Teledetección: Agua y desarrollo sostenible. XIII Congreso de la Asociación Española de Teledetección.
Calatayud, 23-26 de septiembre de 2009. pp. 77-80. Editores: Salomón Montesinos Aranda y Lara Fernández Fornos

TÉCNICAS DE CLASIFICACIÓN EN EL ENTORNO DE WEKA PARA LA DETERMINACIÓN DE CULTIVOS DE REGADÍO (CÍTRICOS) EN LIBRILLA, MURCIA (SE ESPAÑA)

J. C. González, M. Castellón y M. J. Castejón.

Fundación Instituto Euromediterráneo del Agua. Complejo de Espinardo, Carretera Nacional 301. Espinardo 30100 Murcia. jcgonza@f-iea.es

RESUMEN

El estudio para la determinación de cultivos de regadío (cítricos) en el municipio de Librilla se hizo mediante la clasificación de imágenes Landsat por diversos algoritmos de clasificación implementados en WEKA, un completo banco de herramientas para minería de datos y aprendizaje automático, de libre distribución.

En un primer paso las imágenes fueron procesadas y georreferenciadas dentro del sistema de información geográfica GRASS, para luego ser exportadas en formato ASCII, junto a los datos de un muestreo de campo previo realizado sobre una imagen de alta resolución. Los datos organizados adecuadamente en una matriz fueron introducidos en WEKA por medio de la interfaz Explorer (Explorador) y clasificados en la interfaz Knowledge Flow (Flujo de Conocimiento). Los algoritmos probados son Logistic Model Trees, Naive Bayes y Stacking.

De la matriz resultante del proceso de clasificación en WEKA, se toma la última columna, que es la predicha por el sistema, misma que es organizada e importada por GRASS para su visualización y análisis mediante una matriz de confusión.

En base a los resultados obtenidos mediante los diversos algoritmos de clasificación usados en este trabajo se determina su utilidad en un proceso de clasificación para la determinación de áreas de cultivos.

ABSTRACT

The study to determinate irrigated lands (citruses) in Librilla was carried out using Landsat images classification by various algorithms implemented in WEKA. It is an open source software which contains a collection of tools for data mining tasks and for developing new machine learning schemes.

Firstly, images were processed and georeferenced by the geographic information system GRASS. Later they were exported as ASCII files, as well as the results of a previous field sampling made from a high resolution image. Arranged data in a matrix were introduced to WEKA using Explorer interface and classified in the Knowledge Flow interface. Logistic Model Trees, Naive Bayes and Stacking were the algorithms tried.

From the matrix resultant of classification process in WEKA, last column which is the one predicted by the system, was arranged and imported by GRASS to be visualized and analyzed using a confusion matrix.

According to the obtained results, the utility of different classification algorithms used in this study is analyzed trough a classification process to determinate irrigated lands.

Palabras clave: WEKA, Logistic Model Trees, Naive Bayes, Stacking, clasificación, Landsat, cítricos.

INTRODUCCIÓN

Para la determinación de cultivos es probada la validez del uso de la teledetección por medio de imágenes de satélite y mediante la aplicación de diversos procesos de clasificación.

Dada la importancia del cultivo de cítricos en la región de Murcia, se ha escogido como una zona primera de estudio el término municipal de Librilla ubicado dentro del Valle de Guadalentín, con un clima enmarcado dentro del dominio mediterráneo, con una relación entre precipitación y temperatura que establece una característica de aridez, pero que a pesar de estas condiciones, dentro de su actividad económica destaca la agricultura, con cultivo de frutales bajo condiciones de regadío, primordialmente cítricos.

Para el estudio se han utilizado imágenes captadas por el satélite Landsat 7 sensor ETM +, las cuales han sido debidamente procesadas en el sistema de información geográfica GRASS, por medio del cual se ha exportado la información a la herramienta de minería de datos WEKA, para la aplicación de los diversos algoritmos de clasificación implementados en la misma, cuyos resultados han sido introducidos en el SIG para su visualización y análisis.

Se han probado los algoritmos de clasificación Logistic Model Trees, basado en

árboles de decisión, Naive Bayes considerado una forma sencilla de red Bayesiana y Stacking, un meta clasificador.

El objetivo del trabajo ha sido el probar, comparar y analizar los resultados de estos algoritmos de clasificación y determinar cuál es el que más se ajusta a las condiciones de la zona de estudio para la discriminación del cultivo de cítricos.

METODOLOGÍA

Delimitada el área de estudio mediante una máscara correspondiente al término municipal de Librilla, se procedió a cargar en el sistema de información geográfica GRASS, las imágenes Landsat 7 ETM+, debidamente georreferenciadas, correspondientes a las fechas 15 de noviembre de 2004 y 10 de mayo de 2005, con las cuales en un inicio se procedió a calcular el índice de vegetación normalizada (NDVI), para ser introducido como una variable en el proceso de clasificación, por ser un claro indicador de la actividad fotosintética de la planta (Bannari et al. 1995).

Los datos de verdad terreno necesarios para el proceso de clasificación fueron recolectados mediante trabajo de campo y fotointerpretación sobre una imagen del satélite Quickbird y digitalizados y rasterizados en el SIG de trabajo. Las clases identificadas fueron: 1 cítricos, 2 pinar, 3 urbano, 4 natural, 5 herbáceo, 6 desnudo y 7 balsas.

El muestreo para realizar el contraste de los resultados fue realizado sobre la imagen de alta resolución estableciendo una red de puntos aleatorios distantes entre sí 500 metros, estableciendo en éstos, una cruz con puntos equidistantes 5 pixeles y comprobando si coinciden con áreas de cultivo de cítricos o no.

Luego todas las capas raster fueron exportadas en formato ASCII y reunidas en una sola gran matriz para su introducción en WEKA.

WEKA es el acrónimo de Waikato Environment for Knowledge Analysis y nace del esfuerzo de un grupo de investigadores del Machine Learning Laboratory de la Universidad de Waikato en Nueva Zelanda, como software de código abierto bajo los términos de la GNU GPL y es un banco completo de herramientas para aprendizaje automático y minería de datos (Inza et al. 2006)(Soman et al. 2006).

La introdución de los datos se ha realizado por medio de la interfaz Explorer (Explorador) y las clasificaciones mediante la interfaz gráfica Knowledge Flow (Flujo de conocimiento), que permite diseñar y configurar una secuencia de procesamiento de los datos (Figura 1).

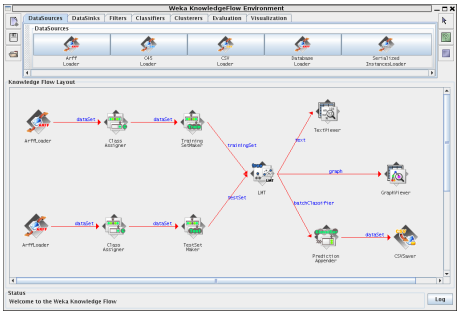

Figura 1.- WEKA-Diagrama Flujo de conocimiento.

Los algoritmos de clasificación probados en este estudio fueron:

Logistic Model Trees

Los árboles de inducción y la regresión logística lineal, son considerados dos métodos populares de clasificación, cada uno con sus ventajas y desventajas.

Así el algoritmo Logistic Model Trees (LMT), combina modelos de regresión logística con árboles de inducción, por lo que es un modelo análogo de árboles para problemas de clasificación.

Un modelo logístico de árbol consiste básicamente en una estructura de árbol de decisión estándar con funciones de regresión logística en las hojas. Como un modelo de árbol, es un árbol de regresión con funciones de regresión en las hojas.

Como en un árbol de decisión ordinario, una prueba en uno de los atributos esta asociada con cada nodo interno. Para un atributo nominal con k valores, el nodo tiene k nodos secundarios y los casos son clasificados hacia abajo en una de las k ramas dependiendo del valor del atributo. Para atributos numéricos, el nodo tiene dos nodos secundarios y la prueba consiste en comparar el valor del atributo con un umbral: un caso es ordenado hacia abajo en la rama izquierda si su valor para el atributo es más pequeño que el umbral, y en la rama derecha en el caso contrario.

De una manera formal un modelo logístico de árbol consta de una estructura de árbol compuesto por un conjunto de nodos internos o no terminales N y un conjunto de hojas o nodos terminales T (Landwehr et al. 2005) (Witten y Frank 2005).

Naive Bayes

El algoritmo Naive Bayes (NB), ampliamente usado en procesos de clasificación, se lo considera como una forma especial, o como el modelo más simple de clasificación basado en una red Bayesiana (Hernández et al. 2004) (Lowd y Domingos 2005) y dentro del campo de las máquinas de aprendizaje y minería de datos, es reconocido como uno de los algoritmos más eficientes y efectivos de aprendizaje inductivo (Zhang 2004).

El presente algoritmo centra su fundamento en la hipótesis de que todos los atributos son independientes entre sí, conocido el valor de la variable clase. El algoritmo representa una distribución de una mezcla de componentes, donde cada componente dentro de todas las variables se asumen independientes. Esta hipótesis de independencia da lugar a un modelo de un único nodo raíz, correspondiente a la clase, y en el que todos los atributos son nodos hoja que tienen como único origen a la variable clase (Hernández et al. 2004) (Lowd y Domingos 2005).

En varias situaciones se ha demostrado que el algoritmo en cuestión, trabaja mejor en dos casos: cuando los atributos son completamente independientes, como es lógico esperar dada su premisa, y cuando los atributos son funcionalmente dependientes, lo que ya es menos evidente; y llegando a presentar sus peores resultados en situaciones intermedias entre estos dos extremos (Rish 2001).

Stacking

Stacking (S) es un meta clasificador, de estructura bastante sencilla, que se basa en la combinación de modelos, construyendo un conjunto con los generados por diferentes algoritmos de aprendizaje. Como cada uno de los modelos se aprende con un mecanismo de aprendizaje diferente, se logra que los modelos del conjunto sean distintos (Hernández et al. 2004).

Una de las formas de combinar los clasificadores es usando voto mayoritario, sin embargo, esto tiene sentido cuando todos los clasificadores se desempeñan con una precisión aceptable.

Para producir una clasificación, este utiliza un meta algoritmo que aprende según las salidas de los clasificadores en los que se basa. En términos generales, partiendo de los datos se construyen n clasificadores distintos. Las salidas de estos se usan como atributos de un nuevo clasificador (Morales y González 2007).

Así el meta clasificador busca descubrir la mejor manera de cómo combinar los resultados de los clasificadores base (Witten y Frank 2005).

En el caso en estudio, se usaron como clasificadores base los algoritmos LMT y NB.

Las clasificaciones realizadas fueron multitemporales y multiespectrales con adición de los índices de vegetación de cada una de las fechas, y los datos resultantes de las mismas fueron introducidos en GRASS, y recodificados en un mapa de cítricos y no cítricos, y se los cruzó en una matriz de confusión con el muestreo realizado para el efecto y validados mediante el estadístico Kappa (Congalton 1991) y el coeficiente de fiabilidad.

RESULTADOS

Los resultados obtenidos en los procesos de clasificación probados, se resumen en la siguiente tabla (Tabla 1), en la cual se aprecia que el mayor porcentaje de fiabilidad se logra con el algoritmo Logistic Model Trees con un valor de 90.70 %, corroborado con un valor de Kappa de 0.82 y da una estimación de 2590 hectáreas de cítricos. La menor fiabilidad se obtiene con el algoritmo Naive Bayes con un porcentaje de 84.95 %, con un valor de Kappa de 0.76 y una estimación de 2072 hectáreas.

Tabla 1.- Resultados clasificaciones.

Algoritmos	Kappa	Fiabilidad %	Estimación ha
LMT	0.82	90.70	2590
NB	0.76	84.95	2072
S	0.75	86.28	2445

El algoritmo Stacking, no logra superar en calidad a otros algoritmos, a pesar de su condición de meta clasificador.

El mapa resultante de la clasificación que mejor resultados presenta, es el siguiente (Figura 2)

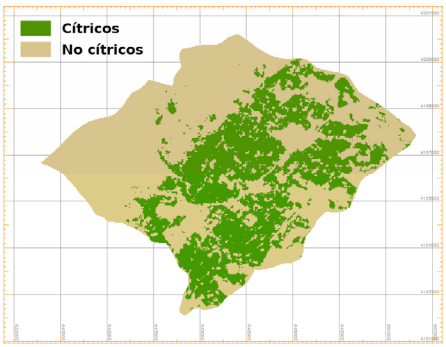

Figura 2.- Mapa clasificación LMT.

CONCLUSIONES

El clasificador que mejores resultados presenta para la discriminación de cítricos, según las condiciones presentes en el término municipal de Librilla es el basado en el algoritmo Logistic Model Trees, implementado en la plataforma WEKA.

Se corrobora la utilidad de las imágenes Landsat 7 para el fin indicado.

BIBLIOGRAFÍA

Bannari, A., Morin, D., Bonn, F. and Huete, A. 1995. A review of vegetation indices. *Remote Sensing Reviews*, Vol 13: 95-120.

Congalton, R. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, Vol 37: 35-46.

Hernández, J., Ramírez, M.J. y Ferri, C. 2004. *Introducción a la Minería de Datos*. Pearson Prentice Hall, Madrid.

Inza, I., Armañanzas, R. y Santafé, G. 2006. Una aproximación al software WEKA. *Aprendizaje Automático: Conceptos básicos y avanzados. Aspectos prácticos utilizando el software WEKA*. Pearson Eduación S.A., Madrid.

Landwehr, N., Hall, M. and Frank, E. 2005. *Logistic Model Trees*. Machine Learning, Volume 59, Issue 1-2 : 161-205.

Lowd, D., and Domingos, P., 2005. Naive Bayes Models for Probability Estimation. *Proceedings of the 22 International Conference on Machine Learing*, Bonn.

Morales, E. y González, J. 2007. Stacking o Stacked generalization. *Aprendizaje 2*. Disponible en: http://ccc.inaoep.mx/~emorales/Cursos/Aprendizaje2/principal.html.

Rish, I. 2001. An empirical study of the naive Bayes classifier. *IBM Research Report*, Yorktown Heights.

Soman, K., Diwakar, S. and Ajay, V. 2006. Machine Learning with Open Source and Commercial Software. *Insight into Data Mining: Theory and Practice*. Prentice Hall of India Private Limited, New Delhi.

Witten, I., and Frank, E. 2005. *Data Mining:Practical Machine Learning tools and Techniques*. Morgan Kaufman Publishers, San Francisco.

Zhang, H. 2004. *The Optimality of Naive Bayes*. American Association for Artificial Intelligence.

AGRADECIMIENTOS

Expresamos nuestro agradecimiento a la Fundación Instituto Euromediterráneo del Agua, por las becas que permitieron esta investigación y al Instituto Universitario del Agua y del Medio Ambiente por poner a disposición las imágenes de satélite.