ANALYSIS OF THE ADDITION OF QUALITATIVE ANCILLARY DATA ON PARCEL-BASED IMAGE CLASSIFICATION.

Recio, J.A., Hermosilla, T., Ruiz, L.A., Fdez-Sarría, A.

Departamento de Ingeniería Cartográfica, Geodesia y Fotogrametría.
Universidad Politécnica de Valencia. Camino de Vera s/n, 46022 Valencia, Spain.
jrecio@cgf.upv.es

KEY WORDS: Database updating; object-oriented; classification trees; high resolution imagery.

ABSTRACT:

Parcel-based classification of high-resolution images is one of the most reliable alternatives for the automatic updating of land cover/land use geospatial databases. Each parcel can be characterized by means of a set of features extracted from the image, its outline, the contextual relationships with its neighbours, etc. Qualitative information about the former land use contained in the geospatial database to be updated can also be considered, since this is often related to the current land use of the parcel. In this study, we analyse the effect that the addition of the class contained in the geospatial database as a descriptive feature has on the final classification accuracy. Since the inclusion of descriptive features as discrete ancillary data requires the employment of classifiers that are able to deal with this type of data, we chose the C5.0 algorithm, which allows us to include this type of information to create classification trees. Several accuracy degrees of the database information have been simulated in order to study the influence of this parameter on the classification accuracy. In all cases, the addition of this information as a feature increases the overall accuracy of the classification. The more precise the geospatial database information is, the more it is used in the different rules that compose the classification tree, and the more accurate the final classification is. These results have special relevance for the automatic updating of geospatial databases.

1. INTRODUCTION

Numerous studies have demonstrated that the integration of ancillary data improves the classification results obtained using only remote sensing images. Ancillary data are of a diverse nature, but have been mainly composed by digital elevation model derived data (slope, aspect), cartographic and topographic data (roads, parcel limits, hydrological networks, etc.) or historical data of land uses/land covers. Information regarding temperature, pluviometry, geology, natural catastrophes, etc., has been also included depending on the case to be analyzed. Topographic information, for example, improves the image classification accuracy when the study is performed on a local scale. On the other hand, climatologic data are more useful on regional or continental scales (Skidmore, 1989).

The integration of the ancillary data into the classification has usually been divided in three categories: before, after or during classification. Integration before classification can be done through stratification, where ancillary data are used for dividing zones which have to be analyzed in a different way (Strahler et al., 1978; Katila and Tomppo, 2002; Leránoz et al., 2007). Many authors have employed cartographic limits (Walter, 2004; Berberoglu and Curran, 2004; Ruiz et al., 2007) in order to segment the image and to create objects to be classified in an object-oriented classification.

Some authors have used ancillary data after classification in order to improve or correct the results of the classification. Land use, pluviometric (Cohen et al., 2000) or topographic information (Raclot et al., 2005) has been added in order to improve the separability between classes with a similar spectral response. Blaes et al. (2005) proposed an iterative classification where those objects assigned to a different class than the one contained in the database are analyzed in a different manner. Therefore, it can be detected if these differences are due to changes or correspond to errors of the first classification.

Integration of ancillary data during classification can be done in different ways. Many authors (Heipke, 1999; Olsen et al., 2002; Walter, 2004; Blaes et al., 2005) employ the land cover/land use contained in agricultural and cartographic geospatial databases to automatically provide training samples for the classifier. This entails that the information contained in the databases has to be sufficiently accurate and that the elements belonging to a same class have a homogeneous response in the image.

The historical information about the land cover/land use of a location and also its geographical context can be a sign of the current land cover/land use at the time of study. Therefore, it is possible to estimate the a priori probability of each class in a classification using the maximum-likelihood method, improving the overall accuracy of the classification. This is the method employed by Janssen and Middelkoop (1992), Maselli et al. (1995), Pedroni (2001) and Blaes et al. (2005). Heipke (2000) and Pakzad (2002) used the information contained in geospatial databases to condition and restrict the possible class transitions that can occur in a given land use. This methodology requires a prior knowledge of the area to be analyzed.

The easiest and most employed technique to include ancillary data during the classification process is to use it as an additional descriptive feature. This technique is determined by the data type (continuous or discrete), and also by the classifier employed, because discrete data is not tolerated by statistical or distance-based classifiers. Ancillary data derived from digital terrain models, such height, slope or aspect, has been included in many studies (Hoffer et al., 1975; Hutchinson, 1982; Bruzzone et al., 1997; Treltz and Howarth, 2000; Lawrence and Wright, 2001) due to its simplicity of use (Pedroni, 2001) and

the well known improvement in the classification accuracy (Hoffer et al., 1975). Some authors included land use/land cover information contained in the geospatial database as a descriptive feature (Huang and Jensen, 1997; Rogan et al., 2003). This can be done by means of machine learning techniques, such as decision trees or neural networks, which allow us to deal with discrete data and do not require the definition of a-priori probabilities to weight ancillary data regarding to the spectral information (Lawrence and Wright, 2001). The use of land use information as ancillary data enables to extract knowledge related to trends or relations about the evolution of land uses between two updating stages.

The accuracy rate or updating degree of a database is unknown, being conditioned by the time passed since its creation or the last update, as well as by the dynamism of the represented territory. The aims of this study are: (i) to evaluate the convenience of including, during the classification process, the land cover/land use available in the geospatial database as one descriptive feature. (ii) To verify how the updating degree of the information of a database influences the accuracy of the classification and the significance of this feature in the classification. (iii) To identify the causes and the nature of the errors produced, and to follow their evolution as the database updating degree increases. An object-oriented classification of the image has been performed, using cartographic limits obtained from cadastral parcels contained in a geospatial database to define the objects.

2. STUDY AREA , SPECTRAL AND ANCILLARY DATA

The study was performed over a rural area located on the Mediterranean coast of Spain. Nine different classes were considered to obtain the classification: *Citrus orchards*, *Young citrus orchards*, *Buildings*, *Forest*, *Carob-trees*, *Irrigated crops*, *Shrub lands*, *Roads* and *Arable lands* (Figure 1).

The remotely sensed data used in this study are 0.5 m/pixel resolution digital aerial orthophotographs, acquired in August 2005 with a Digital Mapping Camera (DMC) sensor. The spatial resolution is achieved through a fusion process between panchromatic and multispectral bands. The system is composed of three bands in the visible part of the electromagnetic spectrum (0.4-0.58 µm, 0.50-0.65 µm, and 0.59-0.675 µm), one in the near infrared (NIR) (0.675-0.85 µm) and a panchromatic band.

The ancillary data comes from a cadastral geospatial database. The limits of the cadastral parcels are defined in vectorial format, and an associated database contains the land use of each parcel. This information has been modified in a controlled manner. Erroneous land uses were assigned to a number of parcels, so that the updating degree of the database was known in each test, this value ranging from 40% to 90% with increments of 10%.

A dataset of 1350 parcels, 150 per class, has been used. For each class, 50 parcels have been used as training samples, reserving the remaining 100 as evaluation samples.

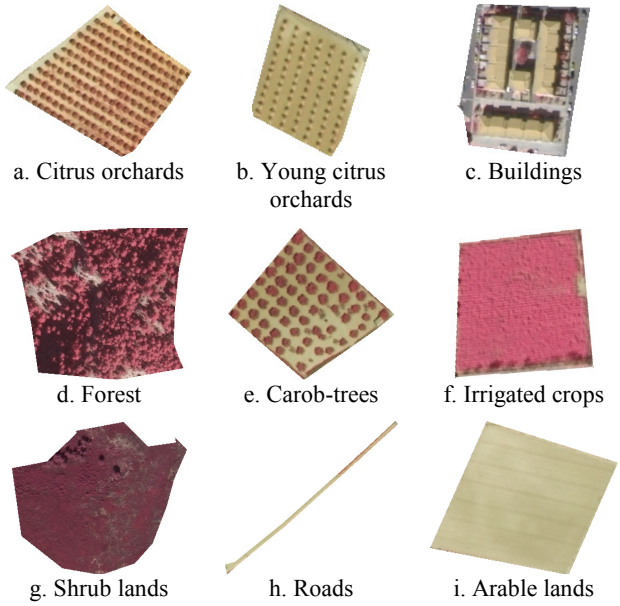

Figure 1. Examples of parcels of defined classes in colour infrared composition.

3. METHODOLOGY

The classification scheme carried out begins creating image objects by using cartographic limits. Each object is processed deriving features related to its spectral characteristics. The land use contained in the database, with different simulated updating degrees, is added as a descriptive feature, and independent classifications for each degree were performed. Objects were classified in one of the predefined classes by applying the decision trees built with the C5.0 algorithm, combined with the boosting method. The results of the classifications are assessed using the evaluation samples.

3.1 Object definition

The definition of the classification objects was based on the cartographic limits obtained from regular cadastral parcels within the geospatial database. A morphological erosion filtering was applied to each object with a circular structuring element with a 5 pixel diameter. This was done in order to avoid the inclusion of pixels that do not belong to the parcel, due to potential geometric errors in the location of the parcel limits. The use of cartographic limits to define objects allows us to relate the features derived from the image with the land use information in the database.

3.2 Descriptive features

Since a high number of features could difficult the interpretation of the results, and as this test is focused on the analysis of the integration of discrete information as a descriptive feature, only spectral features have been derived from the image. These features provide information about the spectral response of objects, which depends on land coverage types, state of vegetation, soil composition, construction materials, etc. Spectral features are especially useful in the characterization of spectrally homogeneous objects, as herbaceous crops or fallow fields. The bands considered were near infrared, red, green and the Normalized Difference

Vegetation Index (NDVI). For each band, the mean and standard deviation values were calculated.

3.3 Classification through decision trees

The objects have been classified by using decision trees. A decision tree is a set of organized conditions in a hierarchical structure, in such a way that the class assigned to an object can be determined following the conditions that are fulfilled from the tree roots (the initial data set) to any of its leaves (the assigned class). The algorithm employed in this study is the C5.0, which is the latest version of the algorithms ID3 and C4.5 developed by Quinlan (1993). This algorithm is the most widely used to deduce decision trees for classifying images (Zhang and Liu, 2005). The C5.0 algorithm can manage several data types, such as continuous or discrete, which highly increases the possibility of adding descriptive features coming from diverse data sources to perform the classification.

The process of building a decision tree begins by dividing the collection of training samples using mutually exclusive conditions. Each of these sample subgroups is iteratively divided until the newly generated subgroups are homogeneous, that is, all the elements in a subgroup belong to the same class. These algorithms are based on searching partitions to obtain purer data subgroups, which are less mixed than the previous group where these come from. For each possible division of the initial data group, the impurity degree of the new subgroups is computed, and the condition which gives the lower impurity degree is chosen. This is iterated until the division of the original data into homogeneous subgroups is carried out by using the gain ratio as splitting criterion (Quinlan, 1993). This criterion employs information theory to estimate the size of the sub-trees for each possible attribute and selects the attribute with the largest expected information gain, that is, the attribute that will result in the smallest expected size of the sub-trees.

Objects were classified using 10 decision trees, by means of the boosting multi-classifier method, which allows for increasing the accuracy of the classifier (Freund, 1995). The methodology followed by the boosting to build the multi-classifier is based on the assignment of weights to training samples (Freund and Shapire, 1997). The higher the weight of a sample, the higher its influence in the classifier. After each tree construction, the weights vector is adjusted to show the model performance. In this way, samples erroneously classified increase their weights, whereas the weights of correctly classified samples are decreased. Thus, the model obtained in the next iteration will give more relevance to the previously wrongly classified samples (Hernandez-Orallo et al., 2004). After the decision tree set is constructed, the class assigned to an object will be done considering the estimated error made in the construction of each tree. The lower the estimated error e, the higher the weight given to a tree, according to the formula:

$$-\log\left(\frac{e}{1-e}\right)$$

The sum of the weights of those trees which assign the same class to one object is computed, giving that object the class with higher value.

4. RESULTS

Seven classification tests were performed. In the first one, only spectral features were included. In successive tests, land use information with different updating degrees was added. Overall accuracies obtained for each test are shown in Table 1. The accuracy reached without considering ancillary data is 77.3%. The addition of ancillary data produces a continuous increase of the overall accuracies, up to 93% for an updating degree of 90%. The addition of the land use information contained in the geospatial database has, in any case, a negative effect on the overall classification accuracy, even when the information provided is mostly erroneous, such as in the cases of updating degrees 40% and 50%.

Analyzing the producer's and user's accuracies (Table 1), it is perceptible how these values are generally increasing as the information about the land use which is added into the classification is more up-to-date. This increment is especially high in those classes which present low accuracy values in the spectral classification. The producer's and user's accuracy values indicate that the addition of ancillary information with a certain updating degree ensures that the classes will be classified with an accuracy at least similar to that updating degree. Therefore the addition of ancillary land use information into the classification does not have a negative effect on the user's and producer's accuracies of the classes with high accuracy values. The addition provides significant information in those classes with low user's and producer's accuracies classifying with other descriptive features.

Updating degree	Overall accuracy	Citrus orchards		Young citrus orchards		Buildings		Forest		Carob-trees		Irrigated crops		Shrub lands		Roads		Arable lands	
		PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
no database	77.3	57	75	64	72.7	82	84.5	95	93.1	83	57.6	96	85.7	94	96.9	45	55.6	80	77.7
40%	78	57	72.2	63	75	86	84.3	95	94.1	80	58.4	95	88	94	96.9	50	62.5	82	73.2
50%	79.7	59	75.6	76	73.8	86	86	95	94.1	80	61.1	95	87.2	94	96.9	48	63.2	84	80
60%	84.7	71	74	77	73.3	88	90.7	95	96	84	75	97	89	97	98	69	81.2	84	85.7
70%	85.0	71	79.8	79	84.9	88	85.4	95	93.1	83	76.9	97	82.9	93	97.9	69	79.3	90	84.9
80%	90.4	78	94	87	89.7	94	91.3	95	90.5	94	85.5	98	90.7	97	95.1	79	90.8	92	87.6
90%	93.0	88	91.7	92	96.8	94	96.9	97	91.5	93	91.2	98	90.7	98	97	87	90.6	90	90.9

Table 1. Overall, Producer's (PA) and user's (PU) accuracies of classifications for each database updating degree (%).

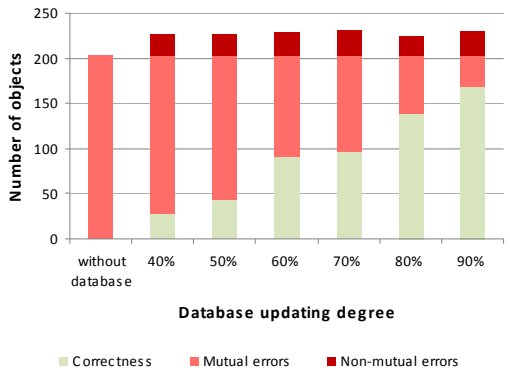

Figure 2. Comparison and development of classification errors as database updating degree increases.

	Database updating degree					
	40%	50%	60%	70%	80%	90%
Decision tree 1	73	73	171	281	450	450
Decision tree 2	0	143	141	305	0	0
Decision tree 3	0	0	0	132	0	450
Decision tree 4	0	0	261	0	450	0
Decision tree 5	0	266	0	450	0	299
Decision tree 6	0	0	127	34	0	0
Decision tree 7	0	0	171	0	450	450
Decision tree 8	0	0	132	450	0	0
Decision tree 9	0	156	65	153	0	450
Decision tree 10	0	159	188	118	450	302
Average	7.3	79.7	125.6	192.3	180	240
Use percentage	1.6%	17.7%	27.9%	42.7%	40.0%	53.4%

Table 2. Number of training objects classified considering ancillary data as a feature, average influence in absolute terms and as a percentage of the total.

In Figure 2 the errors obtained in the classification without ancillary information as a descriptive feature against the errors obtained in the classifications considering this feature with different updating degrees are compared. Two types of errors were defined: mutual and non-mutual errors. Mutual errors represent the objects erroneously classified whether or not ancillary data are used. Non-mutual errors refer to the new errors committed due to the addition of ancillary data as a descriptive feature. From 900 evaluation objects, 204 were erroneously classified without considering the database information. Adding this feature with an updating degree of 40%, 175 errors shared with the previous classification are committed (mutual errors), meanwhile 29 errors are corrected and 23 new errors are produced (non-mutual errors). As the geospatial database updating degree raises, the mutual error rate decreases, which involves an increase of the objects correctly classified. Non-mutual error value in all the cases remains about 10%. The total number of errors made in each classification can be obtained by adding the mutual and non-mutual errors.

As mentioned in section 3.3, one of the main advantages of the C5.0 classifier lies in its potential to select features according to their ability to divide training samples into homogeneous

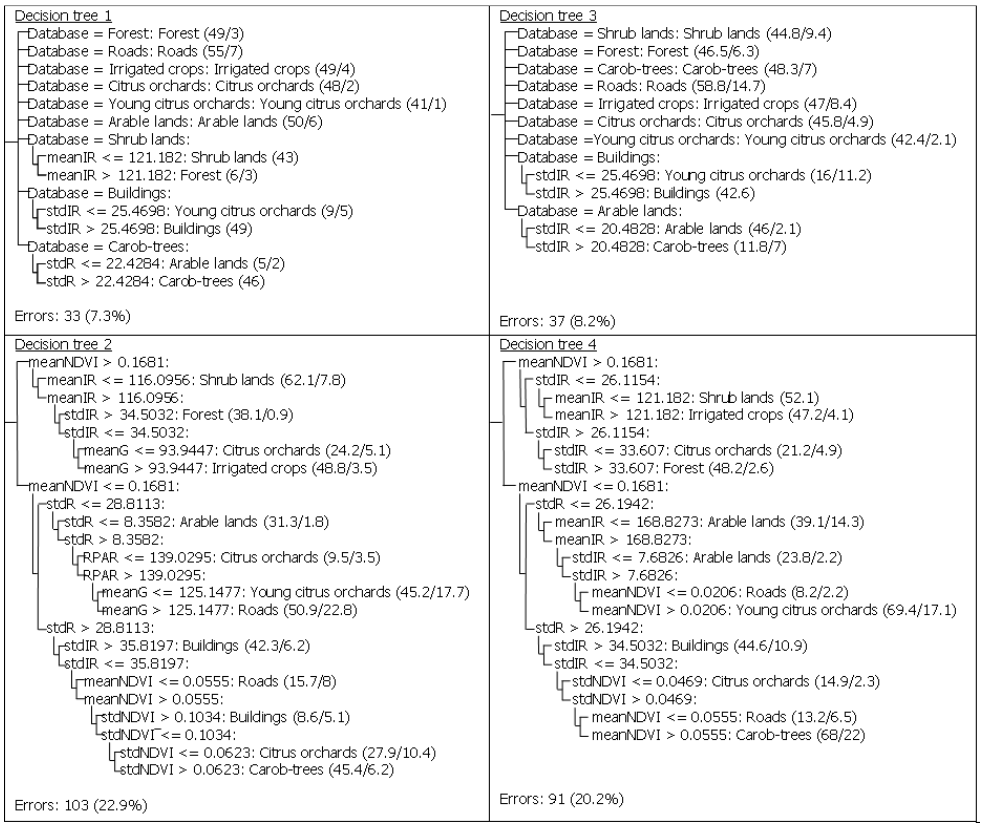

Figure 4. First four decision trees generated when the updating degree of ancillary data is 90%. The feature *Database* represents the land use contained in the database; *meanIR*, *meanR*, *mean* G, *meanNDVI* and *stdIR*, *stdR*, *stdG* and *stdNDVI* are respectively the mean and standard deviation values of objects for infrared, red, green and NDVI bands.

subgroups. Actually, if a feature is less efficient for separating the training samples, it will be employed less frequently in the construction of decision trees. The influence of the land use within the database on the classification accuracy was computed for each of the six tests that employed this feature. For this purpose, the number of training objects classified according to a condition that considered the ancillary data was calculated for each tree. Since the boosting multi-classifier method was used, the average influence of this feature in the ten decision trees was computed as the mean value of the number of training objects classified considering this feature for each tree (Table 2). When a database updating degree of 40% was used, the participation of this feature was extremely low. Only 1.6% of the training objects were classified using a rule which considered this characteristic.

As the database updating degree increases, the percentage of objects classified with this feature grows to a maximum value of 53.3%. In the last test, where the updating degree of the database is 90%, this feature is actually used in six of the ten defined trees, and it is used in the first rule in four of them. When the boosting multi-classifier methodology is used, the first tree generated is usually the one which better adapts itself to the training samples. For an updating degree of 90% (see Figure 4), the first feature employed to create the tree is the land use contained in the database, which means that this is the feature with a higher discriminant level. In the construction of the second decision tree, the weight given to the training samples erroneously classified in the first tree is much higher than the one given to those classified correctly. Subsequently, the second tree is focused on the classification of the samples with non updated land use, so only spectral features are employed for constructing this tree. The second decision tree makes 22.9% assignation errors. Third decision tree recalculates the weights of the training samples based on the classification errors of the second tree, obtaining comparable weight values for samples erroneously and correctly classified. As each training sample has similar weights, the land use feature presents again the highest discriminant capacity and it is used as the most relevant feature to build the third decision tree. This is repeated in the generation of the ten decision trees, where the land use feature is alternately used to produce accurate and non-accurate trees.

5. CONCLUSIONS

This document presents a study of the effect of the use of ancillary data as a descriptive feature for land use classification accuracy. Different updating degrees of this information have been tested. The results show that, in all cases, the addition of land use information increases the overall accuracy of the classification. It is particularly interesting that the inclusion of non updated land use information does not produce a decrease in this accuracy. The user's and producer's accuracies obtained for each class are at least similar to the updating degree of the information contained in the geospatial database.

The use of alphanumeric information as descriptive features requires classifiers that are able to manage this type of data. The C5.0 algorithm allows us to include discrete information and to adequately choose those features which provide a higher separability. The relevance of the land use feature in the construction of the rules is directly related to the updating degree of the database: the less this feature contributes to the separability, the less it participates in the classification rules.

REFERENCES

Berberoglu, S., Curran, P.J., 2004. Merging Spectral and Textural Information for Classifying Remotely Sensed Images. In: *Remote Sensing Image Analysis: Including the Spatial Domain*. Kluwer Academic Publishers, pp. 113-136.

Blaes, X., Vanhalle, L., Defourny, P., 2005. Efficiency of crop identification based on optical and SAR image time series. Remote Sensing of Environment, 96(3-4), 352-365.

Bruzzone, L., Conese, C., Maselli, F., Roli, F., 1997. Multisource Classification of Complex Rural Areas by Statistical and Neural-Network Approaches. Photogrammetric Engineering & Remote Sensing, 63(5), pp. 523-533.

Cohen, Y., Shoshany, M., 2000. Integration of remote sensing, GIS and expert knowledge in national knowledge-based crop recognition in Mediterranean environment. International Archives of Photogrammetry and Remote Sensing, v. XXXIII, Part B7, pp. 280-286.

Heipke, C., Pakzad, K., Straub, B.M., 2000. Image analysis for GIS data acquisition. Photogrammetric Record, 16(96), pp. 963-985.

Heipke, C., Straub, B.M., 1999. Towards the automatic GIS update of vegetation areas from satellite imagery using digital lanscape model as prior information. IAPRS. 32 - Part 3-2W5, 167-174.

Hernández-Orallo, J., Ramírez-Quintana, M.J., Ferri-Ramírez, C., 2004. *Introducción a la minería de datos*. Pearson Prentice Hall. Madrid.

Hoffer, R.M., Staff, 1975, Natural resource mapping in mountainous terrain by computer analysis of ERTS-1 Satellite Data. LARS Information Note 061575, Purdue University, Indiana.

Huang, X., Jensen, J.R., 1997. A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIs Data. Photogrammetric Engineering & Remote Sensing. 63(10), pp. 1185-1194.

Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement. Photogrammetric Engineering & Remote Sensing, 48(1), pp. 123-130.

Janssen, L.L.F., Middelkoop, H., 1992. Knowledge-based crop classification of a Landsat Thematic Mapper image. International Journal of Remote Sensing, 13(15), pp. 2827-2837.

Katila, M. & Tomppo, E. 2002. Stratification by ancillary data in multisource forest inventories employing k-nearest-neighbour estimation. Canadian Journal of Forest Research 32(9), pp. 1548-1561.

Lawrence, R.L., Wright, A., 2001. Rule-based classification systems using classification and regression tree (CART) analysis. Photogrammetric Engineering & Remote Sensing, 67(10), pp. 1137-1142.

Leránoz, A., Albizua, L., Zalba, M., 2007. Nueva metodología de estimación de superficies de cultivos. Actas del XII

Congreso de la Asociación Española de Teledetección, pp. 46-51.

Freund, Y., 1995. Boosting a weak learning algorithm for majority. Information and Computation, 121(2), pp. 256-285.

Freund, Y., Shapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), pp. 119-139.

Maselli, F., Conese, C., De Filippis, T., Romani, R., 1995. Integration of ancillary data into a maximum-likelihood classifier with nonparametric priors. ISPRS Journal of Photogrammetry and Remote Sensing, 50(2), pp. 2-11.

Olsen, B.P., Knudsen, T., Frederiksen, P., 2002. Digital Change detection for map database update. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXIV, part 2, pp. 357-363.

Pakzad, K., 2002. Knowledge based multitemporal interpretation. ISPRS Commission III Symposium. 34, pp. 234-239.

Pedroni, L., 2001. Discriminación de diferentes tipos de bosque tropical mediante imágenes de satélite y datos auxiliares. Revista Forestal Centroamericana, 34, pp. 12-18.

Quinlan, J.R., 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishing, San Francisco.

Raclot, D., Colin, F., Puech, C., 2005. Updating land cover classification using a rule-based decision system. International Journal of Remote Sensing, 26(7), pp. 1309-1321.

Rogan, J., Miller, J., Stow, D., Frankling, J., Levien, L., Fischer, C., 2003. Land cover change mapping in California using classification trees with multitemporal Landsat TM and ancillary data. Photogrammetric Engineering & Remote Sensing. 69(7), pp. 793-804.

Ruiz, L., Recio, J., Hermosilla, T., 2007. Methods for automatic extraction of regularity patterns and its application to object-oriented image classification. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVI, Part 3/W 49A: 117–121.

Skidmore, A.K., 1989. Expert system classifies eucalypts forest types using thematic mapper data and a digital terrain model. Photogrammetric Engineering & Remote Sensing, 55(10), pp. 1449-1464.

Strahler, A.H., Logan, T.L., Bryant, A., 1978. Improving forest cover classification accuracy from Landsat by incorporating topographic information. Proceedings of the 12th International Symposium on Remote Sensing of Environment, 2, pp. 927-942.

Treltz, P., Howarth, P., 2000. Integrating Spectral, Spatial, and Terrain Variables for Forest Ecosystem Classification. Photogrammetric Engineering & Remote Sensing, 66(3), pp. 305-317.

Walter, V., 2000. Automatic change detection in GIS databases based on classification of multispectral data. International Archives of Photogrammetry and Remote Sensing XXXIII Part B4, pp. 1138-1145.

Zhang, S., Liu, X., 2004. Realization of Data Mining Model for Expert Classification Using Multi-Scale Spatial Data. Proc. of ISPRS Workshop on Service and Application of Spatial Data Infrastructure, 26(4/W6), pp. 107-111.

ACKNOWLEDGEMENTS

The authors appreciate the financial support provided by the Spanish *Ministerio de Ciencia e Innovación* and the FEDER in the framework of the Projects CTM2006-11767/TECNO and CLG2006-11242-C03/BTE. We also thank the Spanish *Instituto Geográfico Nacional* for the support provided to carry out this research.