía entre clases) que describen (de forma general) los objetos que pertenecen a la clase. En la figura solo se muestra el nombre de la característica pero cada uno de ellos tiene un valor asociado. También se puede apreciar el arco que une los nodos etiquetados como “manglar” y “agua” y que están etiquetados como “junto_a”, ésta es una relación espacial que forma parte del patrón espacial y dice que una ROI de clase “manglar” se encuentra “junto_a” una ROI perteneciente a la clase “agua”.

Una vez que se genera la red semántica, ésta se utiliza para re-clasificar las imágenes. Este proceso inicia con una pre-clasificación con un algoritmo paramétrico y posteriormente, para cada ROI se calculan las mismas características que se encontraron durante la fase de entrenamiento y se comparan con los diferentes nodos de la red semántica para verificar a qué clase pertenece. En este proceso (la parte procedural de la red semántica) se hace un recorrido por los nodos de la red (búsqueda en profundidad y amplitud) comparando con los valores de los nodos etiquetados

Figura 6. Representación de conocimiento del MMCS con una red semántica



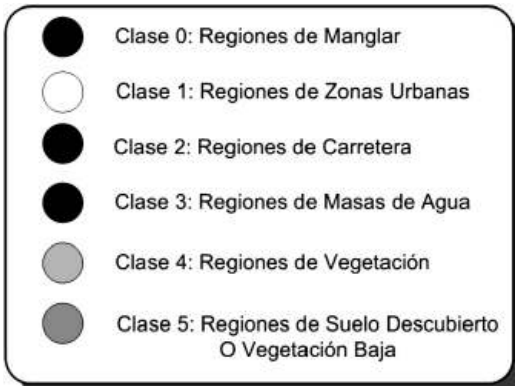
como "concreto" (vea la figura 6, estos nodos se refieren a los valores concretos de las características descriptivas de los patrones) y se le asocia a esta ROI la clase cuyos valores descriptivos se parecen más a ésta. Una vez que se encontraron la(s) clase(s) más parecidas a la nueva ROI, se verifica la información de contexto, que ayuda a rectificar la clasificación generada por los patrones encontrados con los árboles de decisión. Por ejemplo, si se trata de una ROI perteneciente a la clase "Urbano" y la clase más probable resulta ser "Manglar", la red semántica no la clasificará como "Manglar" porque se requiere que se cumpla la condición de que para ser "Manglar" debe haber una ROI adyacente perteneciente a la clase "Agua". Este es un ejemplo de cómo se mejora la precisión de la clasificación utilizando información de contexto.

Para la evaluación del modelo se utiliza la técnica de validación cruzada con 10 pasos, esto es; se dividen los datos de entrenamiento en 10 partes iguales y se realizan 10 pruebas dejando en cada una un décimo de los datos para prueba y el resto para entrenamiento y al final se hace un promedio de los resultados. De esta manera se evita el sobreajuste del algoritmo sobre los datos y se hace una evaluación con menos sesgo. Como se mencionó anteriormente, las pruebas se realizaron con imágenes sintéticas. La siguiente sección muestra los experimentos realizados.

Resultados y Discusión

Esta sección muestra los resultados de clasificación obtenidos con el modelo propuesto. Para obtener una medida cuantitativa de la precisión del modelo se utilizan las Imágenes Multiespectrales Sintéticas IMS descritas en la sección 2. Los algoritmos paramétricos utilizados en el paso de preclasificación (o segmentación en este caso) son: Distancia Mínima a la Media o DMM, Paralelepípedo o PL, Máxima Similaridad o ML y el de Distancia Mahalanobis o DHM (RICHARDS y XIUPING, 1999, MATHER 2004). Los datos de entrada para los algoritmos paramétricos consisten de las medias espectrales por clase por banda. El tratamiento de imágenes satelitales se hizo con el sistema Halcón versión 7.1. Para la preclasificación de las imágenes satelitales con los algoritmos paramétricos se utilizó Matlab. Por último, para la clasificación de las regiones con el modelo MMCS (descrito en la sección anterior) se utilizó C++. El código de colores utilizado se muestra en la figura 7 para las clases Manglar-0, Zona-Urbana-1, Carretera-2, Agua-3, Vegetación-4 y Suelo-Descubierto-Vegetación-Baja-5. La clase "Manglar" agrupa regiones con especies de manglar blanco, negro y rojo o la mezcla de los tres tipos. La clase de "Zona- Urbana"

Figura 7. Esquema de clasificación y su código de colores



principalmente agrupa regiones de medio a completamente ocupadas por edificios, casas o zonas comerciales, principalmente relacionadas con la construcción. En la clase "Carretera" se tomaron en cuenta carreteras pavimentadas y revestidas con concreto. La clase "Agua" incluye masas de agua salada y dulce o la mezcla de ambas. La clase "Vegetación" agrupa regiones de árboles de diferentes tipos mientras que la clase de "Suelo-Descubierto-Vegetación-Baja" agrupa regiones de tipo pastizal y zonas de arena, así como tierra sin vegetación.

En la tabla 1 y en la tabla 2 se muestran los resultados de precisión de la clasificación para la IMS-1 y la IMS-2 respectivamente. En las tablas se compara el resultado de clasificación de los algoritmos paramétricos con el resultado de la re-clasificación de estos resultados utilizando el modelo MMCS. Como se puede apreciar, el modelo MMCS mejora la precisión global obtenida con

Tabla 1. Resultados de clasificación para la IMS-1

	DMM	**DMM MMCS**	**PL**	**PL MMCS**	**ML**	**ML MMCS**	**DMH**	**DMH MMCS**
Precisión Total	95.24	98.09	95.43	98.00	84.37	97.23	84.98	97.04
Est. Kappa	0.94	0.98	0.94	0.97	0.80	0.97	0.82	0.96
Regiones procesadas	735		763		958		860	
Manglar	99.40	100.00	99.40	100.00	94.58	100.00	94.85	100.00
Urbana	99.79	100.00	99.79	100.00	99.51	100.00	99.96	100.00
Carretera	76.01	95.62	77.74	95.24	68.93	89.67	90.45	94.55
Agua	100.00	96.74	100.00	96.02	100.00	98.66	100.00	92.61
Vegetación	99.77	100.00	99.76	100.00	99.88	100.00	99.94	100.00
Suelo desc. Veg. Baja	99.78	96.20	99.76	96.38	39.70	91.82	39.19	92.14

Tabla 2. Resultados de clasificación para la IMS-2

	DMM	**DMM MMCS**	**PL**	**PL MMCS**	**ML**	**ML MMCS**	**DMH**	**DMH MMCS**
Precisión Total	90.64	96.44	86.39	95.54	83.34	96.44	91.68	96.98
Est. Kappa	0.85	0.94	0.81	0.93	0.73	0.94	0.85	0.94
Regiones procesadas	894		928		894		698	
Manglar	100.00	100.00	96.35	0.00	100.0	100.00	100.00	100.00
Urbana	99.00	100.00	99.48	100.00	99.35	100.00	99.91	100.00
Carretera	83.90	94.22	75.18	96.51	63.07	94.22	94.88	94.02
Agua	43.55	100.00	48.80	100.00	100.00	100.00	100.00	100.00
Vegetación	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Suelo desc. Veg. Baja	99.77	78.23	99.78	77.23	15.65	78.23	20.71	77.01

el algoritmo paramétrico pues resulta varios puntos porcentuales arriba de éstos. Sin embargo, en la precisión por clase se obtienen valores más bajos para las clases "agua" y "suelo-descubierto-vegetación-baja" en la IMS-1 y para "suelo-descubierto-vegetación-baja" en la IMS-2. También se puede apreciar que el modelo MMCS mejora notablemente la precisión en la clasificación para las clases "carretera" y "agua" de la IMS-2. Por último, se puede ver la mejora global en precisión de esta comparación en las gráficas de la figura 8 y la figura 9 para la IMS-1 y la IMS-2 respectivamente.

Con estos resultados se demuestra la eficacia del uso del modelo MMCS. Esto es, con el uso de la red semántica que almacena las reglas de clasificación creadas con los árboles de decisión y la información de contexto a partir de relaciones espaciales, se mejora la precisión de la clasificación obtenida con los algoritmos paramétricos.

Figura 8. Gráfica de precisión para la IMS-1

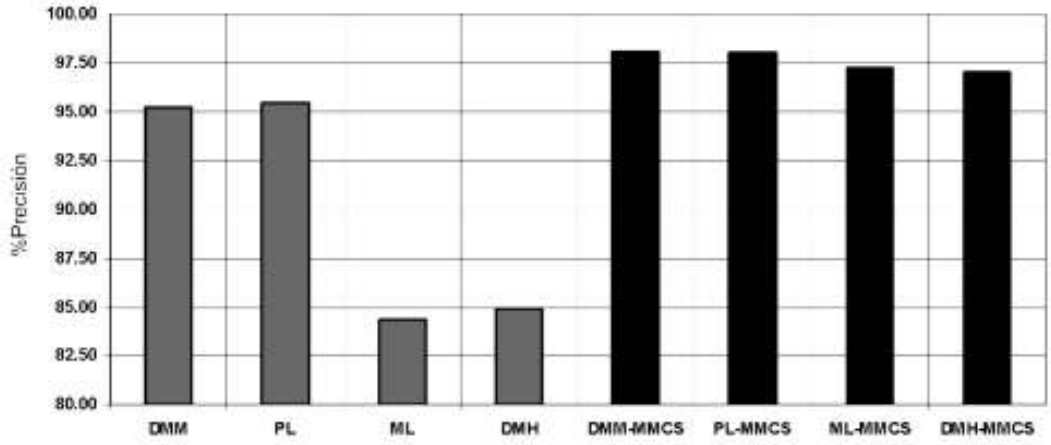
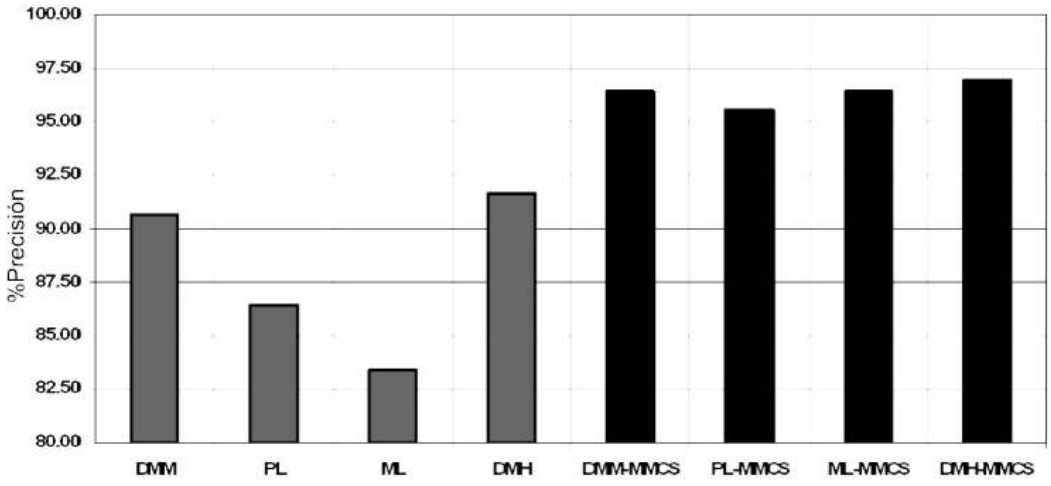


Figura 9. Gráfica de precisión para la IMS-2



Conclusiones y Trabajo Futuro

Este trabajo muestra como el uso de información de contexto a partir de patrones espaciales con la relación topológica de adyacencia se puede mejorar la precisión de la clasificación de imágenes satelitales de los algoritmos paramétricos considerados. Los resultados muestran que la precisión en la re-clasificación supera la obtenida con los algoritmos paramétricos al hacer una post-clasificación con el método MMCS. La mejora de la clasificación varía entre 2.5% y 12.86% para los diferentes algoritmos paramétricos utilizados en la comparación para la IMS-1 y entre 5.3% y 13.1% para la IMS-2.

Como trabajo futuro se plantea incluir más relaciones espaciales para trabajar con más información de contexto y tratar de mejorar más la precisión en la clasificación. Entre las relaciones espaciales que se podrían agregar se encuentran más relaciones topológicas como "intersecta" y "contiene", relaciones de dirección como "norte" y "este" y relaciones de distancia como "cerca" y "lejos". Por otra parte, se propone integrar un algoritmo de segmentación para encontrar las regiones de interés a diferentes niveles de abstracción para poder definir el grado de granularidad y determinar las clases con que se trabajará (clases de alto nivel como agua y vegetación o clases de bajo nivel como diferentes tipos de vegetación).

Referencias

COOK, D. J.; HOLDER, L. B. Substructure Discovery Using Minimum Description Length and Background Knowledge. *Journal of Artificial Intelligence Research*, v. 1, pp 231-255, 1994.

HAN, JIAWEI AND KAMBER, MICHELINE. Data Mining Concepts and Techniques. Morgan Kaufmann, 2001.

HOLDER, COOK, GONZALEZ, AND JONYER. Pattern Recognition and String Matching. Kluwer Academic, 2002.

KOPERSKI, ADHIKARY AND HAN, JIAWEI. SPATIAL DATA MINING: Progress and Challenges. *Research Issues on Data Mining and Knowledge Discovery*, 1999.

MATHER, PAUL M. Computer Processing of Remotely-Sensed Images: An Introduction. John Wiley e Sons, England, 2004.

PECH, SOL, AND GONZALEZ. Graph-Based Knowledge Representation for GIS Data. *Proceedings of the Fourth Mexican International Conference on Computer Science* (ENC), 2003.

QUINLAN, J. R. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, v. 4, pp. 77-90, 1996.

RICHARDS AND XIUPING. Remonte Sensing Digital Image Analysis: An Introduction. Springer-Verlag, 1999.

SAGERER, GERHARD AND NIEMANN, HEINRICH. Semantic Networks for Understanding Scenes. *Advances in Computer Vision and Machine Intelligence*. Springer Verlag, 1997.

Spot Image. http://www.spotimage.fr. Fecha última consulta: 27 de Junio 2006.

KUNZ, SCHILLING, AND VOGTLE. A New Approach for Satellite Image Analysis by Means of a Semantic Network. In W. Forstner and L. Plumer, editors, Semantic Modeling, pp. 20 – 36, Basel, 1997.

LIEDTKE, BÜCKNER, GRAU, GROWE, AND TONJES. AIDA: A System for the Knowledge Based Interpretation of Remote Sensing Data.Third International Airborne Remote Sensing Conference and Exhibition, Vol. II, pp. 313 – 320, 1997.

BÜCKNER, PAHL, STAHLHUT, AND LIEDTKE. GeoAIDA A Knowledge Based Automatic Image Data Analyser for Remote Sensing Data.In CIMA 2001, Bangor Wales, UK, 2001.

MULLER, FEITOSA, MOTA, DA COSTA, DA SILVA, AND TANISAKI. GEOAIDA Applied to SPOT Satellite Image Interpretation. Remote Sensing and Data Fusion over Urban Areas, 2003.