CALIBRATING PROBABILITIES FOR HYPERSPECTRAL CLASSIFICATION OF ROCK TYPES

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

This paper investigates the performance of machine learning methods for classifying rock types from hyperspectral data. The main objective is to test the impact on classification error rate of calibrating the model's output into class probability estimates. The base classifiers included in this study are: boosted decision trees, support vector machines and logistic regression. The standard algorithm for some of these methods provides a non-probabilistic, hard decision as output. For those methods, posterior class probability estimates were approximated by fitting a sigmoid function to the classifier predictions. To perform multi-class classification, a one-versus-all approach was used. The different methods were compared using hyperspectral data acquired from ore-bearing rocks under different environmental conditions. The calibration of class probabilities improved the overall performance for almost all algorithms tested; an improvement of over 10% was observed in some cases.

1. INTRODUCTION

Hyperspectral sensors acquire data in hundreds of narrow, contiguous bands at visible, near-infrared (NIR) and shortwave-infrared (SWIR) wavelengths providing a powerful tool for non-destructive analysis of remote samples. Spectral signature analysis of hyperspectral data can be applied to classify samples into categories and produce land cover maps [1]. The hyperspectral classification problem is characterized by having a large number of spectral bands (high-dimensional features, high correlation), various rock categories (multiple classes), and small number of ground-truth samples (limited training labels). Conventional land cover classification methods allow easy distinction among different materials, e.g., bare soil, vegetation and minerals [2]. However, there are still challenges in providing robust and flexible hyperspectral classification algorithms, especially when targets present high degree of spectral similarity and poor signal-to-noise ratio. Such targets pose a difficult problem causing conventional spectral unmixing or statistical analysis methods to perform poorly.

The timely characterization of geology using hyperspectral sensors can be of enormous value for the mining industry, despite the constraint that it only provides information from the rock surface [3]. An accurate understanding of the geology is important during several phases of the mining process, from exploration to processing and reconciliation. Hyperspectral analysis can be particularly useful in open-pit mine operations where the rocks of interest are exposed. It has the potential to provide fast assessment of the identity and distribution of minerals of interest on a mine bench, resulting in more efficient mining and improving the end-product quality and value.

In this paper, we investigate the effectiveness of calibrating class probability estimates from the output of machine learning algorithms in improving the classification of hyperspectral data into multiple discrete categories. We investigate three algorithms for supervised classification and some of their variants: boosting (with decision trees), support vector machines (SVMs) and logistic regression. Experimental results are presented using hyperspectral data of ore samples collected from an open pit mine in Western Australia. The hyperspectral data sets were acquired under different observational conditions to test the performance of the algorithms.

2. PROBABILISTIC HYPERSPECTRAL CLASSIFICATION

Let us consider that the hyperspectral data is represented by a vector $x_i \in \Re^d$ comprising d spectral bands. The training set is composed of pairs $\langle (x_1, y_1), \ldots, (x_n, y_n) \rangle$ of n labelled examples, in which each instance $i = \{1, \ldots, n\}$ can be assigned to a label y. The target label set can be defined as $y_i \in \{-1, +1\}$, for the binary classification problem, or, in the multi-class case, by assigning each label to an integer $y_i \in \{1, 2, \cdots, C\}$ with the number of classes $C \geq 3$.

In a probabilistic framework, the probability of a class C

This work was supported by the Rio Tinto Centre for Mine Automation and by the ARC Centres of Excellence programme, funded by the Australian Research Council and the New South Wales State Government.

occurring is defined simply as $P(y = C)$. The class probabilities for all classes sum to one $P(y = 1) + P(y = 2) + ... + P(y = C) = 1$. If ahead of making any measurements the classes are equally likely to occur, the prior probabilities of class membership are equal $P(y = 1) = P(y = C) = 1/C$. The posterior probability of class membership is then obtained from Bayes rule: $P(y = C\,|x\,) = p(x\,|y = C)P(y = C)\,/p(x)$. We focus our attention to discriminant approaches, i.e. methods that assign classes based on posterior probabilities with no consideration for the class conditional densities (or distributions) which generate the measurement features.

Most classification algorithms give a hard decision as output. Typically, a set decision threshold is applied on the classifier response, neglecting the relative confidence in the classification. Nevertheless, it is possible to estimate posterior probabilities by using a sigmoid function to map the classifier's output $f(x)$, before the hard decision is made, into $P(y = C\,|x\,)$ [4]. The sigmoid model is calculated in a parametric form as follows: $\hat{P}\,(y = C\,|f(x)\,) = 1/(1 + \exp\,(Af(x) + B))$. The parameters A and B can be calculated by minimizing the negative log likelihood using Newton's method with backtracking [5].

There are several schemes for coding and combining the outputs of binary classifiers to solve the multi-class problem [6]. The two most widely used strategies are the one-versus-all and the one-versus-one approaches [7]. The present study uses a one-versus-all approach which learns a set of binary classifiers $\{f_1, f_2, \cdots, f_C\}$, where the c-th class is assigned to the positive class, while the others are assigned to the negative class. The prediction of the set of binary classifiers is given by majority voting $y_i^* = \arg\max_{c=1,2,\cdots,C}\{f_c(x_i)\}$. In the case of class probabilities being estimated, a similar method using a winner-takes-all scheme can be applied $y_i^* = \arg\max_{c=1,2,\cdots,C}\left\{\hat{P}(y = c|x_i)\right\}$.

2.1. Boosting

Boosting is a machine learning technique for supervised classification that has become very popular due to its sound theoretical foundation, and also due to many empirical studies showing that it tends to yield smaller classification error rates and be more robust to overfitting than competing methods [8]. 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" [9]. 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 Logitboost and GentleBoost [10]. GentleBoost is binary classifier that was designed to be more numerically stable than the standard AdaBoost algorithm. LogitBoost provides multi-class classification by using a symmetric multiple logistic transformation. Two types of weak learners were tested: single node decision trees (also called decision stumps), and 4-node decision trees.

2.2. Support Vector Machines

SVMs have been shown to be effective for nonlinear classification, regression and density estimation problems. Particularly for hyperspectral classification, several studies have reported accurate, robust models using SVMs, which also benefit from the sparseness of the solutions, e.g. [11]. SVMs were introduced for the binary classification problem by fitting an optimal separating hyperplane between the positive and negative classes with the maximal margin. The classical SVM algorithm is based on convex optimization theory, typically quadratic programming involving inequality constraints. An alternative solution for this problem is the sequential minimal optimization (SMO) algorithm [12]. SMO is an efficient approximation method that scales better than the original quadratic programming formulation. It has been reported to perform well in a number of different data sets. Two different types of kernel functions were tested: Gaussian radial basis function (RBF) and d-th degree polynomial.

2.3. Logistic Regression

Logistic regression is a discriminative method for classification, despite its name. It is a linear model that can naturally provide posterior probability estimates. In theory, logistic regression is more numerically robust than linear discriminant analysis since it relies on fewer assumptions [13]. For multiclass classification, posterior probabilities are calculated through a multiple logistic transformation using a softmax function, much in a similar fashion as used by LogitBoost. Despite their limitations, linear models like logistic regression are surprisingly competitive with far more sophisticated methods and are particularly appropriate for high dimensional feature spaces such as hyperspectral data sets.

3. EXPERIMENTS

For the empirical analysis of the algorithms, we collected representative rock samples from an iron ore mine located in the Pilbara region of Western Australia. This study includes both whole-rock samples and cores acquired using a diamond drill. The samples comprise several rock types and ore minerals typically found in that region, specifically: banded iron formation (BIF), martite, goethite, kaolinite (clay), and mixtures of these.

Data were acquired using an ASD (Analytical Spectral Devices Inc.) field spectrometer. The sensor acquires hyperspectral data from the visible (350 nm) to the SWIR (2500 nm) regions of the spectrum at nominal 1 nm intervals. The data sets were downsampled to 2 nm intervals on the visible region and to 6.5 nm in the SWIR in order to approximate the typical spectral resolution of commercially available hyperspectral imaging systems; thus, the total number of bands was reduced to 429.

Table 1. Results of out-of-sample experiment

		LogisticReg	LogitBoost-S	LogitBoost-DT	GentleBoost-S	GentleBoost-DT	SVM-Poly	SVM-RBF
Standard	Acc	0.1846	0.4782	0.4709	0.3721	0.3823	0.5581	**0.5945**
	F1	0.1340	0.4778	0.4844	0.3621	0.3698	0.4936	**0.5519**
	AUC	0.4587	**0.7956**	0.7849	0.6608	0.6697	0.7609	0.7844
Probabilistic	Acc	–	0.4477	0.4666	0.4244	0.4273	0.6541	**0.7369**
	F1	–	0.4560	0.4810	0.4064	0.4350	0.6102	**0.7001**
	AUC	–	0.8127	0.8065	0.7783	0.7789	0.8777	**0.9016**

Table 2. Results of k-fold cross-validation experiment

		LogisticReg	LogitBoost-S	LogitBoost-DT	GentleBoost-S	GentleBoost-DT	SVM-Poly	SVM-RBF
Standard	Acc	0.7162	0.8461	0.8417	0.7500	0.8013	0.7882	**0.8832**
	F1	0.6427	0.8319	0.8246	0.7154	0.7585	0.6976	**0.8455**
	AUC	0.8579	0.9685	**0.9708**	0.8527	0.8831	0.8471	0.9423
Probabilistic	Acc	–	0.8133	0.8515	0.8264	0.8592	0.8286	**0.9301**
	F1	–	0.8025	0.8482	0.8189	0.8485	0.8029	**0.9389**
	AUC	–	0.9653	0.9643	0.9662	0.9710	0.9853	**0.9928**

The hyperspectral data sets were collected under different illumination and physical conditions, in an attempt to reproduce in a controlled manner some of the environmental characteristics of a mine site [14]. Specifically, five sets of hyperspectral data were compiled: a) core samples in artificial illumination (halogen lamp); b) core samples in full sunlight; c) core samples in full sunlight from different angles; d) core samples in shade; e) whole-rocks in artificial illumination.

The experiments were divided in two parts. The first was an out-of-sample analysis. Classification models were trained using hyperspectral data only from core samples under artificial illumination and then evaluated on the other data sets, cases (b)–(e) above. In the second part, the classification algorithms were evaluated using k-fold cross-validation, with $k = 8$. The data sets for cross-validation were selected using stratified random sampling. For each algorithm, the model parameters were optimized to maximize the overall performance on each case. The metrics chosen for the analysis are accuracy, F1-measure (a weighted combination of precision and recall) and area under the ROC curve (AUC).

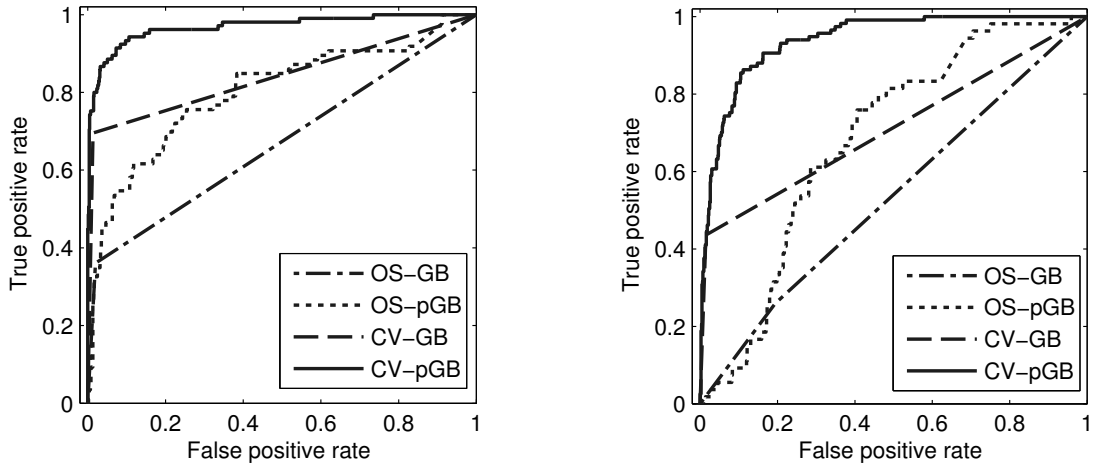
A summary of the results is presented in Table 1 and Table 2, for the out-of-sample and cross-validation experiments respectively. The results presented are averages of all individual classes and the top scores for each metric (rows) are highlighted in bold. In the tables, "standard" versions of the algorithms are the ones without calibration to probabilistic outputs. Note that the logistic regression output is a probability estimate by nature. Decision stumps (S) and 4-node decision trees (DT) were used as weak learners for boosting. The SVMs used n-degree polynomials (Poly) and Gaussian radial basis functions (RBF) as kernels.

To illustrate the variation in AUC performance, ROC curves for two of the most representative rock types (shale and BIF) using GentleBoost (GB) and its probabilistic calibrated version (pGB) are shown in Fig. 1. Both out-of-sample (OS) and cross-validation (CV) analysis are plotted. It is noticeable that the hard-decisions (non-probabilistic outputs) produce poor ROC curves since they can only produce a single point in the space. The probabilistic approach, on the other hand, is more amenable to ROC analysis by allowing for several decision thresholds and, thus, more points in the curve.

Apart from LogitBoost, all the other algorithms benefited from the calibration of their predicted output. Especially for the AUC metric, the performance of the algorithms was improved significantly, by more than 10% in some cases. This indicates not only an improvement on the binary classification for individual classes, but also on the final decision of the multi-class classifier; the winner-takes-all approach using probabilities was far superior to the voting scheme using hard decisions. The LogitBoost with decision stumps did not improve but had its performance slightly degraded. This seems to be due to the the fact that LogitBoost already uses the winner-takes-all approach and probability estimations internally in its standard training algorithm.

4. CONCLUSIONS

This paper compared seven different variants of machine learning algorithms for hyperspectral classification of rocks. The encouraging results demonstrate the importance of calibrating class probability estimates from the classifiers' outputs. In the hyperspectral data set tested, logistic regression did not achieve the same performance level as the other more sophisticated algorithms. It seems the learning of the models was hindered by the high-dimensional feature space and the limited number of samples; the latter was evident especially in the out-of-sample experiment. Despite producing the best performance overall, SVMs using Gaussian RBFs require ex-



(a) Shale (b) Banded iron formation

Fig. 1. ROC curves for GentleBoost classification using decision stumps on two representative rock types

tensive parameter tuning for each case to achieve the levels of performance reported. On the other hand, SVMs using first-order polynomials and the standard multi-class LogitBoost can be trained very efficiently (orders of magnitude faster on the same data set), and both are not nearly as sensitive to the sole parameter they require to be tuned.

Future work includes investigating methods to integrate spatial information to the spectral analysis in order to improve the accuracy of maps showing the spatial distribution of minerals. Further tasks, such as fusion with different sensors, should benefit from the calibrated class probabilities adopted for hyperspectral classification.

5. REFERENCES

[1] R. N. Clark, G. A. Swayze, K. E. Livo, R. F. Kokaly, S. J. Sutley, J. B. Dalton, R. R. McDougal, and C. A. Gent, "Imaging spectroscopy: Earth and planetary remote sensing with the usgs tetracorder and expert systems," *Journal of geophysical research*, vol. 108, no. E12, pp. 5.1–5.44, 2003.

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

[3] H. M. Rajesh, "Application of remote sensing and GIS in mineral resource mapping - an overview," *Journal of Mineralogical and Petrological Sciences*, vol. 99, no. 3, pp. 83–103, 2004.

[4] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, P. J. Bartlett, B. Schlkopf, D. Schuurmans, and A. J. Smola, Eds. Cambridge: MIT Press, 2000, pp. 61–74.

[5] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, no. 3, pp. 267–276, 2007.

[6] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Machine learning*, vol. 47, no. 2–3, pp. 201–233, 2002.

[7] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[8] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. of the 23rd International Conference on Machine Learning*, 2006, pp. 161–168.

[9] 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.

[10] 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.

[11] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.

[12] J. C. Platt, *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*, ser. Advances in kernel methods: Support Vector Learning. MIT Press, 1999, ch. 12, pp. 185–208.

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

[14] S. Schneider, R. Murphy, S. T. Monteiro, and E. W. Nettleton, "On the development of a hyperspectral library for autonomous mining," in *Australasian Conference on Robotics and Automation*, 2009.