Automatic detection of high-voltage power lines in LiDAR surveys using data mining techniques

Chasco-Hernández Daniel[1], Sanz-Delgado José Antonio[2], García-Morales Víctor[3], Álvarez-Mozos Jesús[1]*

1) Department of Engineering, Public University of Navarre, Los Tejos, Campus Arrosadia, 31006, Pamplona, Spain
2) Department of Statistics, Computer Science and Mathematics, , Public University of Navarre, Los Pinos, Campus Arrosadia, 31006, Pamplona, Spain
3) Department of Engineering and Territorial Systems, TRACASA, Cabárceno 6, 31621, Sarriguren, Spain
*Corresponding author: jesus.alvarez@unavarra.es

Abstract

The correct classification of power lines in LiDAR point clouds has attracted the interest of the mapping community in the last years. The objective of this research is the detection and automatic extraction of high-voltage transmission lines from LiDAR data using data mining techniques. With this aim, a Single Photon LiDAR (SPL) survey acquired over the region of Navarre (Spain) in 2017 was used, with a mean point density of 14 pt/m^2. Different data mining techniques were evaluated, including decision trees (C4.5 and CART) and ensemble learning algorithms (Random Forests, Bagging and AdaBoost). Fifteen test sites were studied corresponding to areas with high-voltage power lines over different conditions regarding the underlying vegetation and topography. For these sites 92,104 LiDAR points were identified as power lines and more than 4M points as not power lines using existing cartography. This dataset was randomly split in train and test sets and then balanced two obtain a similar amount of data for the two classes. The results obtained show the importance of balancing the training data with improvements in accuracy of ~10% with respect to the imbalanced case. Accuracies higher than 87% were obtained in all balanced cases, with particularly successful results for ensemble learning techniques, being AdaBoost the technique with the highest accuracy 91%. These results suggest that the combination of SPL surveys and data mining tools can be successfully used for the operational mapping of high voltage power lines.

Keywords: LIDAR, data mining, Single Photon Lidar, power lines, supervised classification

1 Introduction

Power line mapping is a costly task for the organizations in charge of making cartography [1]. Therefore, methods to automatically identify and map power lines are required in order to update existing maps, create new ones for areas where such maps do not exit, and hence for enhancing the efficiency of cartographic organizations.

Since its invention, Light Detection And Ranging (LiDAR) technology has had a dramatic impact in the cartographic methods and workflows. Indeed, airborne LiDAR is an efficient and cost-effective technique for the rapid and precise acquisition of massive 3D point clouds [2]. Furthermore, it enables the automation of the surveying process and the production of maps with a level of detail unattainable with previous cartographic techniques [3]. Before the onset of LiDAR, both topographic surveys and photogrammetric restitution processes required an enormous amount of expert hand labor, so producing detailed scale cartography was very costly [4]. With LiDAR the production of high resolution Digital Terrain Models (DTM) and cartography can be done much faster, cheaper and in a more objective (with less human intervention) and precise manner.

Data mining is a set of techniques and approaches designed to extract knowledge and patterns from large datasets [5], so they are ideal tools to efficiently process LiDAR datasets. The aim of this work is to apply state of the art data mining algorithms to automatically identify high voltage power lines in LiDAR point clouds, so as to obtain a classified point cloud that can produce a power line map. To attain this objective a Single Photon LiDAR (SPL) dataset was processed, which was acquired over the region of Navarre (Spain) in 2017. SPL is a very recent and innovative technology with significant improvements in terms of the number of points acquired. In the next sections, the materials and methods used are described and the obtained results are shown, finally some conclusions and recommendations are outlined.

2. Materials and methods

An airborne LiDAR survey was carried out in Navarre (Spain) between September and November 2017 using the Leica SPL100 sensor. The SPL technology splits the laser beam into a 10x10 array achieving a much higher point density than its predecessors. Some other key differences with the more conventional LiDAR systems used so far are its operation wavelength at 532 nm, the possibility to record as many as 10 returns per laser pulse and its faster recovery time (1.6 ns). On the other hand its circular scanning pattern creates a very irregular point distribution with a higher point proportion at the borders of the swath, and the obtained point clouds are very noise with a large amount of noise points <50m above the ground. For this project, a flight height of 3,900-6,300 m(asl) was set with an airspeed of 200 knots, leading to a swath width of 2.3 km and an average point density of 14 pt/m^2. The complete survey over Navarre comprised more than 500 billion points and ~50 TB of data, and was freely distributed as 1x1km tiles in LAS 1.4 format (ftp://ftp.cartografia.navarra.es). The

LAS files contain information on UTM X, Y and Z in ellipsoidal elevation considering EPSG:25830 (ETRS89 UTM 30 North) reference system, additionally, for each point the LiDAR intensity, return number and RGB reflectance are given (the latter obtained from a RCD30 half-format camera). Finally, each point was automatically assigned to a class using TerraScan software following the ASPRS standard class definition [6].

Using these SPL100 point clouds, two-class supervised classification algorithms were fit to predict whether a point corresponded to the class 'high-voltage power line' (class=1) or not (class=0). The point attributes used as descriptive features for the classification were the following (Table 1):

Table 1. Point attributes used as input for the classification

Attribute	*Description*
Z	Point elevation
Intensity	Intensity of the laser return
Return number	Number of return of the point
Number of returns	Total number of returns for the pulse
R	Digital number of the red channel
G	Digital number of the green channel
B	Digital number of the blue channel

A ground truth dataset was built to train and test the classification algorithms. For this, fifteen 1x1km tiles were selected on areas with high-voltage power lines present over different land uses (6 over forest, 4 over agriculture and 5 over urban areas) (Fig. 1). Then, existing power line cartography was overlaid (Fig. 1) and a 20 m buffer width around the lines was established to mask out the remaining areas. Using a proprietary algorithm (TerraScan), points belonging to power lines were classified. The algorithm is based on a recursive fitting of the catenary curve for all the points in the cloud [7], so it was very expensive in terms of computing time. Besides, it required a significant number of parameters that had to be set for each area, being the outcome very sensitive to some of them, so visual inspection of the results was mandatory and parameter values had to be fine-tuned for each area to avoid false positive and incorrect results. This process was very labor intensive but necessary to obtain a detailed training dataset to build the models upon.

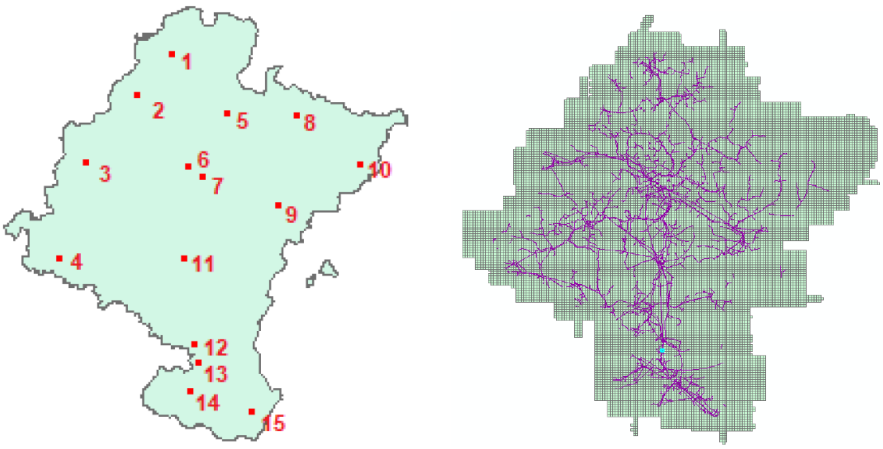

Fig. 1. Left: Location of the fifteen tiles used to obtain the train and test datasets. Right: High-voltage power line cartography overlaid over the SPL LiDAR tiles.

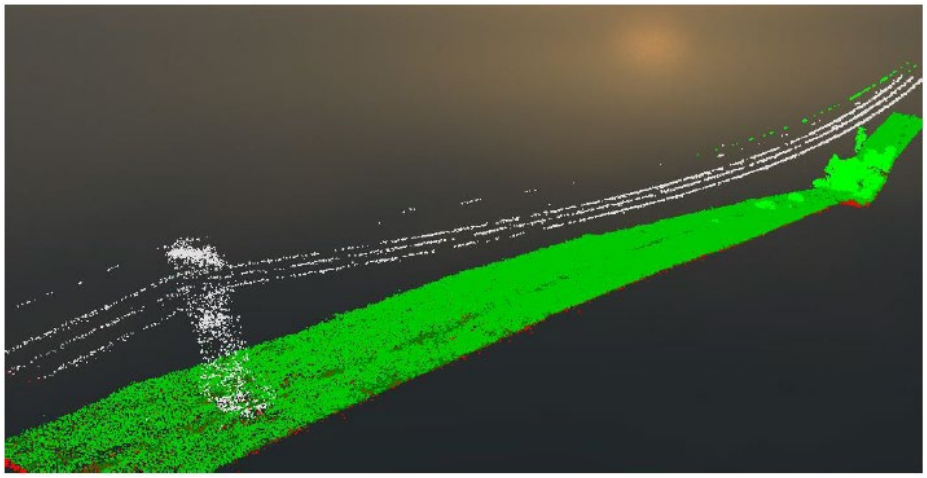

Fig. 2. Sample result of the training dataset used.

A dataset of 4,527,727 points was built out of which 92,104 corresponded to power lines. This dataset was saved as a csv file and randomly split in 75% for training and 25% for testing using the 'hold out' method, and thus ensuring that instances used for training the algorithms are not used for evaluating them and vice-versa. Different data mining approaches were used for classification:

1. Decision trees
   a. C4.5
   b. CART
2. Ensemble learning techniques
   a. Random Forests

b. Bagging
c. AdaBoost

A decision tree (DT) is a classification algorithm that represents a set of organized decisions following a hierarchical structure depending on a set of attributes or descriptive features [5]. DTs are automatically built through a learning process based on the training dataset that is successively split in nodes and branches until leaves are reached (a terminal node where all the elements belong to the same class). At the nodes the dataset is split using the attribute that maximizes an impurity criterion, this criterion was the Gini index in CART and the Information Gain Ratio in C4.5. The obtained DT were pruned to avoid overfitting [5].

Ensemble learning techniques follow the assumption that an ensemble of several classifiers (e.g., DTs) outperforms the results of each classifier alone [8]. Bagging is an ensemble method whereby a classifier is built using new sets of instances obtained applying the Bootstrapping technique, yielding a lower variance prediction if compared to each of the subsets alone and avoiding overfitting [9]. Adaboost updates a first simple DT (weak classifier), by training it iteratively so as to improve the previous version by giving a higher and lower weight to instances incorrectly and correctly classified, respectively [2].

Finally, Random Forest (RF) [10] is an ensemble learning technique that combines the Bagging concept [9] and the Random Subspace concept [11]. On a RF each tree is built upon a different subset of data and using a random subspace of attributes to select the best splitting attribute at each node. This way, the forest is composed of DT that are varied since they learn from a different subset of data and a different subspace of attributes. The final prediction is obtained by majority voting of the different DTs. The generalization error of RF generally depends on the proportion of the final vote assigned to the winning class and on the correlation between the DT that compose the forest.

Supervised data mining methods are severely affected by data imbalance [12], i.e., they tend to over-predict the most frequent classes in the training set. Therefore, in severely imbalance cases like this, training data needs to be balanced first, so as to obtain a training set with the same number of class 1 and class 0 instances. Two approaches can be followed: under-sampling the most frequent class or over-sampling the less frequent one. In this work an over-sampling approach called SMOTE was applied that creates synthetic instances of the minority class by interpolating two instances of the original dataset [13]. Then to reduce the eventual noisy instances created in the borders of the two classes the Tomek Links (TL) method was used [14].

Finally, the test dataset was used to evaluate the quality of the predictions obtained, using a set of performance metrics calculated from the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{4}$$

$$Geometric\ \ mean = GM = \sqrt{TPR \times TNR} \tag{5}$$

where, TP corresponds to the true positive predictions (power lines correctly classified as power lines), TN true negative, FP false positive and FN false negative predictions. The Accuracy (1) represents the proportion of correctly classified points. The Recall or True Positive Ratio (2) corresponds to the proportion of power line points classified as such. The Precision (3) represents the proportion of points classified as power line that actually corresponded to this class. The specificity represented the True Negative Ratio (4), and finally, the Geometric Mean (5) represented the proportion of correctly classified points for both classes. The latter can be used as an overall performance metric to evaluate the behavior of each algorithm.

3. Results and discussion

The application of the trained algorithms to the test data produced in general quite accurate results (Fig. 3). After evaluating the confusion matrices for each case the performance metrics were computed (Table 2). We must point out that SMOTE followed by Tomek Links (SMOTE+TL) was applied only to balance the set of training points, which was subsequently used to train the different models. The performance of the trained methods was always obtained using the original set of test points (without being balanced).

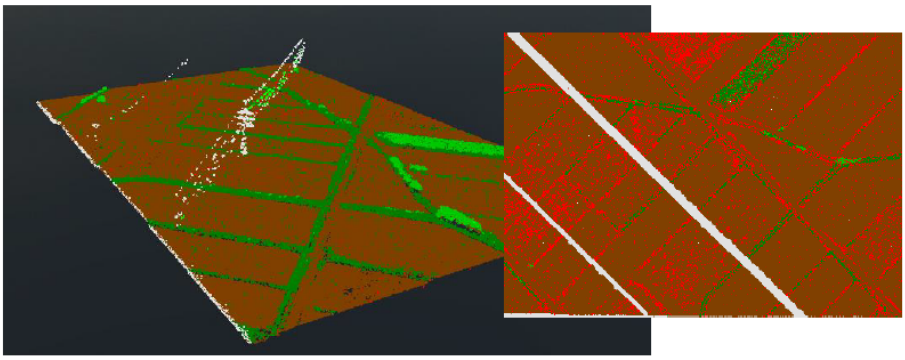

Fig. 3. Sample of a subset of a test area with points classified as power lines represented in white.

Table 2. Performance metrics in testing (all in %) obtained for each classification algorithm using both the imbalanced and balanced datasets

Classification algorithm	*Imbalanced problem*				*Balanced problem (SMOTE+TL)*			
	Acc.	*Rec.*	*Prec.*	*Spec.*	*Acc.*	*Rec.*	*Prec.*	*Spec.*
C4.5	95.2	62.8	59.7	97.3	98.7	76.9	66.7	99.2
CART	95.6	65.0	56.1	97.2	98.7	76.9	65.0	99.1
Random Forests	99.3	65.6	65.9	98.0	99.3	80.4	85.2	99.7
Bagging	96.6	62.9	72.2	98.6	99.2	81.7	81.0	99.6
Adaboost	97.1	66.2	75.7	98.8	99.1	84.5	82.9	99.5

In general, it can be observed that results improved after balancing, with the largest improvements for *Recall* and *Precision* metrics, so both commission and omission errors for the class of interest diminished after balancing. Ensemble algorithms were more benefited by balancing than C4.5 or CART DT. So, it can be said that SMOTE followed by TOMEK LINK was a successful balancing method. Having said that, it can be observed (Table 2) that all algorithms made accurate predictions after balancing the points, with the ensemble learning techniques leading to slightly higher metrics, especially in *Recall* and *Precision*, whereas *Accuracy* and *Specificity* obtained high rates in all cases.

As already explained, the Geometric Mean is the metric that offers a compromise between the correct identification of class 1 and class 0 points, and can be used as a reference for selecting the best performing method (Table 3). Although differences were minor the Adaboost algorithm was the first in the ranking, followed by Bagging, Random Forests at an intermediate position and simple DT algorithms C4.5 and CART a step behind. The main difference was that these last methods had a higher rate of false positives.

Table 3. Classification results in testing in terms of Geometric Mean value obtained for each classification algorithm using both the imbalanced and balanced datasets

Classification algorithm	*Imbalanced*	*Balanced*
C4.5	78.15%	87.36%
CART	79.49%	87.32%
Random Forests	80.14%	89.53%
Bagging	78.73%	90.22%
Adaboost	80.90%	91.73%

4. Conclusions

Several data mining techniques were implemented on a high-voltage power line detection application based on SPL LiDAR point cloud data. The main conclusion is that adequate results were obtained in most cases. Yet, class imbalanced appeared as a key issue and significant improvements were obtained in all cases after training data were balanced, with largest improvements in *Recall* and *Precision* metrics, especially in ensemble techniques. Ensemble techniques achieved higher precision metrics than simple decision trees like C4.5 and CART, in particular in the avoidance of false positives. Although differences were not large the highest accuracies, in terms of the *Geometric Mean* metric, were achieved by Adaboost followed by Bagging.

Using machine learning techniques for detecting high-voltage power lines can be interesting not only for providing accurate results but also for being less computationally expensive than other techniques relying on the computation and fitting of geometric constraints.

Future research efforts in this line should explore the generalization ability of the techniques investigated, validating them over a larger amount of test points. Also, new attributes obtained from LiDAR data, in particular contextual variables between each point and its neighbors should be investigated for their eventual added value.

Acknowledgments

The Government of Navarre and Tracasa are acknowledged for the provision of the SPL LiDAR data, the TerraScan software license and their expertise.

References

1. Zhu, L., Hyyppä, J.: Fully-automated power line extraction from airborne laser scanning point clouds in forest areas. Remote Sensing, 6(11), 11267–11282 (2014).

2. Axelsson, P.: Processing of laser scanner data—algorithms and applications. ISPRS Journal of Photogrammetry and Remote Sensing, 54, 138–147 (1999).
3. Jwa, Y., Sohn, G., Kim, H.B.: Automatic 3D powerline reconstruction using airborne LiDAR data. International Archives for Photogrammetry and Remote Sensing, 38, 105–110 (2009).
4. Liston, R.L.: Photogrammetric methods for mapping resource data from high altitude panoramic photography. Photogrammetric Engineering and Remote Sensing 48(5), 725–732 (1982).
5. Rokach, L., Maimon, O.: Data mining with decision trees. Theory and applications. 2nd edn. World scientific Publishing, Singapore (2008).
6. American Society for Photogrammetry and Remote Sensing (ASPRS): LAS specification version 1.4 – R13 available at https://www.asprs.org last accessed 2019/04/05.
7. TerraSolid:. TerraScan User's Guide, available at https://www.terrasolid.com last accessed 2019/04/05.
8. Scikit-Learn User Guide. Retrieved from http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf last accessed 2019/04/05.
9. Breiman, L.: Bagging predictors. Machine Learning, 24(2), 123–140 (1996).
10. Breiman, L.: Random Forests. Machine Learning, 45(1), 5-32 (2001).
11. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844 (1998).
12. Chawla, N. V., Japkowicz, N., Kolcz, A. Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations, 6 (1), 1–6 (2004).
13. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357 (2002).
14. Visa, S., Ralescu, A. Learning imbalanced and overlapping classes using fuzzy sets. Workshop on Learning from Imbalanced Datasets II (ICML '03), 91–104. (2003).