Boosted Band Ratio Feature Selection for Hyperspectral Image Classification

Zhouyu Fu Terry Caelli Nianjun Liu Antonio Robles-Kelly
NICTA*, RSISE Bldg. 115, Australian National University, Canberra, ACT 0200, Australia

Abstract

Band ratios have many useful applications in hyperspectral image analysis. While optimal ratios have been chosen empirically in previous research, we propose a principled algorithm for the automatic selection of ratios directly from data. First, a robust method is used to estimate the Kullback-Leibler divergence(KLD) between different sample distributions and evaluate the optimality of individual ratio features. Then, the boosting framework is adopted to select multiple ratio features iteratively. Multiclass classification is handled by using a pairwise classification framework. The algorithm can also be applied to the selection of discriminant bands. Experimental results on both simple material identification and complex land cover classification demonstrate the potential of this ratio selection algorithm.

1. Introduction

The development of image sensor technology has made it possible to capture image data in hundreds of bands covering a broad spectrum of wavelength range. The rich information available in hyperspectral imagery has posed significant opportunities and challenges for feature extraction and classification. Many algorithms have been proposed for this purpose, such as Principle Component Analysis, (Linear) Discriminant Analysis, Decision Boundary , Projection Pursuit, and kernel methods[1]. All these algorithms treat the raw pixel spectra as input vectors in high dimensional spaces and look for linear or nonlinear mappings to the feature space (often with reduced dimensionality) by optimizing certain criterion, leading to statistically optimal solutions to classification.

An alternative way is to use simple features that are physically meaningful. One such feature that has received much attention in the remote sensing community is the band ratio - the ratio of spectral values between two different bands. The important property of such ratios is that some materials can be identified by simply observing a single ratio. For example, green vegetation can be differentiated from soil and other surface covers by the Normalized Vegetation Index(NDVI) - the ratio between a near infrared band and a visible red band. This has been used extensively for the estimation of vegetation coverage over the surface[2]. Another advantage of the band ratio is its invariance to shading, as the geometry factor related to shading is constant for different bands. This is an attractive feature for terrestrial hyperspectral imaging, where the surface geometry of the object under study plays a significant role in what is detected by the camera.

However, there is still a lack of technical justification for using band ratios. Ratios are typically chosen from empirical observations or from domain knowledge. Further, no algorithms have been reported that can automatically derive the optimal ratios from spectral data. In this paper, we exploit ratios for feature selection and classification by learning the optimal ratio features. Besides selecting a single ratio for coarse detection, our algorithm is capable of combining multiple ratios to achieve more accurate classification. To do this, we adopt a boosting framework to select ratio features iteratively. A robust method is proposed to estimate the Kullback-Leibler divergence (KLD) between different sample distributions and the ratio feature with maximum KLD is selected at each iteration. Finally, we apply a Support Vector Machine(SVM) to the selected ratio features for training the classifier. The algorithm can be naturally generalized to handle the classification of multi-class samples by casting it into a pairwise classification framework. Moreover, the above procedures can also be applied to the selection of optimal bands.

The remainder of this paper is organized as follows. Section 2 describes our algorithm for feature selection and classification. Section 3 presents the experimental results. In the last section we conclude on the work presented here.

2. Algorithm Description

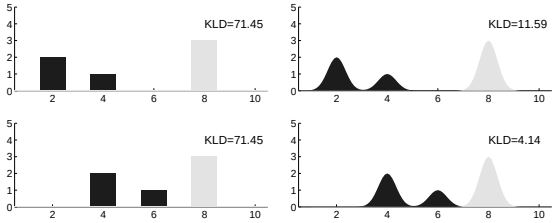
The following two sections are focused on binary classification. Generalization to multiclass cases is addressed in Section 2.3.

2.1. Optimal Criterion for Ratio Features

Instead of directly dividing the value of one band by the other, we use an alternative definition of the band ratio given by

$$r(\lambda_i, \lambda_j) = \frac{x(\lambda_i) - x(\lambda_j)}{x(\lambda_i) + x(\lambda_j) + \epsilon} \quad (1)$$

*National ICT Australia is funded by the Australian Governments Backing Australia's Ability initiative, in part through the Australian Research Council.



(a) (b)

Figure 1. (a) Histograms for two sample distributions. (b) Modified histograms of (a).

where λ_i and λ_j are two arbitrary bands where the ratio is taken, ϵ is an infinitesimal term to ensure numerical stability. As x's are nonnegative values, ratios defined in the above form are bounded in the closed interval of $[-1, 1]$ so that the following estimation of ratio distribution will be computationally feasible.

Suppose each raw spectrum contains D bands, and, hence, there are $D \times (D-1)/2$ possible combinations of two-band ratios. For each candidate ratio, we can compute the distributions of ratio values for positive and negative samples. We then assume that the optimal ratio should maximize the distance between the distributions for both, the positive and negative classes. There are a number of criteria that can be used to measure the distance between two distributions, like the Bhattacharya distance, various types of correlation coefficients and the Kullback-Leibler Divergence (KLD). The first two measures are based on the assumption of Gaussian distributions, while the KLD can be applied to any type of distribution. As ratio distributions are usually too complex to be modelled by a single Gaussian, we use the KLD as the criterion for measuring the distance between two sample classes:

$$\mathrm{KL}(p^+(r), p^-(r)) = \int_r p^+(r) \log \frac{p^+(r)}{p^-(r)} \mathrm{d}r \qquad (2)$$

where $p^+(r)$ and $p^-(r)$ are the Probabilistic Distribution Functions (*PDF*) for positive and negative samples, respectively.

Since Equation (2) does not have a closed form solution for arbitary PDFs, we use histograms to approximate $p^+(r)$ and $p^-(r)$ - a conventional solution for low-dimensional PDFs-. However, we note that the histogram representation causes problems with sparse data where many bins are occupied by only a few samples. In this case, some terms in Eq. 2 could be zero at some intervals. Adding an infinitesimal value to the zero term overcomes this problem numerically but adds to the inaccuracy of estimation. This can be illustrated by the toy example in Figure 1(a). Here, we show two different sample distributions where grey bins are occupied by positive samples and black bins are occupied by negative samples. The two situations only differ by the distance between positive bins and negative bins. The one with closer distance between the two classes should have greater KLD, nonetheless, the divergences computed from histogram approximation are the same for both cases. This is due to the quantization effect related to the binning involved in computing the histogram. We can think of the histogram as an approximation of the true PDF in the form of gate functions. The tails of the distribution are cut off outside the current bin. If we approximate each occupied histogram bin with a Gaussian function, the tail distribution will be maintained without affecting the estimation accuracy due to the exponential decay of the function. This is equivalent to running Kernel Density Estimation on the histogram bins using a Gaussian kernel. In this case, KLD can be still be derived in closed form as follows,

$$\mathrm{KL}(h^+(r), h^-(r)) = \sum_{i=1}^{m} h^+(r_i) \log \frac{h^+(r_i)}{h^-(r_i)} \qquad (3)$$

$$h^+(r) = \sum_{i=1}^{m} n_i^+ \exp\left(-\frac{(r-r_i)^2}{2\sigma^2}\right) / Z^+(r)$$

$$h^-(r) = \sum_{i=1}^{m} n_i^- \exp\left(-\frac{(r-r_i)^2}{2\sigma^2}\right) / Z^-(r)$$

where m is the number of bins, r_i is the center of the ith bin, $n_i^+(n_i^-)$ is the number of positive(negative) samples falling into the ith bin, σ is set to half bin width, $Z^+(r)$ and $Z^-(r)$ are normalization terms. As a result, the KLD computed from the modified histogram representation corresponding to the above two cases correctly reflect the divergence of positive and negative sample distributions by taking into account the margins between them. See Figure 1(b).

The above computations can be naturally extended to handle weighted samples by replacing the bin count $n_i^+(n_i^-)$ with the cumulant of the sample weights falling into the i^{th} bin, i.e. $\sum_{x_j \in bin(i)\ \&\ y_j=+1(-1)} w_j$.

2.2. Feature Boosting

We adopt the boosting framework for the selection of multiple ratio features. Boosting is the generalized term for a class of algorithms that combine weak learners into a strong classifier by iteratively selecting the optimal weak learner and updating sample weights. In many cases, a single ratio feature is insufficient to discriminate between two classes of complex spectra. In our problem, each individual ratio feature is regarded as a weak learner. Hence boosting should be able to select a set of ratio features and satisfactory classification can be achieved by combining the selected ratio features.

Here, we use the Realboost algorithm proposed in [4] for feature selection. Compared to the classical Adaboost algorithm[3], Realboost uses real valued output for the weak learners rather than hard classification decisions, and is better suited to the framework in which histograms are

viewed as weak learners[5]. When tested on our data, we also found that Realboost converges much faster than Adaboost.

The procedure for feature selection using Realboost and the histogram-based classifier as a weak learner are listed in Figure 2.2. Details of weight and coefficient updating rules is beyond the scope of this paper and can be found in [4].

Given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, where $y_i \in \{-1, 1\}$
Initialize sample weights $W_1(\mathbf{x}_i) = 1/N$, T.
For t=1 . . . T:

- For any ratio feature $r(\lambda_i, \lambda_j)$,
  - Build ratio histograms $h^+(r)$ and $h^-(r)$ for positive and negative samples weighted by W_t
  - Compute KLD via Eq. 3
- Select the ratio feature $r^{(t)}$ with maximum KLD
- Construct the weak learner $h_t(r) = \frac{1}{2} \log \frac{h^+(r)}{h^-(r)}$
- Choose α_t by minimizing $Z_t = \sum_{i=1}^{N} D_t(\mathbf{x}_i) \exp(-\alpha_t y_i h_t(r_i^{(t)}))$
- Build the strong classifier from α_r and $h_r.(r = 1...t)$ $H(\mathbf{x}) = sign\left(\sum_{r=1}^{t} \alpha_r h_r(\mathbf{x})\right)$ Stop if training error reaches zero
- Update the sample weights: $W_{t+1}(\mathbf{x}_i) = \frac{W_t(\mathbf{x}_i) \exp\left(-\alpha_t y_i h_t(i)\right)}{Z_t}$

Output features: $r^{(1)}$,$r^{(2)}$,. . . ,$r^{(t)}$($r^{(T)}$)

Figure 2. Boosted ratio feature selection

Realboost was originally proposed as a classification algorithm. Here, we only use it for feature selection since boosting algorithms are much likely to overfit the training data and, as a result, they do not generalize well. Furthermore, the use of a soft margin classifier still cannot guarantee good generalization performance. This led us to a different, yet simple and effective approach, by only using Realboost for feature selection and then using a SVM to classify the selected ratio features. It is also worth noting that the generalization behavior of the SVM is better understood in theory and tested in practice.

2.3. Pairwise Classification Framework

The above procedure is only applicable to two-class cases. However, any multiclass categorization problem can be converted to a number of binary classification problems. In this paper, we adopt a pairwise classification framework for the conversion, which is also called 'one-against-one' classification. A binary classifier is built for any two classes and the final classification result is obtained by voting on the results of all binary classifiers. The procedure is:

Given $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, where $y_i \in \{1, \ldots, K\}$

- For i,$j = 1 \ldots K$, $i \neq j$
  - Obtain positive samples from class i and negative samples from class j
  - Run boosted ratio feature selection algorithm
  - Train a SVM on the selected ratio features for training samples
  - Predict labels for $\mathbf{x}_i$, i=1. . . N
- Combine the pairwise classification results by majority voting

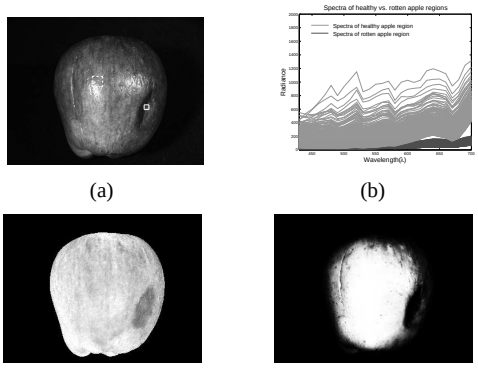
---

3. Experimental Results

In this section, we illustrate the utility of our method for purposes of ratio feature selection.

The first of our experiments is a binary classification example based on a single optimal ratio. An image of diseased apple was captured against a black background using an OKSI hyperspectral camera. The image contains 28 bands sampled at 10nm steps over the visible range (430nm-700nm). We manually selected one diseased area and one healthy region of the apple for training. The pseudocolor image is shown in Figure 3(a), superimposed by regions where pixels were selected for training. To increase the difficulty, some specular pixels were also included in the training region. The large variation of healthy apple spectra in the training region is shown in Figure 3(b). Figure 3(c) shows the surface map inferred by our algorithm, where healthy and diseased areas are mapped onto the apple surface, with brighter pixels indicating healthy regions and darker pixels the diseased tissue. A single band ratio has been automatically chosen here, which is between 670nm, the absorption band, and 700nm, the peak value. It is quite impressive that the selected ratio feature is quite insensitive to changes in incident radiance due to the variation of surface normal. For comparison, we also applied the linear SVM classification directly on the training spectra and obtained the surface map shown in Figure 3(d). From the figure, we can conclude that the mapping result is extremely sensitive to the shading effects.

A more complex multiclass example is shown in the second of our experiments. An image captured by the AVIRIS sensor with 220 bands over the visible and near infrared range was used.The image, available at http://dynamo.ecn.purdue.edu/ biehl/MultiSpec/, maintained by Prof. Landgrebe and his group, covers an agricultural area at NW Indiana. A total number of 10366 pixels were labeled over 16 different terrain types. The ground truth map was shown in Figure 4(b), where each terrain is depicted in a distinct color and the unlabelled pixels were left blank. We performed classification on all 16 classes using 191 bands for each pixel spectrum with water absorption and noisy bands removed. For each class,



(a) (b) (c) (d)

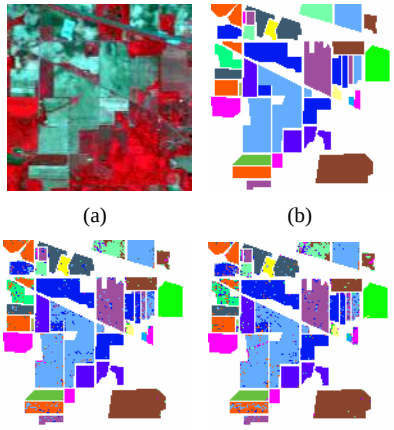
Figure 3. (a) Image of apple superimposed with training regions. (b) Spectra of training samples. (c) Mapping result of our method. (d) Mapping result of SVM.

25% of the labelled pixels were randomly selected for training and the remaining 75% pixels for testing. Four algorithms were compared, including Linear Discriminant Analysis (LDA), our ratio feature selection algorithm followed by SVM (RS+SVM), the same algorithm applied to the spectral bands (BS+SVM) and SVM running directly on raw spectra. To ensure a fair comparison, all the algorithms were properly regularized with parameters chosen by cross validation. Also, the same pairwise classification framework, as described in Section 2.3, was used. We employed the LIBSVM package for training SVMs, which can be found at [6]. The test was repeated five times using different random training samples. The comparison results are listed in Table 3. Here, we have included the mean test error and variance for the five runs, average number of features used in each binary classification and average time spent for prediction on an Pentium 4, 2GHz PC with 512MB of RAM.

From Table 3, its clear that running the SVM directly on all bands achieved best accuracy. Our ratio selection algorithm finished second best. This is a quite surprising result, as the band ratio is generally believed to be a naive measure that can only be used for coarse detection. However, here we show that the correct combination of ratios can perform much more complicated classification than otherwise expected. The overall performance of the RS+SVM framework is quite promising. This is even more important since it achieves a comparable result to that obtained by running the SVM directly, nonetheless it uses much fewer features. It is also less computationally intensive. Thus, it

Table 1. Performance evaluation of competing algorithms on land cover classification

Method	Avg Error	Var	#Features	Avg Time
LDA	19.70%	0.68%	191	2.87s
BS+SVM	14.82%	0.28%	4.35	9.16s
RS+SVM	10.97%	0.22%	3.64	9.31s
SVM	8.47%	0.22%	191	55.90s



(a) (b) (c) (d)

Figure 4. (a) Pseudo-color image of the scene. (b) Ground truth map. (c) Map of land cover inferred by SVM. (d) Map of land cover inferred by our ratio selection algorithm.

offers an ideal trade-off between accuracy and efficiency. Moreover, considering the ratios are much more invariant to reflectance factors, they can be very useful when photometric and geometric effects are dominant.

4. Conclusions

Band ratios have been used for many years in the remote sensing community to identify terrain cover types. In this paper, we have shown the potential of spectral band ratio features for accurate pixel classification and noted its photometric invariant properties. We proposed a principled algorithm for the automatic selection of ratio features. Most importantly, these band ratios are easily linked to domain knowledge. Our future work will focus on terrestrial imaging spectroscopy and investigate the use of the band ratio invariance properties under different photometric conditions.

5. Acknowledgement

The authors are indebted to Dr. Lei Wang for his advice and suggestions on the material presented in this paper.

References

[1] D. Landgrebe, "Hyperspectral Image Data Analysis," *IEEE Signal Process. Mag.*, vol. 19, pp. 17-28, 2002

[2] T. Lillesand, W. Ralph, *Remote Sensing and Image Interpretation*, 4th ed., John Wiley and Sons, 2000

[3] Y. Freund, R. Schapire, "Experiments with a New Boosting Algorithm," *Intl. Conf. on Machine Learning*, 1996

[4] R. Schapire, Y. Singer, "Improved Boosting Algorithm Using Confidence-rated Predictions," *Machine Learning*, vol. 37, no. 3, pp. 297-336, 1999

[5] C. Liu, H. Shum, "Kullback-Leibler Boosting," *Intl. Conf. on Computer Vision and Pattern Recognition*, 2003

[6] C. Chang, C. Lin, "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001