Problem-Based Band Selection for hyperspectral images

Mateus Habermann, Vincent Frémont, Elcio Hideiti Shiguemori

PROBLEM-BASED BAND SELECTION FOR HYPERSPECTRAL IMAGES

Mateus Habermann [1,2]*, Vincent Fremont* [1]*, Elcio Hideiti Shiguemori* [2]

[1]Sorbonne Universités, Université de Technologie de Compiègne,
CNRS, Heudiasyc UMR 7253, CS 60319, 60203 Compiègne cedex, France.
[2]Institute for Advanced Studies, Brazilian Air Force, Brazil.
email: mateus.habermann@hds.utc.fr

ABSTRACT

This paper addresses the band selection of a hyperspectral image. Considering a binary classification, we devise a method to choose the more discriminating bands for the separation of the two classes involved, by using a simple algorithm: single-layer neural network. After that, the most discriminative bands are selected, and the resulting reduced data set is used in a more powerful classifier, namely, stacked denoising autoencoder. Besides its simplicity, the advantage of this method is that the selection of features is made by an algorithm similar to the classifier to be used, and not focused only on the separability measures of the data set. Results indicate the decrease of overfitting for the reduced data set, when compared to the full data architecture.

Index Terms— Band Selection, Deep Learning, Artificial Neural Networks, Feature Selection, Binary Classification. .

1. INTRODUCTION

Deep Learning-based classifiers are powerful due to the sequence of consecutive layers, which are capable of extracting complex features from the data set, what yields better results when compared to shallow structures. Deep architectures normally have several parameters whose learning demands a big quantity of training data, which are not always available. This problem becomes more acute in the classification of hyperspectral images, because they have tens (sometimes hundreds) of dimensions and those images normally are not always available in quantities large enough to provide an appropriate training for the algorithm. In fact, it is known that insufficient training data may lead to overfitting, and, in order to decrease the amount of algorithms parameters, one resorts to dimension reduction of the input data. This may be achieved by either feature selection or feature extraction, which share a common objective, but are different in the way they achieve it. A common representative of feature extraction is Principal Components Analysis, which performs a linear combination of the original features to generate new ones. The feature selection methods, on the other hand, select the best bands —according to some criterion—amongst the original data set.

In [1], it is proposed a feature selection method for hyperspectral images, which uses boosted decision trees. Several decision trees are generated, and the features most used by those trees are selected. In [2], feature selection was performed by means a perceptron neural net with step function. After the training, features with the smallest interconnection weights were discarded. This approach reduces the processing time, compared to features selection based on support vector machines. It also avoids the evaluation of multiple feature subset combinations, what is common on wrapper approaches. In [3], a comparison between wrapper and filter methods for feature selection is made. It is proposed a filter-based forward selection algorithm, which has some characteristics of wrapper method. Indeed, the proposed method uses boosted decision stumps. In a successive processing, the features that correctly predicts the class label are chosen to be part of the reduced feature set.

In this paper, we propose a filter-based algorithm for band selection, considering binary classification. It is a simple method based on a single-layer neural network, which enables the selection of bands linked to the most important weights for a given binary classification.

In section 2, the main techniques for feature selection are described. In section 3, one can find the details of the proposed method, whose results are exhibited in section 4. Finally, the conclusion can be found in section 5.

2. FEATURE SELECTION

A hyperspectral image captures the physical properties of the scene being imaged. Each band is related to a wavelength, whose reflectance may vary significantly from one object to another, permitting, this way, the distinction amongst differ-

This work was carried out in the framework of the Labex MS2T and DIVINA challenge team, which were funded by the French Government, through the program Investments for the Future managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). We are also thankful for the support provided by Brazilian Air Force and Institute for Advanced Studies (IEAv).

ent classes. Thus, depending on the application, certain bands are more important than others [4].

The high correlation between contiguous spectral bands, *per se*, is enough to justify the dimensionality reduction of the data. Another major benefit provided by this reduction, though, is the possibility of having less classifier parameters, when using a deep neural network, for instance.

It is known that the ratio $\frac{|X|}{l}$ —the bigger, the better —is a valid token for assessing the likelihood of overfitting, where $|X|$ is the cardinality of the training data set X, and l is the quantity of the classifier's parameters [5]. Since it is not usually possible to increase $|X|$, one resorts to decreasing l in order to maximize that ratio. For a fully connected deep neural network, decreasing the input layer size means decreasing l of its architecture.

As for the feature selection, two methods are normally employed: *i)* wrapper-based feature selection; and *ii)* filter-based feature selection.

- **Wrapper-based feature selection:** in this case, the selection of features is performed by the classifier, during the training phase. It means that more appropriate features are likely to be chosen, since it is the classifier itself that chooses them. On the other hand, however, this method is computationally slow, because at each new combination of features, the classifier must be trained again.

- **Filter-based feature selection:** under this approach, the feature selection is performed before the training of the classifier. In fact, it may be seen as a data preprocessing step. The positive aspect of this method is its speed in relation to wrapper-based approaches. A drawback is that the most appropriate features may not be selected, because the feature selection process has no relation with the classifier.

In this work, it is proposed a filter-based band selection, due to its simplicity. Since the feature selection method and the classifier are both based on neural networks —each one with its own depth —, it is believed that the selected features be more appropriate for the classifier. This way, it is possible to have some wrapper benefits in a filter-based environment.

3. METHOD'S DESCRIPTION

The method consists of using a single-layer neural network (see Fig. 1) for an initial binary classification. Thus, each band of the input vector is directly linked to the output vector, what makes it easier to assess the importance of the features by analyzing its correspondent weight magnitude.

After this procedure, the most important bands may be selected and used in more powerful classifiers.

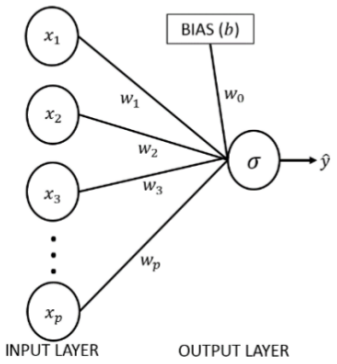

Fig. 1. Architecture used for band selection.

3.1. Mathematical model of the neural network

Given an input vector $x \in \mathbb{R}^p$, where p is the quantity of spectral bands, its estimated class $\hat{y}$ is given by

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

σ is called sigmoid function and $z = W^T x + b$, where W and b are the weights and bias of the neural network, respectively.

The cost function is the cross-entropy, and the training of the network is made by stochastic gradient descent, with back-propagation algorithm.

The proposed method is set up in a way that the actual label y^i, which is associated to the input vector x^i, receives the value 0 to indicate the *target*, and 1 if it is a *distractor*[1].

In practice, given an input vector x^i, either of the following situations always holds:

- $z < 0 \implies \sigma < 0.5 \implies \hat{y} \leftarrow 0$
- $z \geq 0 \implies \sigma \geq 0.5 \implies \hat{y} \leftarrow 1.$

Therefore, the signal of z determines whether an input data should be assigned to the class 0 (target) or 1 (distractor). Since the input data is truncated into 0 and 1, the weights w_j with $j = 1, ..., p$, of the equation

$$z^i = x_1^i w_1 + x_2^i w_2 + ... + x_p^i w_p + b \tag{2}$$

play a fundamental role in determining the signal of z^i, and, consequently, the estimate $\hat{y}^i$ for x^i.

In Fig 2 it is shown an example of the weights values after the training of the algorithm. In the region where the mean value of the distractor's spectral signature is higher than that of the target, the weights have a positive value. Conversely, in the region where the target has bigger spectral values, the weights are negative. Those positive and negative weights are responsible for the signal of z to be positive or negative, respectively. As the weights with absolute values near zero are

[1]This term is largely used in the Saliency Maps area, meaning every image element but the target.

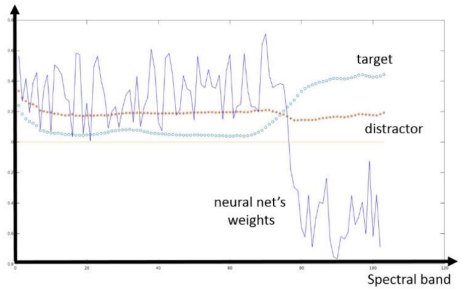

Fig. 2. Spectral signatures of target and distractor. The weights

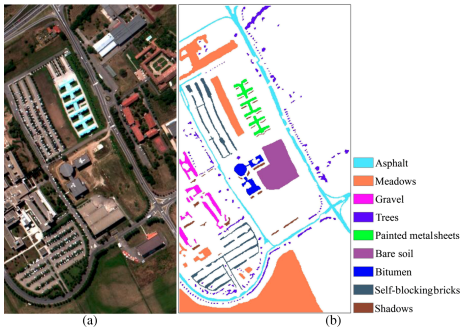

Fig. 3. Pavia university and its classes [7].

of little importance for (2), one can infer that the weights with the biggest absolute values are the most important ones, and, thus, by analogy, one may conclude that the bands relates to those important weights are the most relevant for this classification task. Thus, the output of the proposed method are the selected spectral bands.

4. RESULTS

In this work we use two hyperspectral images acquired by the ROSIS sensor during a flight campaign over Pavia, Italy. The first image (Fig. 3) depicts the city's center (103 bands) and will be used as the training image. The second image (Fig. 4) shows the university (102 bands) and will be used as the testing image. The geometric resolution for both images is 1.3 meter. For the training of the deep learning-based classifier, we use the Theano library [6].

When it comes to band selection, the five biggest positive and five biggest negative (in absolute values) bands are chosen. After the selection of those 10 bands, we proceed to a binary classification by means of deep learning. More precisely, we use a stacked denoising autoencoder with five hidden layers.

Table 1 shows the architectures, which follow a funnel-like shape. It means that the $(l+1)^{th}$ layer has, at most, the same quantity of neurons than the l^{th} layer. By doing so, it is possible to enforce the encoding process between consecutive

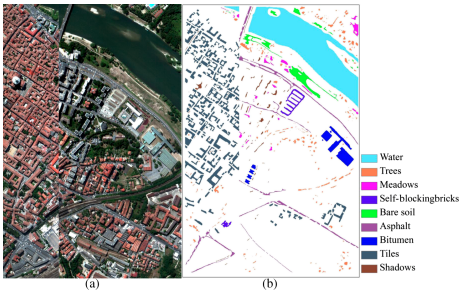

Fig. 4. Pavia center and its classes [7].

Table 1. Deep architectures. From input until output layers.

All bands	102 : 90 : 60 : 30 : 15: 5 : 2
Selected bands	10 : 8 : 6 : 5 : 4 : 3 : 2

layers.

Both data sets (full and reduced data) are classified, and, at the end, we compare the classifier's accuracy over a data set not used during the training phase.

Four different classes from the data set were chosen to be the targets. The criterion for this choice was the similarity of spectral signatures between the training and testing images, considering the same target. Table 2 shows the targets to be classified.

During the training phase of the classifier, the data set (Pavia university) is split into two subsets: *training data* and *validation data*. After each training epoch, the error rate of the algorithm is measured taking into account both training and validation data sets. At the end of the training phase, we keep the parameters learned at the epoch whose validation error is the smallest. Then, the testing set (Pavia center) is used in order to asses the generalization capability of the classifier.

Sometimes the algorithm may achieve very low error rates during the training, but yields high error rates for the test data. This may happen due to overfitting, that is, the algorithm gets specialized in the training data, but cannot perform well when confronted with new data sets.

In Table 3 there are the validation and testing accuracies, measured with validation data and testing data, respectively. The *full data* is the original data, and the *selected data* is the set reduced by the proposed method.

Table 2. Targets to be classified.

Trees	Target 1
Meadows	Target 2
Self-blocking bricks	Target 3
Bare soil	Target 4

Table 3. Accuracy results for training and testing data, for both full and selected data set.

	Target 1	**Target 2**	**Target 3**	**Target 4**
Full	val. 1.52%	val. 4.68%	val. 0.2532%	val. 2.78%
data	test 0.057%	test 50.19%	test 2.12%	test 69.57%
Selected	val. 1.77%	val. 3.92%	val. 0.0%	val. 30.63%
data	test 0.038%	test 50.0%	test 2.11%	test 50.0%

4.1. Remarks about the results

According to Table 3, the *validation* error rates are smaller than the *test* errors (exception made for Target 1). In fact, the classifier learns its parameters with the training and validation data and check their validity with the test data. It is expected, thus, that the error rate be smaller for the validation data.

What is not desired, though, is the high discrepancy between the validation error rate and the test error rate. This may indicate the presence of overfitting.

In the case of classification of the Pavia data set, for Targets 2 and 4 the difference between validation and test error rate is really big. On this, there are two reasonable explanations: *i)* there is a problem in the data set (calibration issues, or wrong assignment of pixels to classes, for example); or *ii)* overfitting.

For the case *i*, there is nothing to be done in a Pattern Recognition level. In relation to assumption *ii*, according to Table 3, for all the targets the test error rates for the *reduced data* are smaller than that of the *full data*. It can be credited to a lesser occurrence of overfitting in the selected data set case.

Since the amount of training data is the same for both cases, the relation $\frac{|X|}{l}$ is bigger for the selected data case, rendering it less susceptible to overfitting.

5. CONCLUSION

Deep Learning has been employed with success in classification of hyperspectral images. Its successive layers make it possible to extract abstract features from data, letting the resulting projection of data points more easily separable by a hyperplane.

Normally, the quantity of parameters of a deep architecture is big, what demands that the training data cardinality be huge, in order to ensure a good generalization of the classifier. However, in many cases the quantity of available training data is not enough to avoid overfitting.

One way to minimize this problem is to reduce the input layer size, when it comes to deep neural networks. This reduction may be achieved by means of feature selection.

The present work proposed a filter-based band selection for binary classification of hyperspectral data. The selection of features is performed by a single-layer neural network. The bands connected to the biggest weights, either positive or negative, are selected. Then, the new data set with selected bands is classified by a deep neural network. Comparing the accuracy results between the selected and full data sets, it is possible to infer that the presence of overfitting was less severe for the selected data set. Therefore, this fact vindicates the validity of the proposed method.

Despite the positive results, some improvements are still necessary. The next step is to expand the present method from binary to a multi-class feature selection. Furthermore, we will seek to devise a way to determine the deep architecture in an automatic fashion, in order to ease the task of a future end-user.

6. REFERENCES

[1] S. T. Monteiro and R. J. Murphy, "Embedded feature selection of hyperspectral bands with boosted decision trees," in *2011 IEEE International Geoscience and Remote Sensing Symposium*, July 2011, pp. 2361–2364.

[2] Arroyo G Meja-Lavalle M, Sucar E, "Feature selection with a perceptron neural net," in *In: Proceedings of the international workshop on feature selection for data mining, pp 131135*, 2006.

[3] Sanmay Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, ICML '01, pp. 74–81, Morgan Kaufmann Publishers Inc.

[4] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9–15, Mar 2017.

[5] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition, Fourth Edition*, Academic Press, 4th edition, 2008.

[6] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010, Oral Presentation.

[7] Jianwei Gao, Qian Du, Lianru Gao, Xu Sun, and Bing Zhang, "Ant colony optimization-based supervised and unsupervised band selections for hyperspectral urban data classification," *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 085094, 2014.