Using Remote Data Mining on LIDAR and Imagery Fusion Data to Develop Land Cover Maps

Jorge García-Gutiérrez[1], Francisco Martínez-Álvarez[2], and José C. Riquelme[1]

[1] Department of Computer Science, University of Seville, Spain
{jgarcia,riquelme}@lsi.us.es
[2] Area of Computer Science, Pablo de Olavide University, Spain
fmaralv@upo.es

Abstract. Remote sensing based on imagery has traditionally been the main tool used to extract land uses and land cover (LULC) maps. However, more powerful tools are needed in order to fulfill organizations requirements. Thus, this work explores the joint use of orthophotography and LIDAR with the application of intelligent techniques for rapid and efficient LULC map generation. In particular, five types of LULC have been studied for a northern area in Spain, extracting 63 features. Subsequently, a comparison of two well-known supervised learning algorithms is performed, showing that C4.5 substantially outperforms a classical remote sensing classifier (PCA combined with Naive Bayes). This fact has also been tested by means of the non-parametric Wilcoxon statistical test. Finally, the C4.5 is applied to construct a model which, with a resolution of 1 m^2, obtained precisions between 81% and 93%.

Keywords: Data mining, remote sensing, LIDAR, imagery, LULC.

1 Introduction

Remote sensing has been a very important tool to study the natural environment for long. It has been applied to lots of different tasks such as species control [1] or landscape control [2]. These techniques are of the utmost importance to reduce costs since they increase the products development speed, helping thus the experts to make decisions. Due to the remote sensing relevance, many researchers have invested much time to find how to develop new algorithms to improve the quality of the results.

One of the most important products in remote sensing is land use and land cover maps (LULC). They are used to develop policies in order to protect specially interesting areas from both environmental and economic points of view. The automatic generation is a very desirable feature since they are generated for covering large zones. To efficiently solve this problem, supervised learning is usually applied from a small quantity of data previously classified by human experts [3]. Hence, multispectral images, hyperspectral images and orthophotography have been widely used in many applications although they have their own

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part I, LNAI 6096, pp. 378–387, 2010.

limitations, e.g., the most useful information is usually on the visible spectrum band which is easily affected by shadows. The apparition of new sensors has caused that a new research line appears, which tries to overcome imagery problems fusing them with new technologies. LIDAR (LIght Detection And Ranging) is one of these new sensors. It is an optical technology that measures properties of scattered light in the near infrared to find range and/or other information of a distant target. The method to determine distance to an object or surface is to use laser pulses. The range to an object is determined by measuring the time delay between transmission of a pulse and detection of the reflected signal. In this way, LIDAR is able to extract the heights of objects and its fusion with imagery boosts any remote sensing technique so that it has been exploited with several purposes [4], [5].

Fusion among sensors increases data size. In this context, intelligent techniques are a must if an automatical process is required and particulary, remote data mining is a very suitable tool to deal with problems associated to big size data. With this in mind, some authors have started to use intelligent techniques like artificial neuronal networks (ANN) [6] or support vector machines(SVM) [7]. But the most used method is still based on classical statistics methods and concretely, the traditional principal component analysis (PCA) and the application of maximum likelihood principle [8], i.e., a Naive Bayes classifier.

In this work, a supervised method to obtain LULC automatically is shown and a comparison between two techniques for supervised learning in remote sensing is established with two purposes:

- Show the quality of models when intelligent techniques are applied on LIDAR and imagery fusion data.
- Show the importance of using well-known machine learning algorithms in the remote sensing context that outperforms most classical statistics procedures.

The rest of the paper is organized as follows. Section 2 provides an exhaustive description of the data used in this work. Section 3 describes the methodology used, highlighting the feature selection and model extraction processes. The results achieved are shown in Section 4 and, finally, Section 5 is devoted to summary the conclusions and to discuss future lines of work.

2 Data Description

The study area is located in the north of Galicia (Spain) as depicted in Figure 1. It is basically composed by a small residential zone and a forest zone, whose dominant species is *Eucalyptus Globulus*. In geomorphologic terms, in spite of the altitudes varying between 230 and 370 meters, the relief of the zone is quite accentuated.

The LIDAR data were acquired in November 2004 with *Optech's ALTM 2033* from a flight altitude of 1500 meters. The LIDAR sensor works with a laser wavelength of 1064 nm and the beam divergence was set to 0.3 mrad. The pulsing frequency was 33 kHz, the scan frequency 50 Hz, and the scan angle

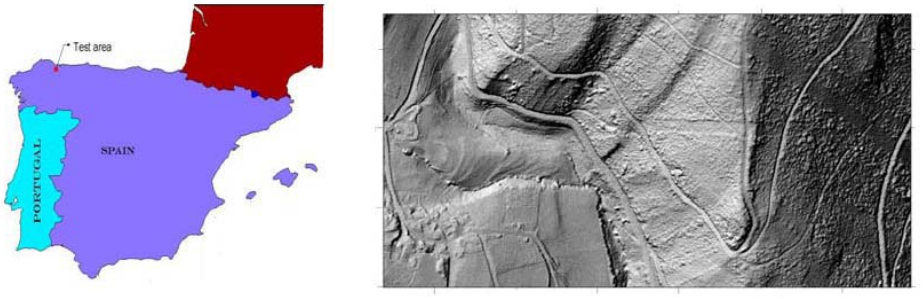

Fig. 1. Study area location and digital elevation model (DEM) used in this work

10 degrees. The first and last return pulses were registered. The complete study area was flown in 18 strips and each strip was flown three times, which gave an average measuring density of about four points per square meter.

A digital elevation model (DEM) was extracted from the LIDAR data using an adaptative morphologic filter method [9] to rectify object heights. The resulted DEM is illustrated in Figure 1. Moreover, a previous orthophoto is used to extract features from the visible spectrum band. It was taken with a resolution of 0.5 meters of the same zone and with similar atmospheric conditions to the moment of LIDAR acquisition flight.

Leaning on the orthophoto and with the help of previous knowledge about the study zone, a training base formed by 5570 pixels was selected (5% out of total, approximately) and classified into 5 different classes: road, farming land or bare earth, middle vegetation, high vegetation and buildings. In addition, in the same previous study, 317 specially interesting pixels were selected and classified to make up a hold-out test base according to the traditional testing in remote sensing.

3 Methodology

In order to classify LULC, a general method based on remote data mining techniques is applied. First, a raster matrix is built with a resolution of 1 meter. Every raster cell represents a squared meter pixel of the study area which contains several different measures extracted from LIDAR data (based on signal intensity, distribution and height) and images (visible spectrum bands). Then, a feature selection phase reduces the total number of features to build a supervised learning model. Later, an intelligent technique is used to generate a model. In this work, two different classifiers are used to make a comparison: a C4.5 decision tree [10] and a classical principal components analysis (PCA) combined with a Naive Bayes classifier which uses the maximum likelihood principle, whose combination was first used in [11]. An overall view of the whole classification process will be described in detail in the following subsections.

3.1 Feature Selection

All generated pixels have 63 different features based on LIDAR data and visible spectrum bands, which are in the geographical zone limited by the pixel itself. These features can be classified as intrapixel or interpixel. The intrapixel features are those which are calculated with data found within a pixel, whilst the interpixel features are those which are characterized as defining a relation between each pixel and its eight adjacent neighbors. With these features, characterization of the terrain is attempted, formalizing the visual differences or morphologies of the different classes. Most of them have been extracted from literature [12] and classical remote sensing applications. However, some are original of this work like SNDVI and EMP.

The Normalized Difference Vegetation Index (NDVI) is a simple numerical indicator that can be used to analyze remote sensing measurements and assesses whether the target being observed contains live green vegetation or not. In Equation 1, NDVI is calculated from the red band value (R) and the near infrared band value (NIR) which is not in the visible spectrum.

$$NDVI = \frac{NIR - R}{NIR + R} \tag{1}$$

In this work, a new attribute SNDVI (Simulated NDVI) has been generated using the intensity (I) from LIDAR as near-infrared value which approximates the real NDVI value, which cannot be calculated since no NIR information is available in LIDAR data. This parameter is calculated as follows:

$$SNDVI = \frac{I - R}{I + R} \tag{2}$$

LIDAR point density is another important characteristic to be analyzed. EMP feature counts the number of empty pixels that surrounds the current pixel in a eight-adjacent neighborhood. It helps to detect water areas because most of laser energy does not reflect on water and responses from flooded zones are not registered by the sensor.

Features from pixels in the training set are submitted to a process of selection. In this case, a correlation based feature subset selection (CFS) has been applied. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy among them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. The 21 final selected predictor variables after CFS execution are listed in Table 1.

3.2 Model Extraction

The next phase consists in executing the classification algorithms. Two kinds of approaches are proposed to extract the model: the C4.5 algorithm [10] and a combination between a PCA and a Naive Bayes classifier [11]. The two models

Table 1. Twenty-one candidate predictor variables after the feature selection phase

Variable	Description	Type
SNDVIMIN	Simulated NDVI minimum	Intrapixel
SNDVIMEAN	Simulated NDVI mean	Intrapixel
SNDVISTDV	Simulated NDVI standard deviation	Intrapixel
IMIN	Intensity minimum	Intrapixel
IMAX	Intensity maximum	Intrapixel
IMEAN	Intensity mean	Intrapixel
HMIN	Height minimum	Intrapixel
HMAX	Height maximum	Intrapixel
HMEAN	Height mean	Intrapixel
HSTD	Height standard deviation	Intrapixel
HCV	Height coefficient of variation	Intrapixel
PEC	Penetration coefficient	Intrapixel
PCT31	Percentage third return out of first return	Intrapixel
IRVAR	Intensity red band variance	Intrapixel
IRMEAN	Intensity red band mean	Intrapixel
IGVAR	Intensity green band variance	Intrapixel
IGMEAN	Intensity green band mean	Intrapixel
IGSKE	Intensity green band skewness	Intrapixel
IGKURT	Intensity green band kurtosis	Intrapixel
SLP	Slope	Interpixel
EMP	Empty LIDAR surrounding pixels	Interpixel

are extracted by means of the data mining environment WEKA [13]. For the C4.5 execution, the J48 implementation from Weka is selected. In both cases, the method is executed with the parameters set as default by WEKA.

Although both methods have similar computational costs, there are significant differences when each one expresses the model. On the one hand, decision trees are more intuitive for non-experts users and for this reason has been widely used in many different industrial and engineer applications and even in remote sensing. On the other hand, when the training set is more complex, generated trees are more difficult to read because branches and nodes are increased in number whilst a PCA and Naive Bayes approach maintains a more constant number of attributes and levels.

Bearing in mind, an analysis of functional aspects such as the precision on the results is necessary to know which one is more suitable for this study zone.

4 Results

In order to evaluate how much improvement can be achieved with each intelligent technique application, two different kinds of testing were carried out. First, a cross-validation on training data. Second, a hold-out testing on specially interesting pixels which is the common testing in classical remote sensing. The training base and the test set were extracted as referred in Section 2.

4.1 Comparison of Methods

After the execution of the 10-fold cross-validation tests, it can be stated that the C4.5 decision tree obtains better results. In Tables 2 and 3, the total and partial precisions as well as the kappa index of agreement (KIA) obtained for the two techniques on the 5570 training pixels are shown. The two techniques obtain high accuracy, but the application of the decision tree produces an improvement of almost fifteen percentile points. Furthermore, the potential of decision trees to indicate which features in the original set are more interesting –which allows a new automatic selection of attributes– must be highlighted.

It is well-known that just a cross-validation result or a hold-out test and a rank method among techniques is not enough to confirm than a technique is better than another or viceversa, even when many authors have used this technique in their studies. For this reason, to evaluate the statistical significance of the measured differences in algorithm ranks, a procedure suggested in several works [14] for robustly comparing classifiers across multiple datasets is used. In this work, there is only one dataset because LIDAR data has high costs to be

Table 2. Cross-validation summary of the tests on C4.5 and confusion matrix

User class \ C4.5	Roads	Farming lands	Middle vegetation	High vegetation	Buildings
Roads	219	13	13	2	0
Farming lands	9	2111	81	8	0
Middle vegetation	10	80	1389	77	14
High Vegetation	6	8	61	1130	3
Buildings	1	0	13	10	312
Producer's accuracy	0.898	0.954	0.892	0.921	0.948
User's accuracy	0.887	0.956	0.885	0.935	0.929
Total accuracy	0.927				
KIA	0.897				

Table 3. Cross-validation summary of the tests on PCA + Naive Bayes and confusion matrix

User class \ PCA + NV	Roads	Farming lands	Middle vegetation	High vegetation	Buildings
Roads	212	13	10	1	11
Farming lands	4	2073	107	25	0
Middle vegetation	10	639	696	169	56
High Vegetation	2	7	144	1034	21
Buildings	23	0	8	14	291
Producer's accuracy	0.845	0.759	0.721	0.832	0.768
User's accuracy	0.858	0.938	0.443	0.856	0.866
Total accuracy	0.773				
KIA	0.677				

obtained. So, the training set is randomly split in five subsets. Then, a 10-fold cross-validation is made for every subset. At the end, there are 50 measures for every algorithm and then, the procedure is carried out. Hence, it is possible to use average ranks that provide a fair comparison of the algorithms analyzing the 50 measures, revealing that, on average, C4.5 ranks first. Given the measured average ranks, it is possible to apply a Wilcoxon test to check whether the ranks are significantly different to consider them different populations or not (which is the expected under the null hypothesis). Leaning on a statistical package (MATLAB), p value for the Wilcoxon test has resulted on a value of $6.7266e-18$ so the null hypothesis is rejected having found that the ranks are significantly different (at $\alpha = 0.05$).

4.2 Model Precision

Once the model developed by C4.5 is extracted, it is applied to every pixel. Then, a new empirical test is done. In this way, a checking process on the test set (317 specially interesting pixels) is carried out and classes which they pertain are evaluated. In Tables 4 and 5, it is possible to observe the results of the test through the confusion matrix, producer and user's accuracies, and the kappa estimator.

Table 4. Hold-out summary on C4.5 and confusion matrix

User class \ C4.5	Roads	Farming lands	Middle vegetation	High vegetation	Buildings
Roads	14	1	0	0	0
Farming lands	2	106	9	0	0
Middle vegetation	3	5	56	15	0
High Vegetation	1	0	10	65	8
Buildings	1	0	3	1	17
Producer's accuracy	0.667	0.946	0.718	0.802	0.68
User's accuracy	0.9333	0.906	0.709	0.774	0.773
Total accuracy	0.814				
KIA	0.7457				

Since the pixels in the test set had been selected for the special difficulty shown to be classified, both methods decrease their global precision. Anyway, C4.5 improves the results by PCA and Naive Bayes in almost 8% which is still a remarkable difference. Moreover, some classes decrease their accuracy dramatically when using PCA and Naive Bayes approach which is inadmissible when dealing with so few classes.

In Figure 2, the resulted global classification next to the training base and the initial input data (orthophoto and LIDAR intensity image) for a C4.5 decision tree model are shown. As it can be appreciated, the LULC map accuracy is very high and its automatic generation is much faster than a manual creation by an expert.

Table 5. Hold out summary on PCA + Naive Bayes and confusion matrix

User class \ PCA + NV	Roads	Farming lands	Middle vegetation	High vegetation	Buildings
Roads	14	1	0	0	0
Farming lands	1	101	13	2	0
Middle vegetation	1	24	29	24	1
High Vegetation	0	1	7	74	2
Buildings	4	0	0	3	15
Producer's accuracy	0.7	0.795	0.591	0.718	0.833
User's accuracy	0.933	0.863	0.367	0.881	0.682
Total accuracy	0.735				
KIA	0.632				

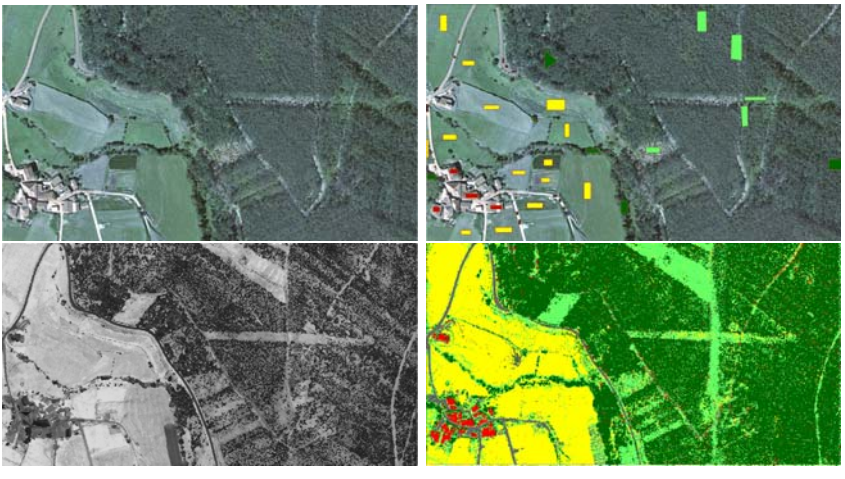

Fig. 2. From up to down and left to right: original orthophoto, training set, LIDAR intensity image and final result. Classes are colored as: urban areas in red, roads in dark grey, farming lands or naked earth in yellow, medium vegetation in light green and high vegetation in dark green.

5 Conclusions and Future Work

An approach based on LIDAR and imagery fusion data as well as the application of intelligent techniques have been tested to classify land coverage of a typical area of the north of Spain in this work. The objective was to evaluate data fusion capabilities and to establish a comparison between two different techniques to classify the fusion data: a classical statistics principal components analysis and a Naive Bayes classifier (based on the maximum likelihood principle) and C4.5, which is a well-known machine learning decision tree generator.

The developed method is based on a pixel-oriented focus which classifies raw data in five different classes. In this way, a series of features were calculated from fusion data (some of them are original in this work), which are associated with each pixel. Thereafter, an attribute selection method was applied to reduce the set of variables to consider. Lastly, the selected classifier extracted a model.

A thorough study was performed to select which algorithm was the best between the two possible ones. The tests carried out selected the algorithm C4.5, which generated a decision tree, as the model with the best fit. The results between 82% and 93% of accuracy depending on the kind of test applied showed that C4.5 outperforms its rival in between eight and fifteen percentile points. They also have demonstrated that different types of terrain can be characterized using intelligent techniques in a multi-staged process using LIDAR and image data and, moreover, robust and well-known machine learning algorithms are perfectly suitable to improve classification results in remote sensing data better than more classical methods in our study zone conditions.

In relation to future works, two problems arise. The first problem is the improvement of the classification method itself, as some problems have already been detected. This can be solved using several techniques joined in ensembles. The second one is related to dependence of results on training set. Some authors have detected problems when a classifier is not trained by well-balance or real data and its effects on the test phase. In addition, outliers (salt and pepper noise) are a great problem due to the imprecise training set definition or intrinsic problems on data, e.g., variability of LIDAR intensity depending on the number of returns per pulse. These problems are much more common than it would be desirable. There are two possible ways to explore in order to solve this problem. On the one hand, the easier option is to introduce an instance selection phase. On the other hand, migrating from a supervised learning to a semi-supervised learning can be another more interesting option in which the system helped by evolutive computation would extract its own training data just taking the number of classes from the user.

Acknowledgments

We would like to thank the Land, Underground and Biodiversity Laboratory of University of Santiago de Compostela for all the support received in the development of this work and especially, to thank Luis Gonçalves-Seco, for all the time he invested that allowed this work to be completed.

References

1. Hyde, P., Dubayah, R., Walker, W., Blair, J.B., Hofton, M., Hunsaker, C.: Mapping forest structure for wildlife habitat analysis using multi-sensor (LIDAR, SAR/INSAR, ETM+, Quickbird) synergy. Remote Sensing of Environment 102, 63–73 (2006)

2. Wang, C., Glenn, N.F.: Integrating LIDAR intensity and elevation data for terrain characterization in a forested area. IEEE Geoscience and Remote Sensing Letters 6(3), 463–466 (2009)
3. Johansen, K., Coops, N.C., Gergel, S.E., Stange, Y.: Application of high spatial resolution satellite imagery for riparian and forest ecosystem classification. Remote Sensing of Environment 110, 29–44 (2007)
4. Suárez, J.C., Ontiverosa, C., Smith, S., Snapec, S.: Use of airborne LIDAR and aerial photography in the estimation of individual tree heights in forestry. Computers & Geosciences 31, 253–262 (2005)
5. Koetz, B., Sun, G., Morsdorf, F., Ranson, K., Kneubuhler, M., Itten, K., Allgower, B.: Fusion of imaging spectrometer and LIDAR data over combined radiative transfer models for forest canopy characterization. Remote Sensing of Environment 106, 449–459 (2007)
6. Nguyen, M.Q., Atkinson, P.M., Lewis, H.G.: Superresolution mapping using a hopfield neural network with LIDAR data. IEEE Geoscience and Remote Sensing Letters 2(3), 366–370 (2005)
7. Mazzoni, D., Garay, M.J., Davies, R., Nelson, D.: An operational MISR pixel classifier using support vector machines. Remote Sensing of Environment 107, 149–158 (2007)
8. Bork, E.W., Su, J.G.: Integrating LIDAR data and multispectral imagery for enhanced classification of rangeland vegetation: A meta analysis. Remote Sensing of Environment 111, 11–24 (2007)
9. Goncalves-Seco, L., Miranda, D., Crecente, R., Farto, J.: Digital terrain model generation using airborne LIDAR in florested area of Galicia, Spain, Lisbon, Portugal. In: Proceedings of 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, pp. 169–180 (2006)
10. Quinlan, J.R.: Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research 4, 77–90 (1996)
11. Mutlu, M., Popescu, S.C., Stripling, C., Spencer, T.: Mapping surface fuel models using LIDAR and multispectral data fusion for fire behavior. Remote Sensing of Environment 112, 274–285 (2008)
12. Hudak, A.T., Crookston, N.L., Evans, J.S., Halls, D.E., Falkowski, M.J.: Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LIDAR data. Remote Sensing of Environment 112, 2232–2245 (2008)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1), 10–18 (2009)
14. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)