Clasificación de Usos del Suelo en Imágenes Satelitales.

Minería de Datos

257366 Ricardo Vanegas Guerrero
257147 Nelson Augusto Cruz Guzmán
257191 Raúl Trujillo García

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE INGENIERIA
DEPARTAMENTO DE INGENIERIA DE SISTEMAS E INDUSTRIAL
BOGOTÁ
2008

Contenido

INTRODUCCION

El siguiente trabajo es un acercamiento a la minería de datos utilizando la metodología CRISP-DM. Se tratara de utilizar los métodos y herramientas vistas en clase para hacer minería de datos geográficos.

La NASA, con el lanzamiento de su último satélite Landsat 7 en 1999, puso en marcha un proyecto de colaboración, permitiendo el acceso al público a una gran base de datos de imágenes satelitales a través de su programa Earth Observing System. Después Google, con su tecnología Google Earth, volvió masivo el uso de los sistemas basados en posición para los usuarios comunes.

Las imágenes satelitales, sirven como base para un gran cantidad de aplicaciones en la industria. Una de estas es la clasificación de los usos de suelos. En este proyecto, probaremos algunas técnicas de la minería de datos, para comparar los resultados arrojados por el software que existe en el mercado y ver la minería de datos como una alternativa a estas herramientas costosas.

Fase 1: Comprensión del Negocio

Contexto

Un sensor remoto es un instrumento que detecta y registra características de los objetos, sin tener ningún contacto físico con este. En esta categoría se encuentran los satélites, que utilizan sus instrumentos para capturar datos. Uno de los usos que se les da a los satélites es la captura de imágenes satelitales.

Los datos de imágenes satelitales, son una de las muchas fuentes disponibles para la representación de un lugar en la tierra. Los modelos del paisaje pueden integrar una gran cantidad de representaciones con diversos tipos de datos espaciales como lo son las imágenes de radar, imágenes multiespectrales, cartografía temática, mapas de usos del suelo, modelos de elevación digital, etc. Lo que se espera en un futuro, es que las imágenes satelitales asuman una gran importancia, dado el nacimiento de una era caracterizada por la integración de técnicas de sondeo remoto.

Las imágenes satelitales usadas para hacer clasificación de los usos del suelo de una zona, se han extendido en gran medida. Existen algunos métodos para realizar dicha clasificación en software comercial, pero se puede probar otro enfoque. En general, este proceso consiste en convertir los datos de tipo continuo almacenados en varias bandas del espectro, en datos nominales que corresponden al tipo de suelo de una zona en particular. En este proyecto se pretende utilizar algunas técnicas de la minería de datos, para compararlo con el método de Isodata Clustering, que utiliza la firma espectral de la imagen para formar cúmulos.

Objetivos del Negocio y Criterios de Éxito

El objetivo del negocio es la correcta clasificación de los datos provistos, para saber la proporción de cada uno de los usos del suelo en el área de estudio. Con estos datos, una compañía puede decidir si comprar o no una determinada porción del terreno, para sembrar algún cultivo, saber donde instalar un sistema de riego para las zonas con poca concentración de nutrientes o agua.

Inventario de Recursos

Los recursos presentes para el presente proyecto están dados por:

- 3 estudiantes de Ingeniería de Sistemas de la Universidad Nacional de Colombia.
- Una imagen Landsat MMS de 2340 x 3380 píxeles (80m/píxel). La imagen original tenía una resolución de 0.3m/píxel, pero por motivos de confidencialidad, se

redujo la resolución.

- Herramientas de Minería de Datos (WEKA).
- Erdas Imagine 9.1.

Requerimientos, Suposiciones y Restricciones

Uno de los requerimientos del proyecto es el lapso de tiempo dado para terminarlo. Se disponen de un lapso de 10 semanas para entregar los resultados finales. También se supone 1 hora de disponibilidad por estudiante por semana para cumplir con el objetivo.

Riesgos y Contingencias

Los principales riesgos asociados a esta técnica, es que se pueden presentar zonas en las cuales las nubes puedan ocultar el terreno. Para solucionar este problema, no se tendrán en cuenta las zonas con alta presencia de nubes. Se podría utilizar una categoría que clasificara nubes, pero en algunas imágenes, existen terrenos con bandas espectrales muy parecidas a las nubes, como por ejemplo el reflejo que se da por el sol en el agua, o por algunos minerales, es por esto que no se utilizará una clase para las nubes.

Otro riesgo asociado, son las zonas, en las cuales existen árboles altos. El problema radica en que lo que se quiere clasificar son los usos del suelo y en una zona boscosa, desde la imagen satelital, solo se verían las copas de los árboles. Por este motivo se seleccionarán ciertas partes de la imagen que no tengan zonas boscosas. Para afrontar este problema, se utilizan imágenes de radar.

Terminología

Imagen Satelital: representación visual de la información capturada por un sensor montado en un satélite artificial. Estos sensores recogen información reflejada para la superficie de la tierra que luego es enviada a la Tierra y que procesada convenientemente entrega valiosa información sobre las características de la zona representada.

Banda Espectral: distribución energética del conjunto de las ondas electromagnéticas. Referido a un objeto se denomina espectro electromagnético o simplemente espectro a la radiación electromagnética que emite (espectro de emisión) o absorbe (espectro de absorción) una sustancia.

Píxel: es la menor unidad homogénea en color que forma parte de una imagen digital, ya sea esta una fotografía, un fotograma de vídeo o un gráfico.

Landsat: serie de satélites construidos y puestos en órbita por EE. UU. para la observación en alta resolución de la superficie terrestre. Los LandSat orbitan alrededor de la Tierra en órbita circular heliosincrónica, a 705 km de altura, con una inclinación de 98.2º respecto

del Ecuador y un período de 99 minutos. La órbita de los satélites está diseñada de tal modo que cada vez que éstos cruzan el Ecuador lo hacen de Norte a Sur entre las 10:00 y las 10:15 de la mañana hora local. Los LandSat están equipados con instrumentos específicos para la teledetección multiespectral.

Costos y Beneficios

Este proyecto no tiene ningún costo de tipo monetario aparte del tiempo gastado por los integrantes del grupo en realizarlo. Los beneficios potenciales del proyecto, es poder usar herramientas de más bajo costo para realizar clasificaciones preliminares en imágenes satelitales. Las técnicas actuales, exigen el uso de algún software de procesamiento digital de imágenes que por lo general tienen un costo de US$70.000.

Objetivos y Criterios de Éxito de la Minería de Datos

Se comenzará haciendo una depuración de los datos que pueden generar problemas. Muchos de estos datos, se pueden ver a simple vista en las imágenes satelitales, pero para evitar este preprocesamiento, de la imagen total, solo se tomará una pequeña porción de 82 x 100 píxeles, que contiene todas las 6 clases posibles. Si el modelo generado clasifica más del 90% de los datos del conjunto de entrenamiento, se podrá decir que se tuvo éxito en esta clasificación, dado que los métodos actuales tienen un poco más de esta precisión.

Valoración Inicial de Técnicas y Herramientas

En principio, el único preprocesamiento que requiere el conjunto de datos es para definir los valores de las bandas espectrales que son de tipo continuo (entre 0 – 255) para el píxel a clasificar y los vecinos más cercanos. Las técnicas de clasificación más comunes de la minería de datos, como los árboles de decisión, las redes bayesianas o las redes neuronales se utilizarán en este conjunto de datos.

Fase 2: Comprensión de los datos

Reporte de Recolección de Datos Inicial

Nuestros datos fueron sacados de la base de datos generada por Landsat Multi-Spectral Scanner image data. Se utilizaron ejemplos de imágenes satelitales provistos en el programa ERDAS IMAGINE 9.1. Las regiones de estas imágenes fueron clasificadas previamente con un estudio de suelos hecho en campo, y luego digitalizado sobre la imagen satelital. El estudio de suelos está a un nivel detallado, pero el campo de clase genérica, provee un buen ítem para usar en el proceso de minería de datos.

Reporte de Descripción de Datos

Un marco de Landsat MSS imagery consiste de 4 imágenes digitales de la misma escena en diferentes bandas espectrales. Dos de estas están en la región visible (correspondiente aproximadamente de la región del verde a la del rojo de el espectro visible) y dos en el (cercano) infrarrojo. Cada píxel es una palabra binaria de 8-bits, con 0 correspondiendo al negro y 255 al blanco. La resolución espacial de un píxel es alrededor de 80m x 80m. Cada imagen contiene 2340 x 3380 de estos píxeles. Está imagen satelital en particular pertenece a una zona de Australia Occidental, cerca a la ciudad costera de Perth.

El número código para las clases son los siguientes:

Numero	Clase[1]
1	Red Soil
2	Cotton Crop
3	Grey Soil
4	Damp Grey Soil
5	Soil With Vegetation Stubble
6	Mixture Class (All Types Present)
7	Very Damp Grey Soil

[1] No hay ejemplos con la clase 6 en este conjunto de datos.

Reporte de Exploración de Datos

Preguntas a Responder

Por medio de la minería de datos se pretende establecer reglas y parámetros para determinar, en una serie de datos, los tipos de suelo en la superficie terrestre. Estos datos corresponden a los pixeles de imágenes satelitales Landsat 4, los cuales están en cuatro bandas espectrales según el tipo de satélite que adquiera la imagen. De estas bandas se encuentran dos en la escala visible y dos en la escala no visible del espectro. Los datos toman valores de 0 a 255 para cada banda, que en términos prácticos corresponderían del negro al blanco. La combinación de las 4 bandas representa 7 diferentes tipos de suelo.

Visualización de Datos

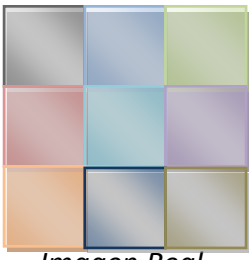


Imagen Real

Proyecciones de la imagen real

Un pixel de la gráfica se verá de la siguiente manera a ser analizado se verá así:

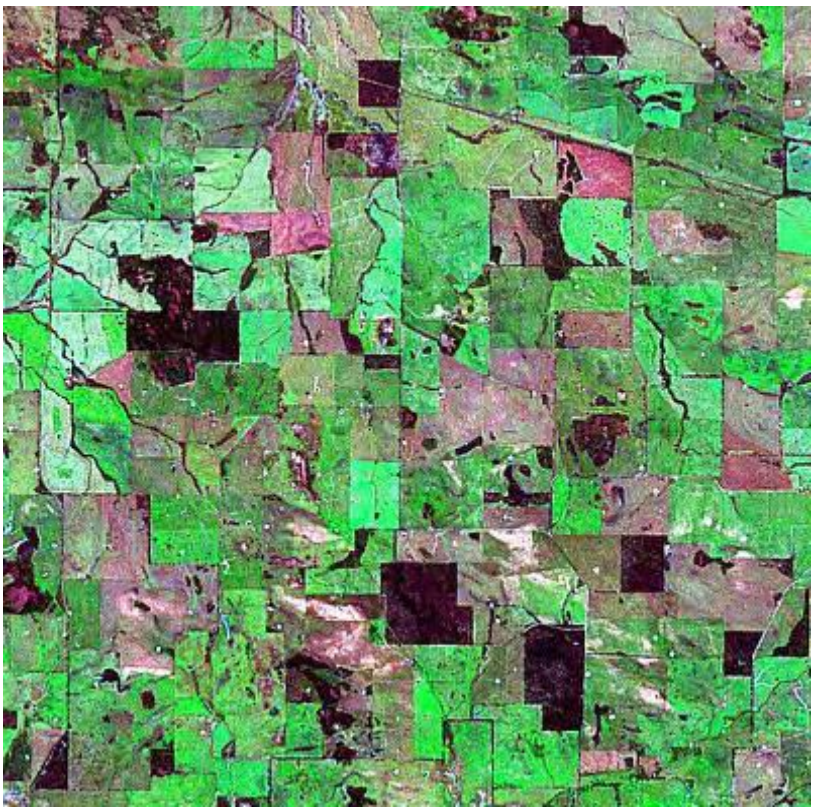
92 115 120 94 84 102 106 79 84 102 102 83 101 126 133 103 92 112 118 85 84 103 104 81 102 126 134 104 88 121 128 100 84 107 113 87 Red Soil

Donde los 9 primeros valores corresponden a niveles de color (del 0 al 255) en el espectro de los colores verdes o rojos del visible, y de igual manera los siguientes 9 valores, los últimos 19 valores hacen referencia a los dos espectros que no están en el visible y el ultimo valor representa el tipo de suelo al que se hace referencia el pixel central.

Atributos

Las clases de los suelos están designadas de la siguiente manera:

1	Tierra roja
2	Cosecha de algodón
3	Tierra gris
4	Tierra gris húmeda
5	Rastrojo de vegetación
6	Clase mezcla (todos los tipos presentan)
7	Tierra gris muy húmeda



Un ejemplo de imagen satelital con diversos tipos de suelo

Hipótesis

Se supone que los pixeles son entidades propias con unidades enteras, los datos has sida preprocesador por sola pación de imágenes que reducen los errores de detención de

suelos por lo cual los datos se encuentra completos, sin errores y hacen parte de una muestra representativa de la población de datos que se van a analizar.

Efectos

Los resultados, reglas o características encontradas en la exploración de datos van a influir directamente con las expectativas del proyecto, ya que inicialmente no se presupone ninguna regla física que determine una regla para la distribución de los suelos.

Teoría

Debido a los cambios climáticos, a los movimientos de las placas tectónicas, a los cambios en la capa de ozono, a la contaminación y a la radiación generada por el sol y otros muchos factores, se cree que debido a todos estos el suelo a cambiado de una forma definida al igual que los suelos cercanos, por lo cual se desea determinar que tipo de reglas se pueden determinar con el análisis de imágenes satelitales de los suelos.

Reporte de Calidad de los Datos

La base de datos de la muestra pequeña se proporcionó por Ashwin Srinivasan el Departamento de Estadística y Ciencia del Modelado La universidad de Strathclyde Glasgow-Escocia, REINO UNIDO y generado por medio datos comprados a la NASA por el Centro australiano para detección remota, y usó para la investigación al Centro para detección remota La universidad de Nueva Gales Sur, Kensington, Australia.

Lo anterior garantiza la realidad de los datos, y confianza que a estos se le atribuyen por porvenir de fuentes científicas de alto nivel, y que garantizan la no manipulación de los datos, adicional a ello los datos fueron pre procesados, corregidos, y escogidos con una prueba de 4435 con una base de prueba de 2000 datos, implicando así que estos sean una muestra representativa de la población analizada, es decir de las imágenes satelitales sobre los tipos de suelo.

Fase 3: Preparación de los datos

Reporte de Descripción del Conjunto de Datos

La base de datos es una (pequeña) sub-área de una escena, consistente de 82 x 100 píxeles. Cada línea de datos corresponde a un vecindario cuadrado de 3x3 píxeles completamente contenidos en el sub-área de 82x100. Cada línea contiene los valores de los píxeles en las cuatro bandas espectrales (convertidas a ASCII) de cada uno de los 9 píxeles en el vecindario de 3x3 y un número indicando la etiqueta de clasificación del píxel central.

En cada línea de datos los cuatro valores espectrales del píxel superior-izquierdo son dados primero seguido por los cuatro valores espectrales del píxel superior-medio y luego por los del píxel superior-derecho, y así sucesivamente con los píxeles leídos de izquierda-derecha y de parte superior a inferior. Así, los cuatro valores espectrales para el píxel central son dados por los atributos 17, 18, 19 y 20. Si uno desea, puede utilizar únicamente estos atributos, para evitar el trabajo de usar los vecinos más cercanos. Los datos son dados en orden aleatorio y ciertas líneas de datos han sido removidas para que no se pueda reproducir la imagen original con este conjunto de datos.

NUMERO DE EJEMPLOS

Conjunto de entrenamiento	4435
Conjunto de prueba	2000

NUMERO DE ATRIBUTOS

36 (= 4 bandas espectrales x 9 píxeles por vecindad)

ATRIBUTOS

Los atributos son numéricos, en el rango 0 a 255.

CLASES

Hay 6 clases: 1, 2, 3, 4,5 y 7.

No hay ejemplos con la clase 6 en este conjunto de datos, han sido removidos debido a dudas sobre la validez de esta clase.

Fase 4: Modelo

Supuestos del Modelo

Se asume que los datos que entren al modelo son confiables, con muy poco ruido y sin datos anormales, por lo cual no tendremos que usar ninguna técnica especial (como el podar el árbol: excepto para encontrar reglas de decisión cortas) para tratar estos inconvenientes. Se asume que el método es el de menos errores de clasificación y al hacer las pruebas es el que mejor modela los datos.

Diseño de Pruebas

Los árboles de decisión son un método fácil, efectivo, que arroja reglas de decisión por inducción para la tarea de minería de datos de clasificación, este método permite manejar datos de alta dimensionalidad. Consideramos este método como una buena aproximación para cumplir nuestros objetivos de clasificación de suelos en una imagen satelital.

- El modelo será probado con un conjunto de datos de prueba de 2000 instancias. El resultado del modelo para el conjunto de datos de prueba nos permitirá evaluar el rendimiento de este y además, la evaluación del modelo se hará en base a la matriz de confusión y los diferentes indicadores de precisión.
- Los datos requeridos para la prueba deben ser compatibles con los datos utilizados para entrenar el modelo, esto se logra utilizando pruebas de la misma fuente de donde se obtuvieron los datos para entrenar el modelo, más no los mismos datos.
- En resumen lo que se hará será entrenar un modelo de árboles de decisión con un conjunto de datos de entrenamiento, luego se probará este modelo con un conjunto de datos de prueba, de la misma fuente de los datos de entrenamiento, y se verificará la precisión con la que el modelo clasifica los datos.

Descripción del Modelo

Modelo: Árbol de Decisión C4.5

Este modelo permite construir un árbol de decisión con base en atributos numéricos (lo que no pasa con el ID3), lo que nos permite pasar nuestros datos numéricos directamente al modelo y cumplir nuestro objetivo de clasificación.

Parámetros: weka.classifiers.trees.J48 -C 0.1 -M 2

confidenceFactor 0,1 (El factor de confianza usado para podar el árbol, valores más pequeños, podan mas ramas del árbol)

minNumObj 2 (El número mínimo de instancias por hoja final en el árbol)

Precisión:

Instancias clasificadas correctamente	1730	86.5 %
Instancias clasificadas incorrectamente	270	13.5 %

Classified as						
a	b	c	d	e	f	**Real Class**
373	13	2	6	0	3	a=grey soil
41	**111**	4	53	1	1	b=damp grey soil
3	1	**195**	15	6	17	c=soil with vegetation stubble
13	45	11	**397**	0	4	d=very damp grey soil
1	3	0	2	**217**	1	e=cotton crop
5	2	13	2	2	**437**	f=red soil

Tabla 1. Matriz de Confusión

El modelo tiene un alto valor de instancias clasificadas correctamente, y es el mejor valor obtenido a comparación de otros modelos de clasificación. Es de destacar que se escogió el atributo e1 como nodo raíz, siendo este un atributo de los que describe el píxel central, que es el que se quiere clasificar.

El modelo muestra la importancia de los atributos centrales, especialmente la del píxel e en la clasificación del tipo de terreno de un dato.

También se puede ver en la matriz de confusión, que la clase *Damp Grey Soil*, no es muy bien clasificada por el modelo. En menor proporción lo es la clase *Very Damp Grey Soil*, pero igual un gran número de instancias fueron mal clasificadas para esta clase.

Evaluación del Modelo

El plan de pruebas se realizó de acuerdo a lo planeado, no se presentaron problemas con los datos o con la técnica usada para crear el modelo.

La tabla nos muestra gran precisión en la clasificación de todas las posibles clases, teniendo la menor precisión la clase *Damp Grey Soil*, y la mayor precisión la clase *Cotton Crop.* En la matriz de confusión presentada anteriormente se puede observar la mayor confusión del modelo en la clasificación de las clases b = *Damp Grey Soil* y d = *Very Damp Grey Soil*. En nuestro modelo todos los errores tienen el mismo costo.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.94	0.039	0.856	0.94	0.896	0.954	Grey soil
0.526	0.036	0.634	0.526	0.575	0.839	Damp grey soil
0.823	0.017	0.867	0.823	0.844	0.922	Soil with vegetation stubble
0.845	0.051	0.836	0.845	0.84	0.92	Very damp grey soil
0.969	0.005	0.96	0.969	0.964	0.98	Cotton crop
0.948	0.017	0.944	0.948	0.946	0.97	Red soil

Tabla 2. Precisión del Modelo

Es de notar que la firma espectral de la clase *Damp Grey Soil* puede ser muy parecida a otras clases y por eso se generan tantos errores con la clasificación de esta clase. Una aproximación mejor puede ser cambiar los costos de la clasificación para que genere menos errores de este tipo. Una dificultad con este enfoque, es que debe ser un experto el que provea los costos de las clasificaciones.

En este modelo podríamos cambiar algunos parámetros, para nuestro caso podríamos variar principalmente los parámetros confidenceFactor para hacer un árbol más pequeño que nos de reglas de decisión más simples, según el valor que le demos y el parámetro minNumObj que nos da un hojas con un mínimo numero de instancias que nos permite deshacernos de hojas poco representativas en el árbol.