Spectral band selection supported by the PSO metaheuristic for prediction of aluminum content of soil samples

Arion de Campos Jr, José Carlos F. da Rocha, Giancarlo Rodrigues

Universidade Estadual de Ponta Grossa (UEPG) – Ponta Grossa, PR – Brazil

arion@uepg.br, jrocha@uepg.br, giancarlops2@gmail.com

Abstract: Near-infrared (NIR) diffuse reflectance spectroscopy (DRS) is a technique that can be employed to make a model of soil nutrient prediction by correlating sample data to the respective reference value obtained by chemical analysis. Such data are organized as attributes of a set of high-dimensional records and making a model from these data involves difficulties that impair its performance. To overcome such difficulties, the use of evolutionary algorithms for Feature Selection has shown to be promising. The aim of this article is to identify, with Particle Swarm Optimization meta heuristic, the relevant wavelengths for predicting the aluminum content of soil samples from the Campos Gerais, PR, Brazil region. Results suggest that, for this scenario, few iterations and small swarm size provide the best subsets.

Keywords: NIR, swarm intelligence, soil nutrients.

1. INTRODUCTION

Near-infrared diffuse reflectance spectroscopy (NIR-DRS) is a method in analytical chemistry that uses the fraction of reflected energy (spectral information) when a known energy level is applied to a sample to identify and quantify its composition. Due to its affordable cost, the fact that it does not use reagents, and that it does not damage the evaluated sample, NIR-DRS has proven to be a viable alternative for estimating the nutrient content of soil. This is an essential activity for soil management.

Estimating nutrient quantities from NIR-DRS data requires a prediction model that associates the spectral information of the sample with the respective reference value. One approach for developing such models is the use of Machine Learning (ML). In this approach, each wavelength of the sample spectrum is an attribute of the dataset used in model training. However, the high dimensionality of the datasets generated by NIR-DRS can lead to overfitting of the models.

Feature Subset Selection (FSS) techniques aim to determine a subset of predictor variables that is effective for the target task. This reduces the dimensionality of the data and reduces the risk of overfitting. The high processing cost of FSS based on wrappers has motivated the use of evolutionary search methods in its implementation (Viniski, 2017). Considering the above, this work evaluates the performance of the Particle Swarm Optimization (PSO) algorithm in identifying relevant attributes for estimating aluminum content, which in high concentrations impairs crop development.

PSO was designed to solve continuous problems, but attribute selection is a discrete problem. One way to overcome this difficulty is to establish a threshold that determines whether or not the attributes will be selected. Since the threshold interferes with the PSO results, it needs to be adjusted for the target task. In view of this, this work incorporated a thresholding scheme into the search procedure. The effect of the PSO

search parameter configuration on the regression procedure was also analyzed. Results suggest that this approach significantly reduces the dimensionality of the database, reducing the risk of overfitting while enabling the definition of a regression model whose predictive performance was superior to that described in related works in the experiments performed.

2. BACKGROUND

Soil is a chemically complex system composed of water, air, and organic and inorganic matter, and is essential for agriculture. One of the nutrients present in soil is aluminum (Al), which in high concentrations compromises the absorption of water and nutrients from the soil and thus harms crop development (Fageria et al, 1988). Therefore, detecting the concentration of Al in soil is important to determine whether any management practices will be necessary.

Conventional soil sampling methods use chemical analyses, which require time to be performed and employ expensive chemical reagents that imply the discarding of the sample after analysis (Rossel et al, 2006). In this context, diffuse infrared reflectance spectroscopy presents itself as an alternative capable of quickly performing simultaneous sampling of several attributes without the disadvantages of conventional methods (Viniski, 2017). This technique focuses a known amount of energy on the sample and measures the amount of reflected energy. Since each molecule has a specific absorption capacity in relation to the energy level provided, it is possible to identify and quantify the chemical composition of the samples by interpreting the reflectance values (wavelengths) recorded in their spectrum.

The estimation of soil nutrient content by NIR-DRS is made from a prediction model and the performance of a model is recognized through its Coefficient of Determination (R^2) and Root Mean Square Error (RMSE), which are descriptive measures of the quality of its adjustment and its accuracy, respectively.

Due to the number of wavelengths, it is necessary to mathematically extract the information from the spectrum and correlate it with the desired attribute so that spectroscopy can be used to estimate soil nutrient content (Rossel et al, 2006). As the size of the set increases, more and more training instances are required to maintain the good performance of the learning algorithm (Russell and Norvig, 2010). One way to overcome such complications is by reducing the dimensionality of the sets.

FSS is a preprocessing technique for reducing the dimensionality of data sets that searches for the minimum subset of its attributes that provides the best possible prediction performance when used in the development of a model (Tan, 2010). To this end, noisy, redundant or irrelevant attributes that could reduce the accuracy of the model are detected and removed. In the FSS based on *wrappers*,

the removal of irrelevant attributes is performed by an iterative procedure that tests different sets of attributes and selects the one that meets previously established accuracy criteria. One of the disadvantages of this approach is its high computational cost.

PSO (Kennedy, 1995) is an evolutionary metaheuristic that simulates the behavior of a flock of birds[1] to solve the search and optimization problems in which it is applied. In this algorithm, the number of dimensions **d** is equal to the number of variables to be optimized in the problem. To search for the optimal solution, at each iteration of the algorithm the particles are repositioned according to the best positions obtained by themselves and their neighboring particles so far. At the end of the algorithm execution, the particle that holds the best positioning/fitness is defined as the optimal solution to the problem. In addition to the control parameters used to update the speed, the number of particles in the swarm and the number of iterations also affect the performance and must be adjusted empirically.

Considering that each ML algorithm adopts a specific approach to propose a model, several NIR-DRS studies applied to the estimation of the nutrient content of soil samples have been proposed (Rossel et al, 2006). FSS, in turn, has proven to be suitable for improving the prediction performance of NIR-DRS models, as it favors the operation of ML algorithms (Viniski, 2017). Related studies commonly aim to propose improvements to the algorithm that favor FSS (Nguyen et al., 2017) or hybrid approaches that perform FSS while optimizing the control parameters of the ML algorithm. Furthermore, almost all of these studies deal with classification tasks rather than data regression.

In general, these publications have not investigated the application of these techniques on NIR-DRS datasets for estimating aluminum content. Furthermore, none of them considered soil samples from the Campos Gerais region.

3. MATERIALS AND METHODS

Two NIR-DRS data sets were used: one to perform the FSS and develop the prediction model - *Dataset 1* - and another to validate its results - *Dataset 2* - as suggested by (Tan, 2010). Each set contains data from samples collected on agricultural properties located in the Campos Gerais region (Piraí do Sul and Ponta Grossa, respectively), which is between 24º and 26º south latitude, 49º and 51º west longitude, with altitudes between 600 and 1,300 meters above sea level.

[1] in the flock, a leading bird (particle) leads the rest of the birds to a food source at a certain speed. If another bird identifies a source with a higher potential than the current leader, it takes the lead and starts to lead the flock in the new direction.

The samples collected on both properties were sent to the physical and chemical analysis laboratory of the ABC Foundation[2], a company that works on the development of research applied to agriculture, so that the measurements could be performed. A spectrometer was used to measure the diffuse reflectance of the samples from both data sets. The equipment takes readings in the range of 400 to 2,500 nm with a 2 nm interval between wavelengths, generating reflectance data for 1,050 wavelengths. The values recorded in the data sets, in turn, are the apparent absorbance of each wavelength, which is the result of the conversion of **log(1/R)** where **R** represents the respective reflectance value.

The reference analysis was performed after the spectral analysis, since it uses chemical reagents that pollute the samples and imply their disposal. The results of this analysis were then attached to the respective spectral data, which finalized the composition of the data set instances. Each data set consisted of 1,050 attributes from the spectral analysis plus the meta attribute obtained by the reference analysis, totaling 1,051 attributes. However, in the presence of 1,050 input attributes, it was not possible to develop even a prediction model from the Multiple Linear Regression (MLR) algorithm, therefore such sets provide the ideal scenario to investigate the applicability of the evolutionary FSS technique with PSO.

To execute the FSS with PSO, it is necessary to elaborate it as an optimization problem that meets the representation manipulated by the algorithm. The particles of the swarm were encoded so that each index of the position vector corresponded to the index of an input attribute of the data set, therefore the dimension of the particles, **d**, was equal to 1,050. Because the FSS is a discrete problem, to use the continuous version of PSO it is necessary to establish a threshold value that determines whether the continuous value existing in the particle index indicates the selection or not of the respective attribute of the data set.

Tran et al., 2016 identified that the threshold value interferes with the number of selected attributes and that the ideal value varies according to the data set. Thus, the value that provided the best subsets was investigated. The candidate values were the same as those used by these authors: 0.05; 0.2; 0.4; 0.6; 0.8 and 0.95.

To establish the fitness function of the optimization algorithm, four steps are defined: ***Step 1****: the indexes of the particle position vector that present a value equal to or greater than the pre-established threshold are identified;* ***Step 2****: a regression equation with the respective selected attributes is elaborated.* ***Step 3:*** *the equation is executed and obtains a regression model together with its respective performance indicators (detailed below). Each particle in the swarm proposes a candidate regression model;* ***Step 4:*** *Finally, the value corresponding to the potential of the solution provided by the particle is returned, which concludes its evaluation.*

Regarding performance indicators, the Akaike Information Criterion (AIC) (Akaike, 1998) allows us to evaluate how well a model fits the data, taking into account its predictive capacity (through its Maximum Likelihood or RMSE) and its complexity

[2] http://fundacaoabc.org/

(number of attributes used). Among the models evaluated, the one with the lowest AIC value represents the best approximate model (Symonds and Moussalli, 2011).

The calculation of the AIC in the fitness function used the RMSE value of the model developed by the regression algorithm[3]. Equation 1 shows how the AIC was calculated, in which **n** is the number of samples used for training, **ln** is the logarithm operation and **m** is the number of attributes of the model. The first term of this equation works with the predictive capacity of the model while the second penalizes its complexity.

$$AIC = n \cdot \ln(RMSE) + 2 \cdot m \quad (1)$$

The *SPSO2011* (Clerc, 2012) version was used in this article because it is considered a standard version. When using PSO to solve an optimization problem, it is necessary to establish its parameters. Several combinations of number of iterations, swarm size and threshold values were evaluated in order to identify the one that provided the best FSS result in *Dataset 1*: Search space size **d**: 1,050; Search space boundaries: 0.0 ~ 1.0; Stopping criterion: Reach the maximum number of iterations (40, 70, 100); Swarm size: 20, 40, 60, 80, 100.

Each combination was repeated 30 times using a different seed for random number generation, which was modified in a controlled manner. This procedure was necessary due to the stochastic nature of PSO and for reproducibility of results. The stopping criterion of the algorithm was to reach the maximum number of iterations.

To recognize the best subset of attributes among the 30 available in the identified optimal combination, each one was validated in *Dataset 2* and then the best one was identified according to the same criteria for identifying the optimal combination. This was the additional selection procedure indicated by Xue et al. (Xue et al., 2012).

The parameter combinations and validation results were statistically compared using the Friedman test with 5% significance to investigate the existence of statistical differences between them. If a difference is found, then the Friedman post-hoc test is used to identify which pairs presented this characteristic. Through these tests and the empirical evaluation of the results, the best combinations and the best subset for predicting the exchangeable aluminum content of the manipulated sets are identified.

4. RESULTS AND DISCUSSION

The optimal combination of the number of iterations, swarm size and threshold value was investigated in order to identify the parameter configuration that allowed the selection of attributes that generated the regression model with the highest prediction indices of aluminum content. In the proposed problem, a solution (global or local) should optimize the model fit and penalize high dimensionality. This is obtained by the AIC index. The results show that the lowest AIC, achieved with the threshold 0.95, were associated with models with only one attribute (expected behavior, due to the high threshold). However, such models proved to be inadequate in terms of R^2 and RMSE

[3] Available in the R software, which seeks to identify a linear relationship between the input attributes and the target attribute, which could not be executed in the presence of 1,050 attributes.

(Symonds and Moussalli, 2011). Considering that the AIC was not related to the other fit indicators, the values of R^2 and RMSE were tested next.

In situations where the Friedman Test found a statistical difference for the R^2 or RMSE values, the *post-hoc* test was used to indicate it between the pairings. This test was first conducted on the R^2 values, but in situations where the pairings did not present a statistical difference regarding this metric, it was also conducted on the RMSE values to identify a single solution.

The first characteristic exposed by the different threshold values refers to the average number of selected attributes, since, in general, the lower the value, the more attributes were selected. Regardless of the number of iterations, when using 20, 40 or 60 particles in the swarm, the **0.6** threshold favored the selection of the attributes that produced the models with the best performance, while with 80 and 100 particles, the **0.4** threshold achieved this feat. The ideal swarm size was identified based on the R^2 and RMSE values of the models produced by the ideal thresholds of each one, which were compared with each other to identify the ideal size to be used in each number of iterations.

The Friedman test found a statistical difference between the R^2 and RMSE values of the models, so the post-hoc test was applied. However, the results of the evaluations and the post-hoc test were identical for each number of iterations and swarm size evaluated. Thus, the swarm size of 20 was selected due to the lower computational cost resulting from the smaller number of particles evaluated. The Friedman test also did not detect a statistical difference between the R^2 and RMSE values for the models generated with 40, 70 and 100 iterations. Based on this result, the configuration with 40 iterations was chosen.

The optimal combination of parameters to run FSS with PSO on *Dataset 1* was the following: 40 iterations, 20 particles in the swarm, and threshold at 0.6. Considering this configuration, after thirty rounds, the best results are presented in Table 1. The experiment highlights the drastic reduction of attributes in the dataset.

Tabela 1. Training and validation performance of the best model generated by combining 40 iterations, 20 particles and threshold 0.6

		Training			Validation		
Iteration	N Atributes	AIC	R^2	RMSE	AIC	R^2	RMSE
10	**22**	**46,05**	**0,61**	**1,24**	**88,61**	**0,50**	**5,96**

Based on the model with the best R^2 in training, Equation 2 presents the regression formula used by this model to predict aluminum content, in which 12.88 is its intercept or adjustment value. Variables starting with the letter **r** refer to the sample's reflectance

at the respective wavelength[4] and the values preceding them are the assigned coefficients.

$$\begin{aligned} AluminumContent = {} & 12,88 + 404,56 * r486 - 940,07 * r554 + 4.821,80 * r594 \\ & - 6.002,88 * r612 + 1.911,41 * r678 - 12.964,86 * r1008 \\ & + 17.804,04 * r1020 - 5.996,58 * r1050 + 2.164,77 * r1152 \\ & + 2.980,35 * r1272 - 5.124,90 * r1276 + 1.180,70 * r1404 \\ & - 1.605,05 * r1430 + 3.095,31 * r1626 - 3.112,30 * r1864 \\ & + 1.236,11 * r2040 + 963,61 * r2086 + 1.592,58 * r2274 \\ & - 755,02 * r2308 - 4.507,30 * r2332 + 2.636,73 * r2368 \\ & + 218,62 * r2460 \end{aligned} \quad (2)$$

Regarding the research published in the area and considering the results obtained, the best identified model obtained a prediction performance superior to those of Rossel et al. (Rossel et al, 2006), developed in the MIR spectral region, and of Terra, Demattê and Rossel (Terra et al., 2015), developed in the Vis-NIR region, with transformed data and with many more training instances available. These authors did not use FSS to develop their models, therefore the potential of FSS is evident.

5. CONCLUSION

In this work, an evolutionary FSS technique was applied to a spectroscopy dataset to select a subset of attributes relevant to predicting the aluminum content of soil samples. The FSS was approached as an optimization problem and its objective was to minimize the average AIC value in the elaboration of the models of the candidate subsets by the RLM algorithm. Due to its performance in the FSS, the continuous version of the PSO was used and required investigation of the best possible parameter configuration.

This investigation indicated 40 iterations, 20 particles in the swarm and threshold 0.6 as the ideal combination, which in addition to providing the best-performing models used the smallest number of iterations and swarm size analyzed. The models obtained through this combination demonstrate the potential of the adopted method, since the most complex model used less than 10% of the total attributes of the original set, while the model with the best performance used 22 input attributes.

The results obtained by applying the evolutionary FSS were positive. Without dimensionality reduction, it would not be possible to obtain a model generated by the RLM algorithm in the presence of the 1,050 input attributes of the data set. However, using this algorithm, only 22 attributes were needed to explain 50% of the data variation (R^2=0.5). This small number of selected attributes compared to the original number of the set highlights the potential of the modeling adopted to perform evolutionary FSS with the PSO algorithm, therefore this is an opportune and suitable technique for FSS in spectroscopy data sets.

Considering that the coefficient of determination obtained was low, some opportunities for future work were identified, which can improve the results of this work. We highlight the adoption of new particle initialization methods and to deal with this

[4] Example: r486 refers to the reflectance value at the wavelength of 486 nm.

research as a multi-objective problem, minimizing the number of attributes and maximizing the predictive capacity, simultaneously.

REFERENCES

Akaike, H. (1998). **Information Theory and an Extension of the Maximum Likelihood Principle.** In Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics), pages 199–213.

Clerc, M. (2012). **Standard Particle Swarm Optimisation**.

Fageria, N. K., Ballgar, V. C., and Wright, R. J. (1988). **Aluminum toxicity in crop plants**. Journal of Plant Nutrition, 11(3):303–319.

Kennedy, J. and Eberhart, R. (1995). **Particle swarm optimization.** In Proceedings of

ICNN'95 - International Conference on Neural Networks, volume 4, pages 1942–1948.

IEEE.

Nguyen, H. B., Xue, B., Andreae, P., and Zhang, M. (2017). **Particle Swarm Optimisation with genetic operators for feature selection.** In 2017 IEEE Congress on Evolutionary Computation (CEC), pages 286–293. IEEE.

Rossel, R. V., Walvoort, D., McBratney, A., Janik, L., and Skjemstad, J. (2006). **Visible,**

near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131(1-2):59–75.

Russell, S. and Norvig, P. (2010). **Artificial Intelligence: A Modern Approach.** Prentice Hall, 3 edition.

Symonds, M. R. E. and Moussalli, A. (2011). **A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion.** Behavioral Ecology and Sociobiology, 65(1):13–21.

Tan, K. H. (2010). **Principles of soil chemistry.** CRC press, 4 edition.

Terra, F. S., Demattê, J. A., and Viscarra Rossel, R. A. (2015). **Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data.** Geoderma, 255-256:81–93.

Tran, C. T., Zhang, M., Andreae, P., and Xue, B. (2016). **Improving performance for classification with incomplete data using wrapper-based feature selection.** Evolutionary

Intelligence, 9(3):81–94.

Viniski, A. D. and Guimaraes, A. M. (2017). **Técnicas de seleção de atributos para mineração de dados de alta dimensionalidade gerados por espectroscopia no infravermelho próximo–NIR.** In Anais SULCOMP, volume 8, Criciuma, SC.

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2016). **A Survey on Evolutionary**

Computation Approaches to Feature Selection. IEEE Transactions on Evolutionary

Computation, 20(4):606–626.