REDUCING HYPERSPECTRAL DATA DIMENSIONALITY USING RANDOM FOREST BASED WRAPPERS

[1]*Nitesh K Poona* & [2]*Riyad Ismail*

[1]Department of Geography and Environmental Studies, Stellenbosch University, Stellenbosch, South Africa, poona@sun.ac.za;
[2]Discipline of Geography, University of KwaZulu-Natal, Pietermaritzburg, South Africa, riyad.ismail@sappi.co.za

ABSTRACT

The random forest algorithm has been widely used for classification of hyperspectral data. To improve model interpretation and classification, the random forest algorithm is often combined with feature selection algorithms. It is within this context that we explore the utility of three random forest wrappers to compute an optimal subset of wavebands to discriminate healthy and stressed *Pinus radiata* seedlings. The Boruta algorithm provided the best classification results using a subset of 17 wavebands of an original 1 769 wavebands. This study demonstrated the value of using wrappers embedded within the random forest algorithm for classification of high dimensional data. in particular, this study highlights the application of the Boruta algorithm for discriminating healthy and stressed *P. radiata* seedlings.

Index terms: Hyperspectral data, random forest, wrappers, *Pinus radiata*

1. INTRODUCTION

Hyperspectral data can provide detailed information on the spectral properties of vegetation. However, the high dimensionality makes data processing difficult [1, 2]. The associated 'curse of dimensionality' often results in reduced classification accuracies arising from the number of samples (n) being many times less than the number of features (p) [2, 3]. In recent years, several authors have advocated the utility of the random forest algorithm [4] for hyperspectral data processing. Within hyperspectral applications, the random forest algorithm has been used for classification purposes in addition to proving a measure of variable importance. Subsequent research has focussed on using the variable importance to improve model interpretation and classification accuracies. However, since the algorithm provides only a ranking and does not automatically eliminate redundant bands [5], researchers have combined the algorithm with various feature selection methods to determine an optimal subset of bands that best explain the phenomena of interest. This paper compares three random forest wrappers in terms of their ability to select of an optimal subset of wavebands that can best discriminate healthy and stressed three month old *Pinus radiata* seedlings. The three wrappers compute the optimal subset of wavebands based on either on the Gini index or mean decrease in classification accuracy. Previous studies have examined the utility of wrappers and filters with random forest but no studies have compared different types of wrappers within a random forest framework.

2. MATERIALS AND METHODS

2.1. Data collection and pre-processing

The main objective of the study was to determine a specific subset of spectral bands that could be used to model asymptomatic stress in three month old *Pinus radiata* seedlings inoculated with the fungal pathogen, *Fusarium circinatum*. An experiment was set up with a healthy class (n = 50) and a stressed class (n = 50). Spectral data was collected over a five week period using an Analytical Spectral Devices (ASD) FieldSpec Pro FR spectroradiometer. The five spectral readings collected per seedling were averaged to a single reading, and the atmospheric water absorption bands (1350-1460nm and 1790-1960nm) were removed [6].

2.2. Random forest

Random forest (RF) is an ensemble of weak unbiased tree-based classifiers (decision trees). RF is similar to bagging but has the additional modification of

selecting only a random subset of candidate features (*mtry*) to determine the split at each node of a tree. Multiple classification trees (*ntree*) are built. As each tree is maximally grown, it makes predictions utilizing an out of bag (OOB) sample for that particular tree. The prediction error then provides an unbiased assessment of the accuracy, since the OOB sample is not used in the training process. RF provides an internal measure of variable importance using the OOB sample. RF was implemented using randomForest library [7] in R statistical software [8].

2.3 Regularized random forest

The premise of the regularization framework is to utilize a regularized version of gain at each node *v*. The coefficient $\lambda_i \in (0,1)$ is used to penalise using a feature $X_i \notin F$ for splitting. A smaller λ_i leads to a larger penalty. The regularized random forest uses $Gain_R\ (X_i, v)$ at each node v and adds new features to F if those features provide new predictive information. Different coefficients of regularization are assigned to variables based on the importance score of an initial random forest. According to Deng and Runger [9] this process allows for the most important bands to have an advantage to enter F. Regularized random forest was implemented using the RRF package in R [8].

2.4 Recursive feature elimination

Díaz-Uriarte and Alvarez de Andrés [10] developed a backward elimination procedure using random forest to select an optimal subset of predictor variables for a microarray application. The method entails fitting the original dataset using random forest and obtaining a permutation based ranked importance for all the variables in the dataset. Subsequently, 20% of the lowest ranked variables are dropped and a new random forest is fitted on the reduced dataset. This process is repeated and the dataset with the lowest OOB error is then selected as the optimal subset of predictor variables.

2.5 Boruta

Boruta is a wrapper built around random forest that evaluates variable importance by creating an ensemble of corresponding artificially added 'shadow' variables randomly sampled from the dataset. Using this extended dataset, Boruta computes and then iteratively compares Z-scores between each variable and the shadow variable. Variable importance is then assessed by comparing variables in the original dataset with variables in the randomised dataset [11, 12]. Many random forest models are run iteratively until variables are classified as *Confirmed*, *Rejected*, or *Tentative* [13].

3. RESULTS

Using the spectral data from week one to train each of the models, and spectral data from week two to test each of the models, the overall classification accuracies using random forest and the three wrapper methods are presented in Table 1. It is clear that Boruta provided the best classification results using just 17 of the 1 769 bands, which is the equivalent of 0.96% of the original hyperspectral dataset. We further evaluated the individual class accuracies for the random forest algorithm and the three wrapper methods (Table 2). Again, Boruta provided better classification accuracies than the other methods considered in this study, although using all the bands with random forest provided slightly better results with the independent test data for the healthy class.

4. CONCLUSIONS

The results show that wrappers used within a random forest framework provide high classification accuracies. More importantly, these wrappers are robust and provide a significant reduction in data dimensionality, making data processing more efficient. The Boruta algorithm has produced the best overall classification results, demonstrating its usefulness in discriminating healthy and stresses *P. radiata* seedlings.

5. ACKNOWLEDGEMENTS

This research was funded by Stellenbosch University Research Subcommittee A, and the DAAD (German Academic Exchange Service) and NRF (National Research Foundation).

Table 1: Classification results for random forest and the three wrapper methods. The number of bands refers to the subset of bands used for classification.

Overall classification accuracies	Training OOB error (Week1)		Test (Week2)		Number of bands
	Accuracy	Kappa	Accuracy	Kappa	
Random forest (no variable selection)	0.81	0.63	0.82	0.64	1769
Regularized random forest	0.81	0.62	0.82	0.64	17
Boruta	**0.85**	**0.70**	**0.84**	**0.68**	17
Recursive feature elimination	0.82	0.64	0.79	0.58	32

Table 2: Classification results for the individual classes. The two classes are healthy (H) and stressed (S).

Class accuracies	Training (OOB error)		Test	
	H	S	H	S
Random forest (no variable selection)	0.86	0.77	0.82	0.82
Regularized random forest	0.83	0.79	0.79	0.86
Boruta	**0.91**	**0.81**	0.81	0.87
Recursive feature elimination	0.85	0.80	0.72	0.91

6. REFERENCES

[1] M.B. Kursa, A. Jankowski, and W.R. Rudnicki, "Boruta – A system for feature selection," *Fundamenta Informaticae*, vol. 101, pp. 271-285, 2010.

[2] M. Pal, and G.M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, pp. 2297-2307, 2010.

[3] F.A. Mianji, and Y. Zhang, "Robust hyperspectral classification using relevance vector machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 2100-2112, 2011.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[5] R. Ismail, and O. Mutanga, 2011, "Discriminating the early stages of *Sirex noctilio* infestation using random forest and shortwave infrared (SWIR) wavelengths," *International Journal of Remote Sensing*, vol. 32, pp. 4249-4266, 2011.

[6] P. Walker, "Guidelines for post processing ASD FieldSpec Pro and FieldSpec 3 spectral data files using the FSF MS Excel template," V03.1, Natural Environment Research Council Field Spectroscopy Facility, Edinburgh, UK, 2009.

[7] A. Liaw, and M. Wiener, "Classification and regression by randomForest". R News, vol. 2, pp. 18-22, 2002.

[8] R Development Core Team. "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org/, 2011.

[9] H. Deng, and G. Runger, "Gene selection with regularized random forest," *arXiv:1209.6425*, 2012.

[10] R. Díaz-Uriarte, S. Alvarez de Andrés. "Gene selection and classification of microarray data using random forest". BMC Bioinformatics, vol 7, no. 3, pp. 1-13, 2006.

[11] M.B. Kursa, A. Jankowski, W.R. Rudnicki. "Boruta – A system for feature selection". Fundamenta Informaticae, vol 101, pp. 271-285. 2010.

[12] M.B. Kursa, W.R. Rudnicki. "Feature selection with the Boruta package". Journal of Statistical Software vol 36, no. 11, pp. 1-13, 2010.

[13] M.B. Kursa. "Important attribute search using Boruta algorithm." In: Package 'Boruta'.

http://cran.r-project.org/web/packages/randomForest/index.html, 2012.