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

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 high 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 where conventional spectral unmixing or statistical analysis methods often 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 effect of calibrating probability estimates from the output of machine learning techniques for improving the classification of hyperspectral data acquired of ore-bearing rocks into discrete categories. The algorithms—naive Bayes, Boosting, and SVM—were assessed using hyperspectral data sets of ore samples collected from an open pit mine in Western Australia. Hyperspectral data sets acquired under different environmental conditions were used to compare 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

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.

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 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, ignoring the relative confidence in the classification. Nevertheless, posterior probabilities can be estimated by using a sigmoid function that maps the classifier outputs $f(x)$, 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 classifiers is given by majority voting $y_i^* = \arg\max_{c=1,2,\cdots,C}\,\{f_c(x_i)\}$.

2.1. Naive Bayes

The naive Bayes classifier is based on the assumption that the inputs are conditionally independent in each class. The estimation of individual class-conditional marginal densities can then be performed by one-dimensional kernel density estimates [8]. Despite this optimistic assumption, this method is surprisingly competitive with far more sophisticated methods and is particularly appropriate for high dimensional feature spaces such as hyperspectral data sets. The naive Bayes classifier can naturally provide probabilistic outputs and handle multiple classes.

2.2. 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 [9]. 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" [10]. 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 two versions called AdaBoost and Logitboost [11].

2.3. 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. [12]. 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. We focus on a different formulation known as Least Squares Support Vector Machines (LS-SVMs) [13], which present lower computational complexity

and may scale better for high-dimensional problems. Three different types of kernel functions were tested: radial basis function (RBF), polynomial (Poly) and neural network (NN).

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 ore minerals typically found in that region, specifically: banded ironstone formation, martite, goethite, kaolinite (clay), and mixtures of these rock types.

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 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.

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. 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.

Two sets of experiments were done. The first was an out-of-sample analysis. Classification models were trained using hyperspectral data from the core samples in artificial illumination, case (a) above, and then these models were evaluated on the other data sets. In the second test, the classification algorithms were evaluated using k-fold cross-validation. The data sets for cross-validation were selected using stratified random sampling. The total classification accuracy for each test is a micro-average of per-class values. For all algorithms, the respective parameters were optimized to maximize accuracy for each test case and the best results were kept. A summary of the results is presented in Table 1 and Table 2 for the two scenarios tested. The statistical results are the averages over all test cases of accuracy, precision, recall and F-measure.

Table 1. Summary of results: out-of-sample analysis, averaged over the different environmental conditions

	LogitBoost(DS)	AdaBoost(DT)	NaiveBayes	SVM(RBF)	SVM(Poly)	SVM(NN)
Accuracy	0.5044	0.5102	0.3968	0.5407	0.5523	0.3052
Precision	0.5865	0.6229	0.3649	0.6995	0.6243	0.4589
Recall	0.5002	0.5051	0.3819	0.5933	0.5300	0.3622
F-measure	0.4932	0.5155	0.3376	0.5919	0.5184	0.3154

Table 2. Summary of results: k-fold cross-validation using all data sets

	LogitBoost(DS)	AdaBoost(DT)	NaiveBayes	SVM(RBF)	SVM(Poly)	SVM(NN)
Accuracy	0.8144	0.8406	0.5852	0.8930	0.9072	0.4585
Precision	0.8211	0.8437	0.5244	0.9003	0.9137	0.5178
Recall	0.7785	0.8181	0.4844	0.8995	0.9065	0.4801
F-measure	0.7958	0.8296	0.4902	0.8989	0.9093	0.4720

4. CONCLUSIONS

In this paper, we evaluated six different variants of machine learning algorithms for hyperspectral classification of rocks. The out-of-sample experiment was the most challenging for the algorithms. The cross-validation analysis indicates that using spectra acquired under different conditions improves the performance of the machine learning models. The naive Bayes and SVM using

neural network kernels did not achieve the same performance level as the other algorithms. Overall, SVMs using polynomial kernels outperformed the other methods.

The learning of the models was hindered by the limited number of samples and high-dimensional feature space. Despite providing less accurate models, boosting requires less parameter tuning to produce competitive results. Both boosting methods presented similar results. In the SVMs, the more complex kernels did not perform as well as the simple polynomial kernel. In some cases a first-order polynomial provided the best model. This seems to be due to the high dimensionality of the input features that may not require a complex mapping to a higher dimensional kernel space to provide good classification. Also, the SVM model performance is dependent on the training method used, the least squares approach in this case, which was not as effective to train the neural network kernel.

Future works include investigating methods to integrate spatial information to the spectral data in order to improve the accuracy of maps showing the spatial distribution of the minerals. This should be facilitated by the probabilistic framework adopted for hyperspectral classification.

5. REFERENCES

[1] Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Raymond F. Kokaly, Steve J. Sutley, J. Brad Dalton, Robert R. McDougal, and Carol 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] John C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, Peter J. Bartlett, Bernhard Schlkopf, Dale Schuurmans, and Alex J. Smola, Eds., pp. 61–74. MIT Press, Cambridge, 2000.

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

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

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

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

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

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

[11] Jerome 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.

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

[13] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.