EMBEDDED FEATURE SELECTION OF HYPERSPECTRAL BANDS WITH BOOSTED DECISION TREES

Sildomar T. Monteiro and Richard J. Murphy

Australian Centre for Field Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney, NSW 2006, Australia
s.monteiro@acfr.usyd.edu.au

ABSTRACT

Feature selection is an important step in hyperspectral analysis using machine learning for many applications, in particular to avoid the curse of dimensionality when there is limited available ground truth. This paper presents an approach to select hyperspectral bands using boosting. Boosting decision trees is an efficient and accurate classification technique that has been applied successfully to process hyperspectral data. The learned structure of the trees can provide insight about which bands are more relevant for the classification. We develop a method that takes into account the improvement obtained by each split of the tree ensemble and calculates a relative importance measure of the input features. The method was evaluated using hyperspectral data of rock samples from an iron ore mine in Australia. We show that by retaining only the most relevant features it is possible to reduce the computational load while retaining classification performance.

Index Terms— Boosting, decision trees, feature selection, hyperspectral data

1. INTRODUCTION

Land cover classification methods exist allowing the identification of different materials from hyperspectral data, e.g., soil, vegetation and minerals [1]. Hyperspectral data is characterized by presenting a high number of spectral bands (high-dimensional features, some highly correlated). In most practical applications, the number of ground-truth samples is small (few number of training labels). Although the problem of selecting relevant features from hyperspectral data sets has received much attention recently, e.g. [2], the most suitable feature selection method to automatically identify spectral features is still not well defined in the literature.

The selection of an a priori optimal set of bands would greatly facilitate classification of hyperspectral data, by avoiding the curse of dimensionality and band-to-band correlation [3]. However, the selection of fewer features is important not only for computational issues. If a small number of input features is sufficient to robustly identify materials of interest, it might be possible to optimize the data acquisition process for the relevant wavelengths. Potential applications are in robotics navigation and mining, in which the use of hyperspectral technology has been restricted due to the complexity and cost of data acquisition. The ultimate goal is to identify narrow regions in the spectrum relevant for specific applications, which would allow for the design and construction of imaging systems based on optical filters specifically tuned to these wavelengths. The optimized system would allow timely and efficient spectral data acquisition by avoiding the burden of acquiring irrelevant bands. Emerging sensor technologies, e.g. [4], provide such a tunable system of multispectral bands.

In this paper, we propose to apply a method using boosted decision trees to rank the input features based on their relevance to the task. Boosting has proved to be a robust method to classify hyperspectral data of rock types [5]. Boosting trees inherits the favourable characteristics of single trees, such as robustness and interpretability. At the same, it mitigates some of the disadvantages of trees, such as low accuracy and high variance. Although decision trees are interpretable, i.e. the internal structure of the learned tree can identify the relevant features, boosted trees require a different analysis. A naive implementation based on feature count has been used for hyperspectral analysis in [6]. However, our approach takes into account the improvement associated with the features being selected by the trees, and extends the analysis to boosted trees. This leads to an improved feature ranking which is embedded in the classification procedure of boosted decision trees.

2. BACKGROUND

Let us consider that the hyperspectral data is represented by a vector $x_i \in \Re^p$ comprising p spectral bands (features). The training set is composed of pairs $\langle (x_1, y_1), \ldots, (x_n, y_n) \rangle$ of

This work has been supported by the Australian Centre for Field Robotics and the Rio Tinto Centre for Mine Automation.

n labelled examples, in which each instance $i = \{1, \ldots, n\}$ can be assigned to a label y. In the multi-class case, the target label set can be defined by assigning each class label to an integer $y_i \in \{1, 2, \cdots, C\}$ with the number of classes $C \geq 3$.

Typically, the number of ground-truth samples n is small compared to number of features p, $p > n$. Moreover, the intrinsic dimensionality of hyperspectral data is often much smaller than its nominal dimensionality. Feature selection methods are categorized as filter, wrapper or embedded [7]. The feature relevance method proposed here is an embedded method. It assigns a score $S(i)$ to each input feature indicating the relevance or importance of that feature for classification. The scores can then be ranked and features can be selected based on a threshold; features with score below the threshold are eliminated.

3. BOOSTING DECISION TREES

The idea of Boosting is to train many "weak" learners on various distributions (or set of weights) of the input data and then combine the resulting classifiers into a single "committee" [8]. A weak learner can be any classifier whose performance is guaranteed to be better than a random guess. There are many different variants of boosting algorithms. In this study, we investigate a version called LogitBoost, which can be used to directly classify multiple classes. Logitboost, uses stagewise optimization of the maximum likelihood through adaptive Newton steps to fit additive logistic regression models [9]. For weak learners, the present study utilizes regression stumps, which can be viewed as binary decision trees with only one node [10].

3.1. Relative importance of features

It is useful to learn the relative importance of individual input variables in predicting the response of the tree ensemble. Assuming that the model learned by the tree is a reasonable approximation of the true function, this interpretation of the tree can give an insight about the underlying relationship between the input variables and the output. For a single tree, a measure of the relevance, as proposed by [11], is

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \delta(v(t) = l) \tag{1}$$

where the sum is over the nonterminal nodes of the tree T and $v(t)$ is the splitting variable associated with node t. The empirical improvement $\hat{i}_t^2$ for choosing that variable is the squared error risk associated with the split [12].

The linear combination of trees must be interpreted differently than a single decision tree. For the ensemble of decision trees obtained by boosting, the relative importance measure can be generalized by simply averaging over the trees [13]. Similarly, for the multiclass case, the importance measure of the separate models for each class can be obtained by averaging over all the classes. Finally, the relevance measure for each input variable is given by the square root of the respective relevant measure averaged out over all trees and all classes. Since the relevance measure of the inputs are relative to each other, their values can then be normalized to fall between 0 and 1 or 100.

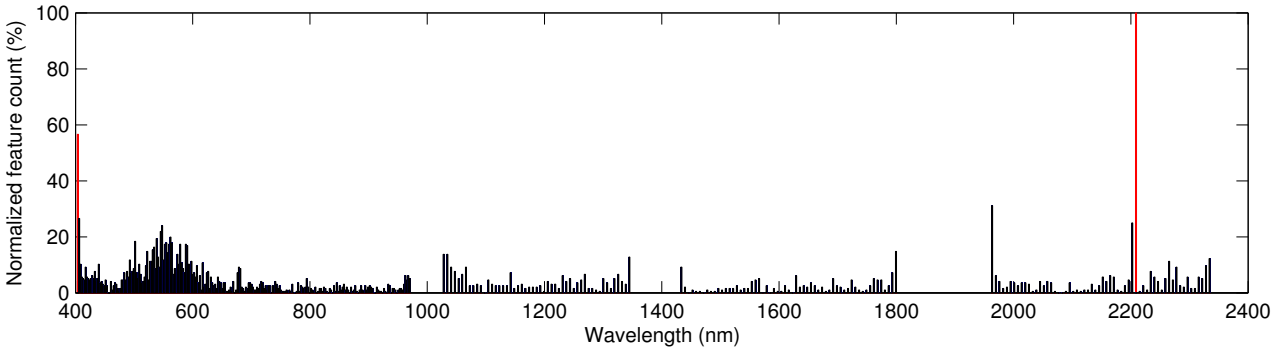
4. EXPERIMENTS

The algorithms were evaluated using hyperspectral data of rock samples collected from an iron ore mine located in the Pilbara region of Western Australia. The samples comprise several types of rocks typically found in that region, including rocks rich in martite, goethite and kaolinite. The hyperspectral data was acquired using an ASD (Analytical Spectral Devices Inc.) field spectrometer comprising wavelengths from the visible (350 nm) to the SWIR (2500 nm) regions of the spectrum at nominal 1 nm intervals. The number of bands was reduced to 429 (wavelengths from 404 nm to 2334 nm) to approximate the typical spectral range of commercially available hyperspectral imaging systems. This was done by downsampling the data to 2 nm intervals on the visible region and to 6.5 nm in the SWIR. Bands with very low signal-to-noise ratio, such as those affected by water-vapour absorption, were removed. The hyperspectral data was collected under several different illumination and physical conditions, in an attempt to simulate some of the environmental conditions expected in a mine site, i.e. under conditionss of direct sunlight, shadow and different vieweing angles.

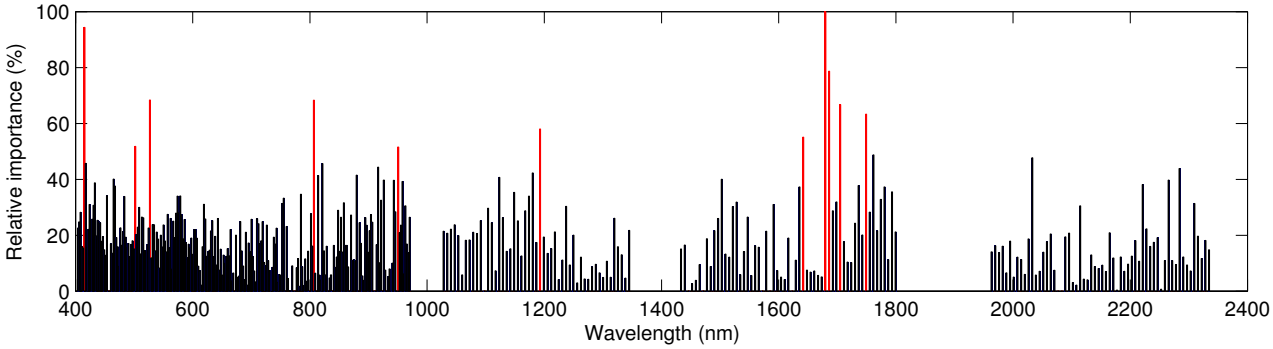
Boosting decision trees require two parameters to be defined, the depth (number of nodes) of the trees and the number of weak learners (trees). In this study, the number of weak learners was kept constant in all experiments to 100. The number of nodes in the trees can vary from one to a full pruned tree. This study was done using trees with a single node, also called regression stumps. The experiments were performed using stratified 4-fold cross-validation.

A simplistic approach to interpret the tree classification is to count the number of times a given feature is being used to split the data; the result of this analysis is shown in Fig. 1(a) for the simultaneous classification of 9 classes of rocks. The relative importance method based on empirical risk was compared with the simple feature count approach. An illustration of the relative importance of features is shown in Fig. 1(b). It is noticeable that the same tree ensemble interpreted by the two different methods can produce very different feature rankings. The selection criterion used was to keep the features that have over 50% normalized count/importance.

Several statistical metrics were calculated to assess the performance of the models: accuracy, precision, recall, F-score, kappa coefficient, and area under the ROC curve (AUC); for more details, see [14]. Each metric can capture a different aspect of the performance of the model. The performance was calculated by combining the results of all folds in the cross-validation and comparing them with ground-truth labels that were analysed by mine geologists. After the most



(a) Normalized count of features used by the trees



(b) Relative importance of features used in the trees

Fig. 1. Comparison of relative importance of features against simple feature count. The most important bands based on a threshold of 50% are highlighted in red.

relevant features were selected based on the 50% threshold, the reduced feature set was used to learn a new boosted decision tree. A summary of the performance using the most relevant features is presented in Tables 1 and 2 , for all 9 classes. The results of the relative importance method are comparable, averages within approximately 10% overall, to the obtained by boosted trees with the same parameters but using all input bands. Note that lowering the threshold to include more features tends to improve the performance of both methods, with the penalty of increased computational load.

5. CONCLUSIONS

This paper presents a method to select relevant features from hyperspectral data by using the internal learned structure of an ensemble of decision trees trained by boosting. To our knowledge, this is the first paper to report on the feature selection of hyperspectral data using the relative importance of features as indicated by boosted trees. Note that the procedure is not restricted to classification problems and can be applied to regression trees as well.

The results are encouraging. The boosted decision trees are able to identify the relevant features while learning a classification model. Therefore, the procedure is very efficient by nature. The most relevant features can be selected from the ranked list of features generated by the method. The reduced number of hyperspectral bands was sufficient to distinguish the 9 classes of rocks. The reduced feature set also allows much faster training than required when using all available bands, which can be advantageous if further fine-tuning of the classifier parameters is desired.

In this study, the relevance threshold indicating the number of features to be retained was predetermined ad hoc. Research is ongoing to allow the method to automatically identify the most suitable number of features for the task; an alternative approach using statistical tests was proposed in [15]. Nevertheless, due to instability in the models, especially in small trees, the feature selection results might also be unstable. Extended experiments are planned to assess this issue; note that the statistical tests method may alleviate this problem.

6. REFERENCES

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques

Table 1. Classification performance of 2 bands selected using the normalized count of features

	WRC	BIF	GOL	MAR	SHL	SHN	NS4	NS3	CHT	Overall†
Accuracy	0.9017	0.8133	0.9072	0.8745	0.8472	0.8810	0.9760	0.9640	0.9138	0.5393 (0.0517)
Precision	0.6369	0.2065	0.5630	0.7633	0.3232	0.3780	0.1000	0.5172	0.2889	0.4197 (0.2154)
Recall	0.7519	0.1624	0.6700	0.8182	0.3048	0.3483	0.0714	0.4412	0.2167	0.4205 (0.2691)
F-score	0.6897	0.1818	0.6119	0.7898	0.3137	0.3626	0.0833	0.4762	0.2476	0.4174 (0.2408)
Kappa	0.6318	0.0782	0.5596	0.7004	0.2278	0.2971	0.0715	0.4577	0.2029	0.3585 (0.2366)
AUC	0.8395	0.5355	0.8031	0.8577	0.6111	0.6433	0.5307	0.7127	0.5896	0.6804 (0.1278)

† The standard deviation in the classification performance over all classes is presented in brackets

Table 2. Classification performance of 11 bands selected using the relative importance of features

	WRC	BIF	GOL	MAR	SHL	SHN	NS4	NS3	CHT	Overall†
Accuracy	0.9312	0.8504	0.9432	0.8963	0.9159	0.9672	0.9934	0.9847	0.9389	0.7107 (0.0448)
Precision	0.7652	0.4180	0.7353	0.7924	0.6111	0.8933	1.0000	0.8846	0.5556	0.7395 (0.1836)
Recall	0.7594	0.4359	0.7500	0.8674	0.7333	0.7528	0.5714	0.6765	0.3333	0.6533 (0.1731)
F-score	0.7623	0.4268	0.7426	0.8282	0.6667	0.8171	0.7273	0.7667	0.4167	0.6838 (0.1560)
Kappa	0.7221	0.3408	0.7107	0.7542	0.6190	0.7992	0.7242	0.7589	0.3865	0.6462 (0.1678)
AUC	0.8599	0.6735	0.8585	0.8877	0.8365	0.8716	0.7857	0.8365	0.6573	0.8075 (0.0855)

† The standard deviation in the classification performance over all classes is presented in brackets

for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, no. Supplement 1, pp. S110 – S122, 2009.

[2] M. Pal and G. Foody, "Feature selection for classification of hyperspectral data by svm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, 2010.

[3] B. Tso and P. M. Mather, *Classification Methods for Remotely Sensed Data*, 2nd ed. CRC Press, 2009.

[4] J. M. Eichenholz and J. Dougherty, "Ultracompact fully integrated megapixel multispectral imager," in *Proc. of the Integrated Optics: Devices, Materials, and Technologies XIII*, vol. 7218, 2009.

[5] S. Monteiro, R. Murphy, F. Ramos, and J. Nieto, "Applying boosting for hyperspectral classification of ore-bearing rocks," in *IEEE Workshop on Machine Learning for Signal Processing*, 2009.

[6] H. R. Bittencourt and R. T. Clarke, "Feature selection by using classification and regression trees," in *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 35, no. 7, 2004, pp. 66–70.

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[8] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the 13th International Conference on Machine Learning*, 1996, pp. 148–156.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[10] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, 2007.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Chapman and Hall (CRC Press), 1984.

[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009.

[15] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *Journal of Machine Learning Research*, vol. 10, pp. 1341–1366, 2009.