On the Relationship Between Class Definition Precision And Classification Accuracy in Hyperspectral Analysis

David Landgrebe
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-1285
voice: 765-494-3486 fax: 765-494-3358 email:landgreb@ecn.purdue.edu

ABSTRACT

Research in recent years into methods for hyperspectral image data analysis has shown that there is a strong relationship between the precision with which classes are defined and the classification accuracy that results. There is also a relationship between these two factors and the complexity of the classifier algorithm used in the analysis. In this paper we illustrate this relationship empirically using a moderate dimensional, moderately difficult classification task. This example is also used to explore the effect of two recently introduced algorithms that are intended to mitigate the effect of use of a limited number of training samples on classifier performance. The results tend to confirm the theory with regard to training sample size vs. classifier complexity. They also show the two algorithms to be moderately useful in improving classifier performance when training data is limited.

INTRODUCTION

It is well known from both theory and practice that classes of land surface cover are usually not adequately represented by a single spectral curve. Materials of practical interest, such as agricultural crops, forest plantations, natural vegetation, soils, minerals, and objects of interest in urban areas, exist in a number of states and are observed in a number of conditions of illumination. It is thus necessary to characterize them not with a single average or typical spectral response, but with a family of responses. Indeed, the characteristics of this family of responses, i.e. how the spectral responses vary about their average value, may be just as diagnostic of the material as that of their average value.

Quantitatively, this fact is apparent from the Bhattacharyya distance, one of the primary quantitative measures of the separation between two classes. Bhattacharyya distance in parametric form is given by

$$B = \frac{1}{8}\left[\mu_1 - \mu_2\right]^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} \left[\mu_1 - \mu_2\right] + \frac{1}{2} \mathrm{Ln} \frac{\left|\frac{1}{2}\left[\Sigma_1 + \Sigma_2\right]\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}$$

where $\boldsymbol{\mu}_i$ is the mean vector of class i and Σ_i is the covariance matrix of the class. It is seen that the first term on the right quantifies that portion of the distance between the two classes due primarily to the difference in mean or average values. The last term on the right, quantifies the component of separation due to the covariance, or how the class varies about its mean. In any given case, either of these two terms can dominate.

By extension of this concept, it might seem desirable to quantify higher order statistics of the class than just these two, referred to as first and second order statistics. Indeed, it is well established that, in theory, a complete description of an arbitrary distribution can be made by the use of statistics of all orders, as in an infinite series. The reason it is customary to use only the first two orders of statistics, the mean vector and the covariance matrix, arises from the practical problem of estimating these two statistics from the data set to be analyzed. There will necessarily be only a finite number of labeled samples available for each class by which to estimate the statistics of the class. Frequently the number available is not only finite but also small, since the labeling of such samples is one of the most onerous and time-consuming aspects of designing a classifier. Further, due to the highly variable and dynamic nature of Earth surface cover, this labeling must be redone for every data set to be analyzed. Indeed, there often is not much information available about the scene to use in labeling the design samples. Thus, since higher order statistics require increasingly larger numbers of samples to arrive at an adequately precise estimate, it would ordinarily not be practical to use statistics beyond the second order.

The problem of adequacy of the number of design samples becomes even more important as the number of the spectral bands becomes large, since clearly it requires more samples to obtain reasonably precise estimates of high dimensional statistics. Indeed, this results in the fact that, if design data are very limited in the case of hyperspectral data, a simpler classification algorithm which does not make use of second order statistics can on occasion provide more accurate results than a more complete one.

This strong, often dominate dependence on the size of the design set has led over the last few years to seeking algorithms which mitigate this dependence as much as possible. Thus in the following we describe a test on this dependence, varying both the design set size and the classifier complexity. It is also of interest to see how robust the algorithms are in the face of various users.

A PERFORMANCE TEST OF ALGORITHMS AND CLASSIFIERS

The data used for the test is a 12 band data set over an agricultural area consisting of 949 scan lines and 220 samples per line. There were 12 analysis teams, each instructed to design a classifier for the following classes: *corn, oats, soybeans, wheat, and forage/hay.* Though there was extensive ground truth available for the area, the analysts, who were all analyzing multispectral data for the first time, were limited to using about 1000 total design samples, regardless of the number of subclasses they chose to define.

The classification algorithms to be used are the

- *Minimum Distance classifier*, which uses only the individual class mean vectors, the
- *Fisher Linear Discriminant,* which uses the individual class mean vectors plus a covariance matrix common to all classes, the
- *Quadratic Classifier,* which uses individual class mean vectors and individual class covariance matrices, and
- *ECHO*[1,2], which, in addition to class mean and covariance matrices, uses spatial information.

In addition to the classification algorithms, two algorithms that tend to mitigate the limited design set size problem were used. These were the LOOC covariance estimator[3] and a Statistic Enhancement[4] scheme. The LOOC estimator examines the sample covariance estimates, the common covariance estimate, as well as their diagonal forms, and their mixtures to determine which would be most effective. Though a covariance matrix estimate would ordinarily be singular if fewer than n+1 samples are used, where n is the number of spectral bands, LOOC provides a usable covariance estimate with as few as 3 samples, regardless of the number of bands. The Statistics Enhancement scheme uses a sampling of unlabeled samples in addition to the labeled design samples to mitigate the limited design set problem as well as improving the classifier's ability to generalize over the entire data set.

The procedure used by each analyst team was to define an appropriate set of classes including any subclasses they felt necessary, but using a total of 1000 samples or less. This will be called the Baseline Training set. Then define a new training set by selecting a single pixel from each training field used for the baseline set. Next enlarge that set to a 2x2, a 4x4, and an 8x8 pixel area in each Baseline Training field. Classify the data with each algorithm and determine the accuracy based upon a specific test set of 70,588 pixels (about one third of the total pixels in the flightline) provided to each analyst. Next, re-compute the training statistics using the LOOC algorithm for each training set and classify with each algorithm again. Finally, apply the Statistics Enhancement algorithm and again classify with each algorithm. The results are presented in the table below. The table gives the average and standard deviation of the test sample performance over the 12 analysts.

Baseline Results. The baseline results for all three methods of training show a steadily increasing accuracy with increasing classifier complexity as one might expect with a reasonably adequate sized training set. The baseline results also show a steadily declining standard deviation with classifier complexity, indicating a desirable degree of robustness relative to analyst variability. It is seen that, for the MD classifier, LOOC provides no change in accuracy over the standard means for estimating class statistics, as should be the case since the effect of LOOC is only on the class covariance, which the MD classifier does not use.

Dropping down to the **1 pixel/field** case, one sees that a significant price is paid for attempting to use the small design set. Note also that, for the standard training case, the price is the greatest for the more complex classification algorithms. Indeed, there is no loss in accuracy at all for the MD classifier, indicating that the problem is with the precision of the estimate of the class covariance matrices. The LOOC procedure provides substantial improvement in this case, and the use of the Statistics Enhancement procedure provides some additional improvement.

As the size of the training set is enlarged, the results generally tend to improve, as seen by moving down the table from the 1 pixel/field case. There tends to also be improvement moving to the right, meaning using more complete (or complex) algorithms. The accuracy improvement with training set size is small for the MD classifier, indicating the estimation precision of the mean vector is not much of a factor in this case. The improvement is more significant for those algorithms utilizing second order statistics.

Generally the best performance occurs in the case of ECHO, which incorporates spatial information in addition to first and second order statistics. The LOOC procedure also continues to be of some value although the marginal value of it decreases as the accuracy approaches its maximum. However, it may be seen as adding a robustness to the process in the following sense. In a practical circumstance, the analyst usually has no way of knowing "how many training samples are enough," and indeed, the number that can be labeled may well be determined by other factors. LOOC tends to mitigate the cost of having too small a number, tending to keep performance nearer to the maximum possible.

The Statistics Enhancement scheme tends to have the same effect, though it is less apparent from this table. Statistics Enhancement acts to improve the generalization abilities of the classifier by adjusting both the mean vector and the covariance matrix so that the composite of class statistics better fit the entire data set rather than just the subset of training data.

And finally, the standard deviation of the various schemes might be interpreted as quantifying just how much variation in results can be expected due to the analyst him/herself. Low values of standard deviation would suggest that a procedure is relatively immune to variations in the technique used by the analyst to quantitatively define the classes that are desired by the user.

Classifier→	Minimum	Fisher Lin.	Quadratic	ECHO
Training	Distance	Discrim	Max Likeli	
Baseline	100-200 pixels/class			
Std. Ave	75.1%	86.9%	91.4%	92.8%
St. Dev.	9.1%	4.0%	1.9%	1.8%
LOOC Ave	75.1%	87.0%	92.0%	93.9%
St. Dev.	9.1%	4.0%	2.3%	2.3%
LOOC-Enh.Ave	74.4%	86.9%	91.5%	94.6%
St. Dev.	7.6%	2.9%	2.5%	2.3%
1 Pixel/field	1-4 pixels/class			
Std. Ave	75.4%	69.8%	76.4%	78.9%
St. Dev.	7.7%	13.2%	*	*
LOOC Ave	75.6%	81.9%	79.9%	84.0%
St. Dev.	8.0%	6.1%	6.1%	8.5%
LOOC-Enh.Ave	69.3%	81.7%	83.6%	87.2%
St. Dev.	10.9%	4.5%	5.5%	6.1%
4 Pixel/field	4-16 pixels/class			
Std. Ave	71.7%	83.6%	75.9%	76.0%
St. Dev.	18.0%	7.7%	6.6%	6.4%
LOOC Ave	71.6%	84.2%	84.0%	85.6%
St. Dev.	18.0%	7.8%	7.7%	7.3%
LOOC-Enh.Ave	73.5%	83.1%	86.3%	88.0%
St. Dev.	8.8%	9.0%	8.8%	10.1%
16 Pixel/field	16-64 pixels/class			
Std. Ave	75.7%	87.2%	88.6%	89.7%
St. Dev.	8.6%	3.5%	4.2%	4.7%
LOOC Ave	75.7%	87.2%	91.7%	93.6%
St. Dev.	8.6%	3.8%	2.2%	2.4%
LOOC-Enh.Ave	73.3%	85.8%	91.0%	94.2%
St. Dev.	7.2%	4.2%	2.6%	2.1%
64 Pixel/field	64-256 pixels/class			
Std. Ave	76.5%	88.3%	92.8%	94.6%
St. Dev.	7.7%	3.2%	1.1%	1.0%
LOOC Ave	76.5%	88.3%	93.1%	95.0%
St. Dev.	7.7%	3.2%	1.0%	0.9%
LOOC-Enh.Ave	72.7%	84.9%	90.2%	93.2%
St. Dev.	6.4%	4.7%	4.5%	5.3%

All of the calculations by the 12 analyst teams were done using MultiSpec, a software system for Macintosh and Windows desktop computers. MultiSpec is available to anyone without cost at

http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/.

The data set used in these tests is available at that site as well, as a part of an example labeled "Multispectral Data Analysis: A Moderate Dimension Example." Documentation for this example contains the ground truth information needed to define training sets and test the results, so that it is possible for anyone to repeat the classifications reported above. It is noted that these data have not been calibrated or adjusted for the atmosphere or any other observational variable, and none is required. It is likely that no improvement in results would accrue by doing so. The intention is to design and put forth analysis procedures that are not complex and are inexpensive to use.

The overall objective of this line of research is to advance the data analysis technology for multispectral and hyperspectral data to the point that the user can give primary attention to the use of results that remote sensing can provide, rather than necessarily focusing on the technique needed to produce them. As the field matures and more complex data becomes more available, it must not be necessary for one to be thoroughly grounded in the fundamentals of signal processing engineering to obtain the best results possible, any more than it is necessary to understand all of the intricacies of the internal combustion engine in order to drive a car.

Acknowledgement. Work reported in this paper was funded in part by the US Army Research Office under grant DAAH04-96-10444. This sponsorship is gratefully acknowledged.

[1] R. L. Kettig and D. A. Landgrebe, "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," *IEEE Transactions on Geoscience Electronics*, Volume GE-14, No. 1, pp. 19-26, January 1976.

[2] David Landgrebe, "The Development of a Spectral-Spatial Classifier for Earth Observational Data," *Pattern Recognition*, Vol. 12, No. 3, pp. 165-175,1980.

[3] Joseph P. Hoffbeck and David A. Landgrebe, “Covariance Matrix Estimation and Classification with Limited Training Data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, no. 7, pp. 763-767, July 1996.

[4] Behzad M. Shahshahani and David A. Landgrebe, “The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon,” *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, No. 5, pp. 1087-1095, September 1994.