MACHINE LEARNING ALGORITHMS TO LAND COVER MAPPING WITH LANDSAT-8

Jonathan Richetti[1], Laiza Cavalcante de Albuquerque Silva[1], Willyan Ronaldo Becker[1], Alex Paludo[1], Humberto João Comineti[1], Jerry Adriani Johann[1]

[1]Agricultural Engineering Department, State University of West Parana, Rua Universitária, 1619 – Cascavel – PR – Brazil
{j_richetti@hotmail.com; laiza.cavalcante@hotmail.com; willyan.becker@outlook.com; paludo.alex@hotmail.com; humbertocomineti@gmail.com; jerry.johann@hotmail.com}

ABSTRACT

Data mining algorithms applied to satellite image can be used to land cover mapping. This brings agility to the process of mapping areas and the accuracy can be assessed. However, with many machine learning algorithms it is hard to assess the best one for a giving task. Therefore, this work aims to test different machine learning algorithms to classify land cover using high-resolution imagery. Four algorithms were tested: Bagged CART, Random Forest (RF), Neural Network, and Model Averaged Neural Network in the Landsat-8 tile path/row 223/078 from December 13, 2017. A sample of 42,676 pixels in eight different categories (city, exposed soil, soybean, corn, turnip, pasture, forest, and water) was used. From all pixels, 25,607 pixels (60%) were used as training set and 17,069 pixels (40%) were used as testing set. The results shown that RF algorithm performed better with overall accuracy of 97% and kappa of 0.946.

Keywords *— data mining, classification, remote sensing, satellite image*

1. INTRODUCTION

Data regarding land use is an important information to all decision makers from all sectors. Such information can be obtained using remote sensing data and data mining techniques. Satellite remote sensing data can provide timely, accurate, and objective information on land [1] and data mining algorithms have been widely used in a wide range of areas, including remote sensing mapping for agriculture purposes [2-4]. Bhojani [2] explains that the classification techniques are designed for classifying unknown samples using information provided by a set of predefined samples.

Studies shown the potential in the use of data mining techniques in remote sensing data for classifying land cover. Xiong et al. [5] used Google Earth Engine for an automated cropland mapping algorithm using the Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) with 250-m and 16-day time-series data for Africa continent showed overall accuracies greater than 89%. Telungtla et al [6] also used Google Earth Engine cloud-computing platform 16-day Landsat data, random forest machine learning algorithms, cropland class was mapped with producer's accuracy of 98.8% (errors of omissions = 1.2%) for Australia and 80% (errors of omissions = 20%) for China.

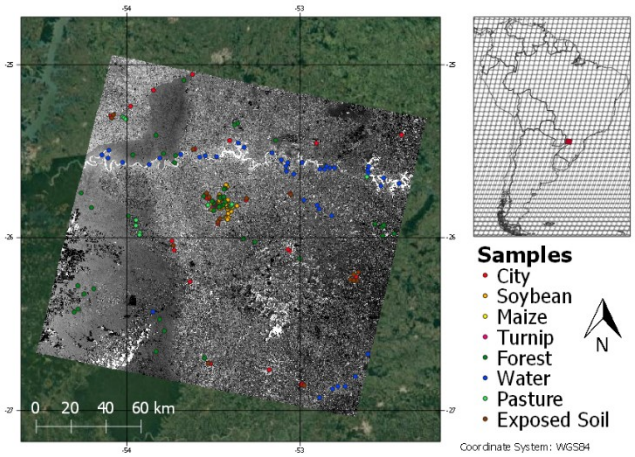

Figure 1. Location of the study area (tile 223/078) and samples used in this study.

More locally, Grzegozewski et al. [7] used a maximum and minimum value of the Enhanced Vegetation Index (EVI) from MODIS to map corn and soybean in Paraná obtaining higher overall accuracy of 86%.

Hence, the aim of this study is to test different machine learning algorithms to classify the land cover using high-resolution imagery. Specific objectives are 1) assess accuracy of different MLA 2) to identify which MLA best performed.

2. MATERIAL AND METHODS

2.1 Study area and dataset

A scene from Landsat 8 OLI path/row 223/078 from December 13, 2017 was used. The Normalized Difference Vegetation Index (NDVI) was calculated and bands 1 to 7 were used as inputs. A sample of 42,676 pixels in eight different categories (city, exposed soil, soybean, corn, turnip, pasture, forest, and water) was used. From all pixels, 25,607 pixels (60%) were used as training set and 17,069 pixels (40%) were used as testing set (Figure 1). The pixels split was done randomly.

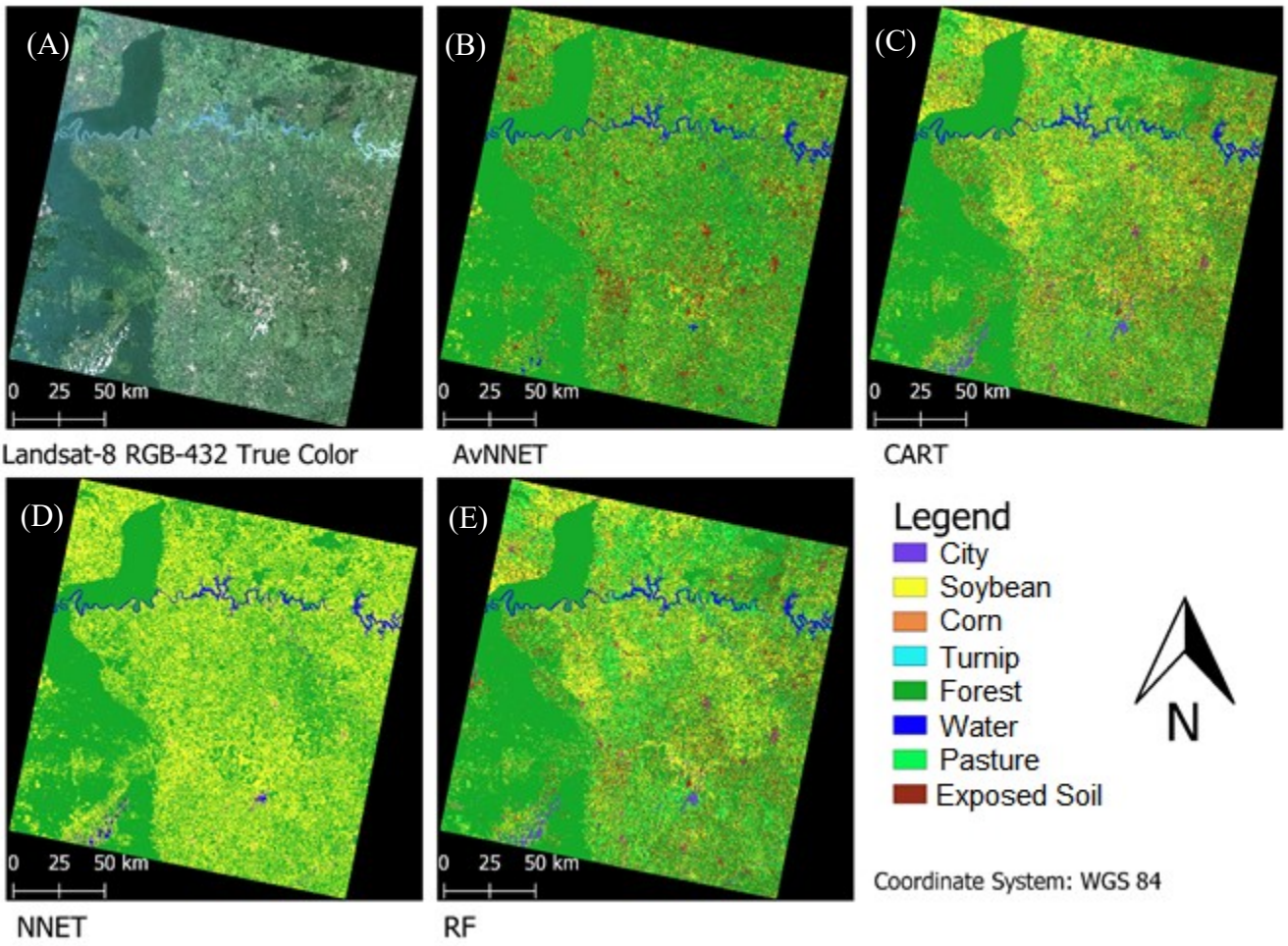

Figure 2. Classified maps from the four different MLA used. True color Landsat-8 RGB-432 (A); AvNNET classification result (B); CART classification result (C); NNET classification result (D); and RF classification result (E).

2.2 Machine Learning Algorithms

Four machine learning algorithms (MLA) were used to classify the Landsat-8 tile. The Bagged CART (CART), Random Forest (RF), Neural Network (NNET), and Model Averaged Neural Network (AvNNET).

The CART is a non-parametric, bagged tree algorithm capable to detect relations among input features and split them into nodes according to their similarity [8]. The RF is an ensemble of decision trees used to classify by bagging technique. The tuning parameters in RF algorithm are the number of variables randomly sampled as candidates at each split (mtry) and the number of trees (ntrees). The NNET is a massively parallel combination of simple processing unit which learn from features and store the knowledge in its connections. The advantage of this algorithm is that it does not need assumptions about the data distribution [8]. The avNNET fits multiple neural network models to the same dataset and predicts using the average of the predictions coming from each constituent model. The class probabilities are averaged to produce the final class prediction [9].

The MLA were performed with no pre-processing and a five folded bootstrapped resampling with 25 repetitions was done for obtaining the optimum tuning parameter. The optimum tuning parameter for each MLA was chosen based on the highest accuracy and kappa values. For RF the mtry tested were 2, 4, and 7 and ntrees were 250, 500, 750, and 1000. For NNET the size tested were 1, 3, and 5 with decay of 0, 0.1, and 0.0001. For AvNNET the same size and decay of NNET was tested with and without bagging.

2.3 Accuracy Analysis

Based on the test set a confusion matrix was calculated for each MLA and the kappa index (kappa – Equation 1), overall accuracy (OA – Equation 2) and the no information error rate were determined. The no information error rate is the largest class percentage in the data. This means that a useful model should do better than predicting the most common class.

$$kappa = \frac{\sum_{i=1}^{r}\frac{x_{ii}}{n} - \sum_{i=1}^{r}\frac{x_i x_i}{n^2}}{1 - \sum_{i=1}^{r}\frac{x_i x_i}{n^2}} \quad (1)$$

$$OA(\%) = \frac{A}{m} * 100 \quad (2)$$

Where: n is the number of observations (sample pixels); A is general correctly classified pixel; m is the number of sampled pixels; r is the number of lines in the error matrix; x_{ij} is observations on row i and column j; x_i is the marginal total of line i; and x_j is the marginal total of column j.

3. RESULTS

The optimum tuning parameters for each MLA was determined (Table 1) based on the training data and used for the classification of the whole tile.

Table 1. Optimal tuning parameters for the tested MLA.

MLA	Parameter	
CART	-	-
RF	Mtry	2
	Ntrees	750
NNET	Size	5
	Decay	0.1
AvNNET	Size	3
	Decay	0.1
	Bag	False

MLA: Machine Learning Algorithms; CART: Bagged CART; RF: Random Forest; NNET: Neural Network; and AvNNET: Model Averaged Neural Network. mtry: number of variables randomly sampled as candidates at each split; ntrees: and the number of trees.

Each MLA was applied to the Landsat-8 tile in study (Figure 2). With the test set the accuracy analysis was performed. The no information error rate was 77% and the OA and kappa of all MLA was higher 95% and 0.89, respectively (Table 2).

Table 2. Accuracy analysis for the tested MLA.

MLA	OA	Kappa index
CART	97%	0.940
RF	98%	0.946
NNET	95%	0.766
AvNNET	96%	0.891

4. DISCUSSION

All MLA presented higher overall accuracy than the no information error rate, which means that the use of the algorithm is better than simply applying the most common class in the image. The best performing MLA was the RF that has been widely used on literature for land-cover classification achieving higher results. Chan et al. [10] used RF for classification with OA of 70%. Feng et al. [11] also applied RF for mapping vegetation with unmanned aerial vehicle obtaining OV higher than 76%. Rodriguez-Galiano et al. [12] used RF for land cover classification obtaining 92% overall accuracy and a Kappa index of 0.92. The results from this study and the reviewed literature shows that the RF presents high accuracy for land classification using satellite image.

5. CONCLUSIONS

This study tested four different machine learning algorithms to classify the land cover using high-resolution imagery. The accuracy of all the different MLA was assessed and was higher than 95% of OA and 0.89 of kappa with the best performing algorithm the random forest. This study shows that data mining techniques can be used with Landsat-8 high resolution imagery for land cover classification.

6. ACKNOWLEDGMENTS

The authors are grateful for the Coopavel Industries for supporting this study by providing the sample data. To UNIOESTE - Cascavel, the Graduate in Agricultural Engineering Program (PGEAGRI) and the GeoScience Research Center for infrastructure and technical-scientific support. The Coordination of Improvement of Higher Education Personnel - Brazil (CAPES), CNPq and Araucária Foundation (FA) for financial support

7. REFERENCES

[1] L. King et al., “A multi-resolution approach to national-scale cultivated area estimation of soybean,” Remote Sens. Environ. v. 195, pp. 13–29, 2017.

[2] Bhojani, S.H. “Geospatial Data Mining Techniques: Knowledge Discovery in Agricultural”. Computer Science, v. 3, pp. 22–24, 2013.

[3] Mankar, A.B.; Burunge, M.S. “Data Mining - An Evolutionary View of Agriculture”. International Journal of Application or Innovation in Engineering & Management, v. 3, n. 3, pp. 102–105, 2014.

[4] Solanki, J.; Mulge, P.Y. “Different Techniques Used in Data Mining in Agriculture”. International Journal of Advanced Research in Computer Science and Software Engineering, v. 5, n. 5, pp. 1223–1227, 2015.

[5] Xiong, J.; Thenkabail, P S., Gumma, M K., Teluguntla P, Poehnelt, J., Congalton, R G., Yadav, K, David Thau. “Automated cropland mapping of continental Africa using Google Earth Engine cloud computing” ISPRS Journal of Photogrammetry and Remote Sensing v. 126, pp. 225-244, 2017, DOI https://doi.org/10.1016/j.isprsjprs.2017.01.019

[6] Teluguntla, P.; S Thenkabail P. S.; Oliphant, A, Xiong, J, Gumma, M K, Congalton, R G., Yadav, K, Huete, A: “A 30-m

landsat-derived cropland extent product of Australia and Chinausing random forest machine learning algorithm on Google Earth Enginecloud computing platform" ISPRS Journal of Photogrammetry and Remote Sensing v. 144, pp. 325–340, 2018, DOI https://doi.org/10.1016/j.isprsjprs.2018.07.017

[7] Grzegozewski, D M; Johann, J A; Uribe-Opazo, M A; Mercante, E Coutinho, A C.: "Mapping soya bean and corn crops in the State of Paraná, Brazil, using EVI images from the MODIS sensor." International Journal of Remote Sensing, v. 37, n. 6, pp. 1257–1275, 2016

[8] Shao, Y.; Lunetta, R. S. "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points". ISPRS International Journal of Photogrammetry and Remote Sensing, v. 70, pp. 78–87, 2012.

[9] Haykin, S. Neural networks: a comprehensive foundation. 2 ed, Prentice Hall Publisher, ISBN-13: 978-0132733502, p. 842, 1998

[10] Chan, J. C. W.; Paelinckx, D. "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery". Remote Sensing of Environment, v. 112, n. 6, pp. 2999–3011, 2008.

[11] Feng, Q.; Liu, J.; Gong, J. "UAV Remote sensing for urban vegetation mapping using random forest and texture analysis". Remote Sensing, v. 7, n. 1, pp. 1074–1094, 2015.

[12] Rodriguez-Galiano, V. F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J. P. "An assessment of the effectiveness of a random forest classifier for land-cover classification". ISPRS Journal of Photogrammetry and Remote Sensing, v. 67, n. 1, pp. 93–104, 2012