Evolutionary Algorithms and Machine Learning Applied to Land Cover/Use Classification

Vinícius Ceccon[1], Pedro Vinicius Yamamoto Agner de Faria[1], Danielli Batistella[2], Vanderlei Aparecido de Lima[3]

[1]Curso Técnico Integrado De Nível Médio Em Agrimensura
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

[2]Departamento de Agrimensura
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

[3]Departamento de Química
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

vinicecconp@gmail.com, pedroviniciusyama@gmail.com,
batistella@utfpr.edu.br, valima@utfpr.edu.br

Abstract. *Land cover/use classification is an important area within Remote Sensing, and it is ordinarily performed with traditional classifiers such as Minimum Distance and Maximum Likelihood. These algorithms presented good results with Landsat-8 images, but they degrade when confronted with higher resolution Pleiades images. To accurately classify higher resolution images, this paper proposes the application of evolutionary filters and Machine Learning classifiers. The filters used were Genetic Search and Multi Objective Evolutionary Search, and the classifiers were Random Forest and Multilayer Perceptron. This conjunction resulted in a model with the best attributes that efficiently classifies the land cover/use, presenting Kappa 0.98.*

1. Introduction

In Remote Sensing, the identification of different areas is traditionally performed through manual recognition. Given this, one of the problems of this field is to design a computer program, that is, an algorithm, which accurately and efficiently classifies specific aspects of images. In this paper, algorithms of this type derive from Machine Learning (ML). ML is an area of Artificial Intelligence that learns from past experience to formulate hypotheses, or models, that are induced by an algorithm. Hypothesis induction represents the data set by bias on certain characteristics [Facelli et al. 2011].

Evolutionary and genetic algorithms, an emerging area of ML in recent years, have largely shown their ability to solve various search and optimization problems. These algorithms use the filter selection bias, which privileges certain attributes considered to be more adjusted according to a fitness function. In this sense, objects with higher fitness ratings are more likely to produce new solutions that have the most qualified attributes [Luger 2013].

Traditional semi-automatic classifiers produce high accuracy classifications of the land cover/use when applied to low resolution images, but degrade when applied to higher resolution images. In this sense, the purpose of this paper is to compare the classification of the land cover/use of high resolution images utilizing traditional semi-automatic classifiers and more sophisticated ML algorithms with evolutionary filters. Also, to determine the most accurate traditional classifier and the conjunction between evolutionary filters and ML classifiers that results in the most accurate classification.

2. Methodology

The activities conducted in this work were performed using the following software: QGIS 3.4.4, ChemoStat, GIMP 2.10.8 and WEKA 3.8.3. In Semi-Automatic Classification Plugin (SCP), a classification plugin within QGIS, Landsat-8 and Pleiades image classifications were performed with traditional semi-automatic classifiers - Minimum Distance and Maximum Likelihood. In WEKA, only the Pleiades picture was rated. Attribute selection was applied with the Genetic Search (GS) and Multi Objective Evolutionary Search (MOES) filters. After that, the classification algorithms Random Forest and Multilayer Perceptron distinguished the image into four land cover/use classes previously defined. Figure 1 shows a representative scheme of the work areas, divided between two main software.

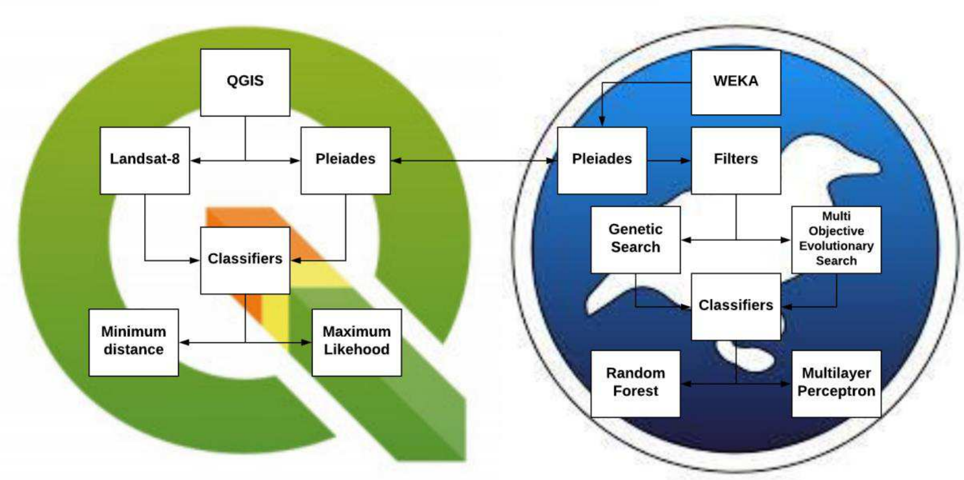

Figure 1. Representative scheme of the work areas.

2.1 Area of study

The area used for classification is a section of the northern portion of the municipality of Pato Branco - PR, mainly constituted of rural area. It was dissociated into four classes of land cover/use: Forest, Agriculture/Pasture, Bare Soil and Urban Area.

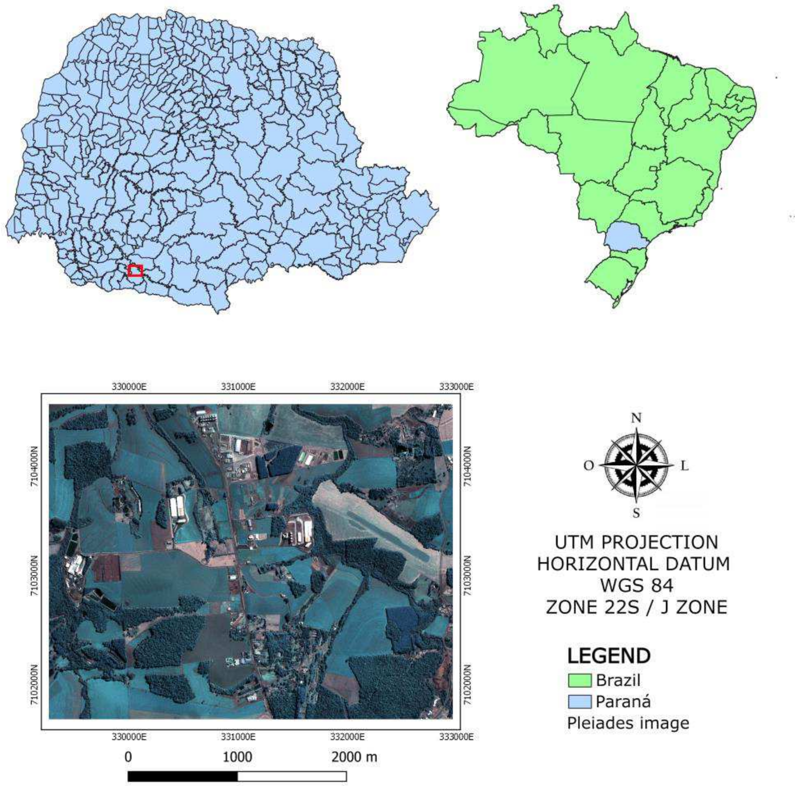

Figure 2. Location of the study area.

The classes are characterized as follows:

- Forest (FO): is composed of dense vegetation. It has rough texture and dark green appearance.
- Agriculture/Pasture (AP): encompasses all types of shallow vegetation and agriculture developed or in advanced development. Its color is light green and may have slight traces of brown.
- Bare Soil (BS): designates areas with surface without any vegetation cover or construction. It has dark or light brown color.
- Urban Area (UA): is all kind of human construction, and incorporates paving areas, residences and sheds. Its color comprises black (asphalt) and white (construction).

2.2 Landsat-8 and Pleiades images

The images analyzed in this work come from the Landsat-8 satellite, whose sensor is OLI (Operational Land Imager), and Pleiades satellites, whose sensor is HiRI (High Resolution Imager). Its RGB bands were merged with the panchromatic band, increasing the spatial resolution from 30 m to 15 m. The spatial resolution of the Pleiades image is 50 cm.

Table 1. Landsat-8 OLI sensor.

Sensor	Espectral Bands	Spectral Resolution	Spacial Resolution	Temporal Resolution	Swath Width	Radiometric Resolution
OLI (Operational Land Imager)	(B1) COSTAL	0.433 - 0.453 μm	30 m	1 day	185 km	12 bits
	(B2) BLUE	0.450 - 0.515 μm				
	(B3) GREEN	0.525 - 0.600 μm				
	(B4) RED	0.630 - 0.680 μm				
	(B5) NEAR INFRARED	0.845 - 0.855 μm				
	(B6) MEDIUM INFRARED	1.560 - 1.660 μm				
	(B7) MEDIUM INFRARED	2.100 - 2.300 μm				
	(B8) PANCHROMATIC	0.500 - 0.680 μm	15 m			
	(B9) CIRRUS	1.360 - 1.390 μm	30 m			

Table 2. Pleiades HiRI sensor.

Sensor	Espectral Bands	Spectral Resolution	Spacial Resolution	Temporal Resolution	Swath Width	Radiometric Resolution
HiRI (High Resolution Imager)	PAN	0.470 - 0.830 μm	0.5 m	1 day	20 km x 20 km and 100 km x 100 km	12 bits
	Blue	0.430 - 0.550 μm	2 m			
	Green	0.500 - 0.620 μm				
	Red	0.590 - 0.710 μm				
	Near Infrared	0.740 - 0.940 μm				

2.3 WEKA classifiers

Within Machine Learning, there is a subdivision of tasks according to the learning model: descriptive, unsupervised learning; and predictive, supervised learning. Therefore, as a classification problem, in which land cover/use classes are the output attributes, two predictive classifiers were selected in WEKA: Random Forest and Multilayer Perceptron.

2.3.1 Attribute extraction

Prior to WEKA, it was necessary to prepare the training samples to train the classification algorithm. In GIMP software, fifty samples for each class were clipped from the study image, resulting in a training set of 200 labeled images. Later using Chemostat software, the grayscale attributes were extracted from these clippings, which produced a file that was later converted to a CSV file and suited to WEKA's file format.

The radiometric resolution of the 12-bit Pleiades image was converted to 8-bit, resulting in 256 shades of gray for each spectral band. In consequence, taking into account the 3 spectral bands used (RGB), the total number of attributes is 769, which comprises 768 grayscale attributes and one attribute for the classes.

2.3.2 Attribute Selection

Attribute selection is a process that identifies the most essential attributes, which improves the performance of the ML model by creating a more concise and less costly model with regards to processing time and data collection. Therefore, this selection seeks the smallest subset of attributes with the best classification accuracy [Pappa 2002b].

Two evolutionary algorithms were used to select the best attributes of the sampled images: Genetic Search (GS) and Multi Objective Evolutionary Search (MOES). The difference between the two algorithms lies in the fact that Genetic Search is a genetic algorithm, an evolutionary algorithms class that uses a tool called crossover to find the space for possible solutions. In contrast, Multi Objective Evolutionary Search is an algorithm based on multi-objective optimization. This optimization expresses a function of local minima and maxima and seeks to optimize or eliminate solutions to find the population of solutions capable of solving a certain problem.

2.3.3 Test Option

The test option refers to how the data set is divided between training set and validation or test set. The first is used to build the model, while the second evaluates the accuracy of the classification. Two test options were used: Cross Validation and Supplied Test Set.

Cross Validation of 10 folds was employed. Since the entire subset is used for validation, the number of classified instances is the same as that of samples, i.e. 200. In the Supplied Test Set, the training and validation sets were separated manually. For this test option, 70% of the labeled images were used for training, and 30% for validation. That being so, the training set was constituted of 140 instances, meanwhile the remaining 60 instances were used for external validation.

2.3.4 Random Forest

Random Forest is a supervised ML algorithm that performs a search in a space of possible solutions according to a hypothesis evaluation function. This type of decision tree-based algorithm performs an attribute selection that identifies the most representative variable for the model, which makes it robust against noise and redundant attributes [Breiman 2001].

Figure 3 illustrates the top-down representative structure of the Random Forest, which is composed of several decision trees. Based on the grayscale that has been evaluated, each tree determines which class they are most likely to belong to, and the most voted class is chosen. The letter of the input attributes symbolizes which band this attribute belongs to - Red (R), Green (G), and Blue (B) - and the number next to it indicates the gray tone, which ranges from 0 to 255. The number of iterations employed in the ratings was 100 and the Seed number was 1.

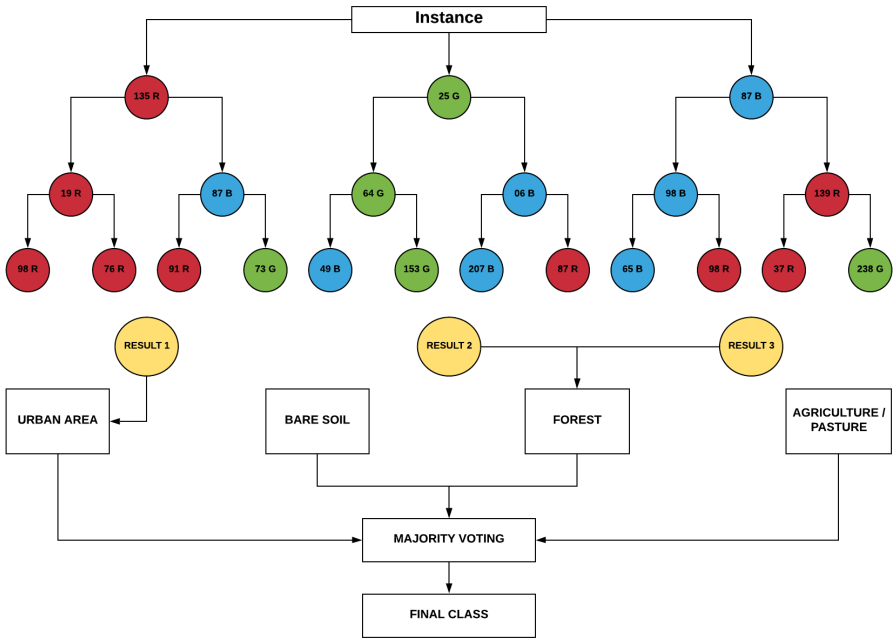

Figure 3. Depiction of a Random Forest utilized in this work.

2.3.5 Multilayer Perceptron

Artificial Neural Network (ANN) is a ML algorithm established on optimization. This kind of algorithm uses a function to find the hypothesis that describes the data and seeks to optimize this hypothesis by minimizing (or maximizing) the objective function. Multilayer Perceptron is an ANN with one intermediate or hidden layer and solves nonlinearly separable problems.

Figure 4 shows a representation of the Multilayer Perceptron developed in this paper. In it, the network layers and connections are expressed. The input layer is represented in green, and is associated with the 768 grayscale attributes. The hidden layer neurons are represented in red and gray and adjust the weights and biases of the connections.

Finally, the output layer is expressed in yellow and gray, and each neuron in this layer is associated with one of the four classes analyzed in this work (UA, BS, AP and FO). 500 epochs were used as the training time, momentum 0.2 and learning rate 0.3.

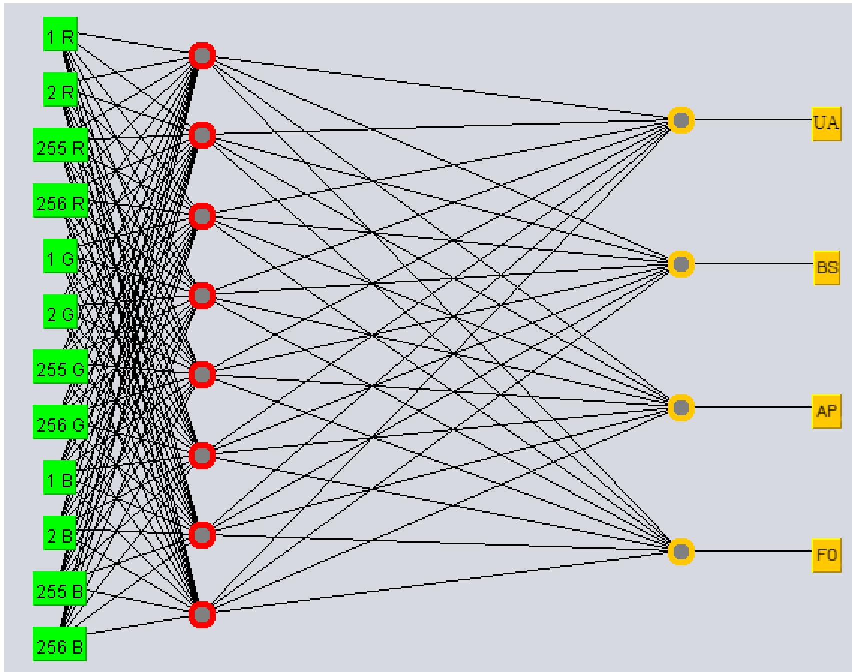

Figure 4. Representation of the Multilayer Perceptron employed in this work.

2.4 QGIS Classifiers

The SCP plugin in QGIS provides a ready-made interface for training sample selection and classification settings. Primarily, it was necessary to select the areas from the image and label them according to their respective classes. In succession, the classifiers chosen from those available in the plugin were: Minimum Distance and Maximum Likelihood.

2.4.1 Minimum Distance

Minimum Distance (MINDIST) is a distance-based classification method, whereas it considers the proximity between data for making predictions. The minimum distance or nearest neighbor algorithm is based on the premise that objects related to the same concept are similar to each other. By calculating the Euclidean distance between the spectral signatures of the training data and each pixel of an image, the algorithm assigns to each pixel the class whose spectral signature is closest.

2.4.2 Maximum Likelihood

Maximum Likelihood (MAXLIKE) algorithm is related to the Bayes theorem, and is a parameter estimator. This classifier calculates probability distributions for classes in the form of multivariate normal distributions, to then estimate whether a pixel belongs to a given class.

3. Results and Discussions

Table 3 compares the accuracy generated in the classification performed by the two classifiers used in QGIS: MAXLIKE and MINDIST. For Landsat-8 and Pleiades images, respectively, MAXLIKE Kappa values were 0.950 and 0.912, while MINDIST Kappa values were 0.841 and 0.680. That being so, for both images the parametric classifier MAXLIKE obtained the best results of the two. In addition, although the spatial resolution of the Pleiades image is higher than Landsat-8's, its classification presented lower Kappa for both QGIS classifiers. This increase in resolution especially affected the Urban Area class, as it showed large decrease in its accuracy. Figure 5 presents the classified images, which evidences the discrepancies in classification.

Table 3. Accuracies of MAXLIKE and MINDIST classifiers.

Image		Landsat-8		Pleiades	
Classifier		MAXLIKE	MINDIST	MAXLIKE	MINDIST
Producer Accuracy [%]	Forest	99.34	99.19	90.84	63.94
	Agriculture/Pasture	95.67	93.60	98.52	97.54
	Bare Soil	92.38	88.41	96.45	73.48
	Urban Area	97.40	75.66	69.11	43.47
User Accuracy [%]	Forest	98.86	95.76	98.73	97.82
	Agriculture/Pasture	91.94	80.40	92.08	80.72
	Bare Soil	93.97	80.45	93.89	91.24
	Urban Area	99.61	97.86	73.69	16.47
Kappa	TOTAL	**0.950**	**0.841**	**0.912**	**0.680**

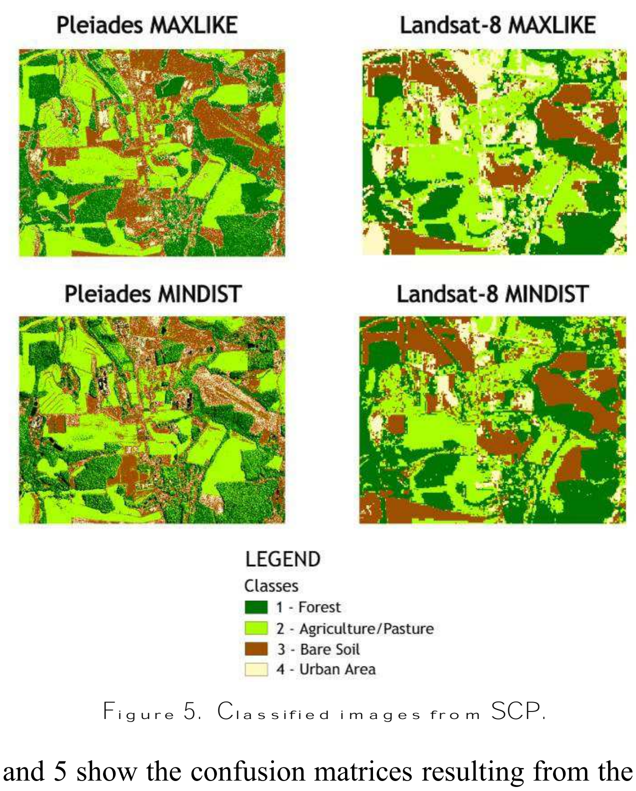

Figure 5. Classified images from SCP.

Tables 4 and 5 show the confusion matrices resulting from the combining of the two filters and two classifiers by WEKA analysis. Cross Validation option test was used in both tables. In table 4, GS was used for attribute selection, and selected 335 relevant attributes from the initial 769. With the attributes selected by this filter, both classifiers showed classification errors in the distinction of the class Urban Area with the classes

Agriculture/Pasture and Forest. Table 5 shows that MOES filter was more rigorous in selection, as it selected 46 attributes from the initial 769. Even after the evolutionary filters were applied, there were classification errors for both classifiers concerning the discrimination between Urban Area and Forest samples. In short, Multilayer Perceptron classifier presented better results when used in conjunction with GS, Kappa 0.9733, and Random Forest was more effective with MOES, Kappa 0.9600.

Table 4. Confusion matrices utilizing GS and Cross Validation.

Filter: Genetic Search					
Classifier	Test Option: Cross Validation			Number of Attributes: 335	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	49	0	0	1
	Agriculture/Pasture	1	48	0	1
	Bare Soil	0	0	50	0
	Urban Area	1	0	0	49
	Kappa:	0.9733	Correctly Classified Instances:	196 (98%)	
Random Forest	Forest	47	0	1	2
	Agriculture/Pasture	0	49	0	1
	Bare Soil	0	0	50	0
	Urban Area	2	0	0	48
	Kappa:	0.9600	Correctly Classified Instances:	194 (97%)	

Table 5. Confusion matrices utilizing MOES and Cross Validation.

Filter: Multi Objective Evolutionary Search					
Classifier	Test Option: Cross Validation			Number of Attributes: 46	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	44	0	0	6
	Agriculture/Pasture	1	49	0	0
	Bare Soil	0	0	50	0
	Urban Area	6	0	0	44
	Kappa:	0.9133	Correctly Classified Instances:	187 (93.5%)	
Random Forest	Forest	49	0	0	1
	Agriculture/Pasture	0	49	0	1
	Bare Soil	0	0	50	0
	Urban Area	1	0	0	49
	Kappa:	0.9800	Correctly Classified Instances:	197 (98.5%)	

In Table 6, both algorithms incorrectly classified two samples from the Urban Area and Forest classes and were the only ones to present errors.

Table 6. Confusion matrices using Supplied Test Set without any filter.

Filter: None					
Classifier	Test Option: Supplied Test Set			Number of Attributes: 769	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	13	0	0	2
	Agriculture/Pasture	0	15	0	0
	Bare Soil	0	0	15	0
	Urban Area	2	0	0	13
	Kappa:	0.9111	Correctly Classified Instances:		56 (93.3%)
Random Forest	Forest	11	0	0	4
	Agriculture/Pasture	0	15	0	0
	Bare Soil	0	0	15	0
	Urban Area	1	0	0	14
	Kappa:	0.8889	Correctly Classified Instances:		55 (91.7%)

The classification errors of algorithms in classifying those classes are presumably due to the proximity of their clipping areas in the image. As the study area was in a rural region, the portions of Urban Area sampled were very close to those of Forest, which impaired the separation of the attributes of these classes. This test option – Supplied Test Set - does not use validation samples in training, and therefore it is possible to evaluate the reliability of previous ratings. However, their prediction accuracy was the worst of all: Multilayer Perceptron and Random Forest algorithms resulted in Kappa 0.9111 and 0.889, respectively. This is not due to the test option, but to the absence of a filter that minimizes noise and redundant attributes. This effect does not take on major proportions for the Multilayer Perceptron algorithm, as it is based on optimization and benefits from a large database. However, it is magnified for Random Forest, because its search model is more sensitive to noisy data.

4. Conclusions

Traditional semi-automatic classification algorithms, Minimum Distance and Maximum Likelihood, available in QGIS and applied in this study, proved to be very effective in discriminating land cover/use. The 15 m spatial resolution image of the Landsat-8 satellite, available for free from INPE, has resulted in very accurate classifications, especially for the parametric algorithm MAXLIKE. However, when these classifiers were confronted with a higher resolution Pleiades image of 50 cm, they were not able to perform so precisely.

In contrast, Machine Learning algorithms have shown to be able to classify the high-resolution image with high accuracy, even higher than that of traditional algorithms applied to the Landsat-8 image.

Analysis of the combinations between evolutionary filters and supervised classifiers shows that the multi-objective filter MOES favors the Random Forest algorithm, while the genetic filter GS generates better results with the Multilayer Perceptron algorithm. Considering Random Forest is a search-based method, providing this classifier with a small number of training attributes causes noise to decrease and, consequently, the model to be improved. On the contrary, GS benefits Multilayer Perceptron as it is a method based on optimization of a function. Thus, this algorithm needs large amounts of attributes for its improvement, and the filter with a higher number of attributes gives the best results.

In this research, the selection of attributes by bio-inspired algorithms effectively eliminated noise, as it selected the most relevant attributes for the land cover/use classification. Even with low operational cost, ML type classifiers were able to generate models that effectively described the data set. In conclusion, it is proved that evolutionary algorithms and search/optimization classifiers together form sophisticated and efficient mathematical machinery for land cover/use classification of high resolution images. Yet, there is space for a future study that applies the models built to the classification of a full extension image.

References

Basgalupp, Márcio Porto (2007). "Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócios", Pontifícia Universidade Católica do Rio Grande do Sul - Faculdade de Informática - Programa de Pós-Graduação em Ciência da Computação.

Breiman, L. (2001). "Random Forests", University of California – Statistics Department, vol 45, pages 5-32.

Congedo, Luca (2016). "Semi-Automatic Classification Plugin Documentation", https://semiautomaticclassificationmanual.readthedocs.io/en/latest/index.html, chap. 3.3. Land Cover Classification.

Deb, Kalyanmoy (2015). "Multi-Objective Evolutionary Algorithms", Evolutionary Computation, Part E, 49, pages 995-1011.

EMBRAPA (2013). "Pléiades – Pleiades Satellite constellation", https://www.cnpm.embrapa.br/projetos/sat/conteudo/missao_pleiades.html?fbclid=IwAR1FD_r4bk_AUZaLyEmsRGcWUopBBgmyKaL9kxi8YBm9Sy6DKSmXP3b9nfs, June 6th, 2019.

EMBRAPA (2013). "Land Remote Sensing Satellite", https://www.cnpm.embrapa.br/projetos/sat/conteudo/missao_landsat.html?fbclid=IwAR1FD_r4bk_AUZaLyEmsRGcWUopBBgmyKaL9kxi8YBm9Sy6DKSmXP3b9nfs, June 6th, 2019.

Facelli, K., Lorena, A. C., Gama, J., et al (2011). "Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina", editora LTC, Rio de Janeiro.

Goldberg, D. E. (1989). "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison Wesley, 1st edition.

Goldberg, D. E., Holland, John H. (1988). "Genetic Algorithms and Machine Learning", vol 3, 3rd edition, pages 95-99. Kluwer Academic Publishers.

Haykin, S. (2001). "Redes neurais: princípios e prática", 2nd edition, Editora Bookman.

IBGE. "Malhas digitais das bases cartográficas de 2015", https://mapas.ibge.gov.br/bases-e-referenciais/bases-cartograficas/malhas-digitais, May 21st, 2019.

INPE. "Catálogo de Imagens", http://www.dgi.inpe.br/CDSR/, February 5th, 2019.

ITCG (2019). "Divisão Político Administrativa do Paraná – 2019", http://www.itcg.pr.gov.br/modules/faq/category.php?categoryid=8#, February 5th, 2019.

Luger, G. F. (2013). "Inteligência Artificial", 6th edition, Editora Pearson Universidades.

Moreira, M. A. (2012). "Fundamentos do Sensoriamento Remoto e Metodologias de Aplicação", 4th edition. Editora UFV, Viçosa – MG.

Novo, E. M. L. de M. (1992). "Sensoriamento Remoto: princípios e aplicações", 2nd edition. Editora Blucher, São Paulo – SP.

Pappa, G. L. (2002). "Seleção de atributos utilizando algoritmos genéticos multiobjetivos", Universidade Pontifícia Católica do Paraná – Programa de Pós-Graduação em Informática Aplicada, Curitiba - PR.

Punia, P., Maninder, K. (2013). "Various Genetic Approaches for Solving Single and Multi-Objective Optimization Problems: A Review", vol 3. International Journal of Advanced Research in Computer Science and Software Engineering, pages 1014-1020.

Santos, J. C., Oliveira, J. R. de F., Dutra, L. V., et al. (2007). "Seleção de atributos usando algoritmos genéticos para classificação de regiões", Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, pages 6143-6150.