Quantifying multivariate classification performance – the problem of overfitting

Brian R. Stallard and John G. Taylor

Sandia National Laboratories, Albuquerque, NM 87185-0980

RECEIVED

AUG 18 1999

OSTI

ABSTRACT

We have been studying the use of spectral imagery to locate targets in spectrally interfering backgrounds. In making performance estimates for various sensors it has become evident that some calculations are unreliable because of overfitting. Hence, we began a thorough study of the problem of overfitting in multivariate classification. In this paper we present some model based results describing the problem. From the model we know the ideal covariance matrix, the ideal discriminant vector, and the ideal classification performance. We then investigate how experimental conditions such as noise, number of bands, and number of samples cause discrepancies from the ideal results. We also suggest ways to discover and alleviate overfitting.

KEYWORDS: imaging spectrometry, multivariate classification, overfitting

1. INTRODUCTION

The multispectral thermal imager (MTI) satellite, sponsored by the U. S. Department of Energy will be launched in the first quarter of the year 2000. Among its 15 spectral bands are 7 that see to the ground and fall in the reflectance spectral region of 0.5 to 2.5 micrometers. Using these bands, the authors are studying the general capability of the satellite to find targets in spectrally similar backgrounds. It has also been of interest to compare MTI's predicted performance to hyperspectral instruments, although specific performance will not be discussed here. Rather, the present paper is devoted to the side issue of the effect of overfitting on predicted classification performance.

In two dimensions, figure 1 illustrates the concept of overfitting. The experimenter constructs a predictive model from a few observations on known samples. The model is then applied to unknown samples in order to predict a quantity of interest, the y axis, from an observation, the x axis. In viewing figure 1, a technically trained person recognizes that the model using two degrees of freedom is probably appropriate, while the model using five degrees of freedom is overfit. By overfit we mean that excess degrees of freedom in the model drive the error closer to zero than is statistically justified.

In order to insure that our absolute and relative predictions of MTI performance are reliable, we began a systematic investigation of the problem of overfitting in multivariate classification problems. This is an old topic, but we believe that the present work is valuable in that it elucidates some often poorly understood notions and couches them in terms that are directly applicable to remote spectral imagery. The focus of this paper is a data modeling study where the ideal classification performance is known and can be compared to classification performance when overfitting is present. We investigate how experimental conditions such as noise, number of bands, and number of samples may cause discrepancies from the ideal performance. In the future, we expect the computer model to be an important tool in comparing various methods for dealing with overfitting.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

2. DESCRIPTION OF THE MODEL

2.1. Preliminary Comments

The objective in multivariate classification, as we employ it, is to reliably group each pixel in a spectral image as belonging to either of two classes, the target or the background.¹ The first step is to determine a discriminant vector (discussed below) having the same dimension as the spectra. The next step is to calculate the pixel score as the dot product of the discriminant vector with the spectrum at each pixel. This reduces the image to a single spectral dimension. The final step is to choose a threshold score and assign all of the pixels one of two colors depending on whether their scores lie above or below the threshold. If the classification is successful, the resulting image shows the target in high contrast.

In order to study the statistics of the classification process one may ignore the spatial relationship of the pixels and chart the pixel scores as a histogram (see figure 2). Clearly, the more separate the classes are in the histogram, the more reliable will be the classification. For convenience in comparing classification performance we define a unitless quantity, the separation metric (SM), which is simply the difference between the class means divided by the width of the class distributions (i.e., sigma in the two identical Gaussian distributions assumed herein).

A more universal gauge of classification performance is known as the receiver operating characteristic (ROC) curve. Figure 2 shows a typical ROC curve where the x-axis is the fraction of background pixels misclassified as target (false alarm rate) and the y axis is the fraction of target pixels correctly classified (success rate). Each point on the ROC curve is determined with a different threshold. When the area below and to the right of the ROC curve is large, there is good discrimination of the target from the background. Because of the parameterized class distributions used in this study, a single number, SM, adequately represents the ROC curve.

2.2. Constructing the Ideal Covariance Matrix

The model assumes a two class problem where each class is populated from a single mode multivariate Gaussian distribution described by a common covariance matrix. Having chosen *three* as the number of fundamental dimensions for the problem, the fundamental covariance matrix is completely specified by the variances:

$$\Sigma_f = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}. \quad (1)$$

By fundamental dimension, we mean those dimensions that have non-zero eigenvalues. Of course, the actual experimental dimension of the data is another matter. It is dictated by the instrument acquiring the spectra. To transform Σ_f to the higher dimension of the experimental covariance matrix, Σ_i (where the subscript *i* stands for *ideal*), we must specify a set of eigenvectors or basis vectors. We have chosen as the basis vectors the first three terms of a Taylor series, as illustrated in figure 3. After digitizing the basis vectors to correspond to the desired experimental dimension, the vectors are both orthogonalized and normalized. The ideal covariance matrix is found by a transformation using matrix of basis vectors, B_{fe} :

$$\Sigma_i = B_{fe}^T \Sigma_f B_{fe}, \quad (2)$$

where matrix multiplication is implied and the superscript *T* indicates the transpose. The transformation renders Σ_i non-diagonal. In this manner we construct the ideal covariance matrix for any experimental spectral dimension. At this point Σ_i represents the within-class variance. Experimental noise is added to

the model through the noise covariance matrix, which assumes a single variance, σ_n^2 , for the independent Gaussian noise in each sensor channel:

$$\Sigma_n = \begin{bmatrix} \sigma_n^2 & 0 & \cdots & 0 \\ 0 & \sigma_n^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}. \quad (3)$$

The dimension of Σ_n is the same as the number of experimental spectral bands. Finally, the ideal-noise covariance matrix in the experimental dimension is:

$$\Sigma_{in} = \Sigma_i + \Sigma_n. \quad (4)$$

While eigen analysis will readily give Σ_f from Σ_i , once the noise is added we can no longer get back exactly to Σ_f .

2.3. Constructing the Simulated Covariance Matrix

We can also construct a covariance matrix, called the simulated covariance matrix, from a different starting point. In this case we begin with simulated experimental spectra that are calculated with the same parameters used in section 2.2. Now we can vary the number of spectra used to construct the simulated covariance matrix and see what happens when there are insufficient samples to estimate it well. The simulated covariance matrix is:

$$\Sigma_s = D_e'^T D_e', \quad (5)$$

where D_e is a matrix of simulated experimental spectral data, having one column for each spectral channel and one row for each individual spectrum. The prime signifies that the spectra in D_e have been centered to their class means. To construct the simulated spectral data set we begin in the fundamental dimension, which we have chosen to be three:

$$D_3 = \begin{bmatrix} R(M_{1,1}, \sigma_{1,1}) & R(M_{1,2}, \sigma_{1,2}) & R(M_{1,3}, \sigma_{1,3}) \\ \vdots & \vdots & \vdots \\ R(M_{1,1}, \sigma_{1,1}) & R(M_{1,2}, \sigma_{1,2}) & R(M_{1,3}, \sigma_{1,3}) \\ R(M_{2,1}, \sigma_{2,1}) & R(M_{2,2}, \sigma_{2,2}) & R(M_{2,3}, \sigma_{2,3}) \\ \vdots & \vdots & \vdots \\ R(M_{2,1}, \sigma_{2,1}) & R(M_{2,2}, \sigma_{2,2}) & R(M_{2,3}, \sigma_{2,3}) \end{bmatrix}, \quad (6)$$

where $R(\text{mean}, \text{sigma})$ is a Gaussian randomizing function producing a scalar, M is the mean spectral intensity, and σ is the Gaussian width of the class. The subscripts on M and σ signify first the class membership and second the spectral band in the fundamental dimension. The dots in the data matrix signify a continuation of data up to the number of samples that may be desired. Next the data in the fundamental dimension is transformed into the experimental dimension and noise is added:

$$D_e = D_3 B_{3e} + R'(0, \sigma_n), \quad (7)$$

where $R'(\text{mean}, \text{sigma})$ is a Gaussian randomizing function producing a matrix of values having the dimensions of the data matrix, D_e .

Figure 3 shows the basis vectors, the class mean vectors, and a set of representative spectra from both classes. The spectra in figure 3 are noise free, i.e., σ_n in equation 7 is zero. The parameters chosen for this study are presented in the following table. They are rescaled for each change in experimental dimension. This is required to compensate for the rescaling of the basis vectors due to normalization that occurs with each change in dimension. The idea is to keep the shape of the mean spectra fixed on a 0 to 1 reflectance scale no matter what dimension is chosen.

	Class 1	Class 2
M_1	0.32	0.30
M_2	-0.04	-0.02
M_3	-0.02	0.02
σ_1	0.021	0.021
σ_2	0.014	0.014
σ_3	0.007	0.007

Table 1. The set of modeling parameters used in this study. The M and σ are the mean spectrum and square root of the variance, respectively, in the three dimensional spectral space. Values are appropriately scaled for use when the first element in each basis vector is unity.

2.4. Constructing the Discriminant Vector

We have chosen to use the classical linear discriminant vector (LDV), also known as the Fisher discriminant,² since it is widely applied and provides optimal classification performance under the assumptions of the model. The definition is:

$$\text{LDV} = \Sigma^{-1}(M_2 - M_1) = \Sigma^{-1}\Delta, \quad (8)$$

where the superscript, -1 , indicates a matrix inversion and the quantity in the parenthesis is the difference between the class mean spectra, or Δ . Note that the LDV may be scaled by a multiplicative factor with no effect on SM. The LDV is intended to optimally separate the two classes when all the spectra are projected on it. This concept is shown for the case of two dimensional spectra in figure 4. The point to be made from figure 4 is that Δ is already a fair discriminant vector for the two classes. Σ^{-1} provides a possible improvement to Δ by rotating it to take into account the shape of the data cloud. This fine tuning of Δ is helpful if, and only if, the data clouds are non-spherical and there is sufficient experimental data to adequately estimate Σ^{-1} . We shall see that it is the careless use of poor estimates of Σ^{-1} that causes the overfitting problem.

It is possible to define several versions of LDV. The ideal LDV_i uses Σ_i from equation 2 and the ideal Δ_i calculated from the noise-free class mean spectra (i.e., equations 6 and 7 with all the sigmas set to zero). The ideal-noise LDV_{in} uses Σ_{in} from equation 4 and Δ_i (not Δ_{in} , which we do not calculate). The simulated LDV_s uses Σ_s from equation 5 and the simulated Δ_s calculated from a full set of simulated data.

3. OVERFITTING AS DESCRIBED BY THE MODEL

Figure 5 shows how noise affects LDV_{in} . For low noise it approaches LDV_i and for high noise it approaches Δ_i . The reason for this is clear from equations 3 and 4. As the noise begins to compete with the within-class variance, the resultant covariance matrix becomes nearer to being proportional to the identity matrix. In the limit of large noise the covariance matrix has no affect on the shape of the LDV. The inverse of the identity matrix is, of course, the identity matrix. The point of figure 5 is to emphasize that for many experimental conditions LDV_{in} may be significantly different from LDV_i and the former is the appropriate standard to which we ought to compare LDV_s .

Figure 6 shows LDV_s compared to LDV_{in} for the conditions of 31 spectral bands, $\sigma_n = 0.03$, and data sets of both 100 and 1000 samples. When we specify a sample size it is the sum of two equally sized classes. It is startling that LDV_s appears to be such a poor approximation to the desired LDV_{in} . The larger sample size improves the match,³ but we have found that very close matches of these two vectors requires data sets with 10^6 to 10^9 samples, depending on the number of spectral bands. It is instructive to compare the inverse covariance matrices associated with the LDV of figure 6. Figure 7 shows the inverse of the ideal-noise and simulated covariance matrices. The 3D plots have the relative value of the matrix elements on the z axis and the spectral bands on the x and y axes, where (1,1) is at the lower left. It is clear that the ideal-noise inverse covariance matrix is very poorly approximated. The roughness in the approximated covariance matrices is alone responsible for the rough LDV of figure 6 since Δ_s is relatively smooth.

With actual experimental data we also find rough LDV and Σ , but only in the model study can we compare them to the ideal results. The problems noted in figure 6 and 7 are not the bottom line. We need to see how the SM is affected. In figure 8 we plot the histograms related to figures 6 and 7. We see that SM_{in} (using LDV_{in}) is about 3.0 and that there is no dependence of SM_{in} on the number of samples used. But the simulated SM_s (using LDV_s) is higher than what we know to be correct and increases dramatically with a low number of samples. This is clearly a case of overfitting. Note that, at least for the parameters we have used, the overfitting in SM_s fades away with additional samples well before LDV_s conforms accurately to its correct shape.

We now have a glimpse of how SM_s depends on the number of samples in the data set. To complete the picture we have calculated the surfaces in figure 9 and 10 that show how SM_s/SM_{in} (a measure of overfitting) depends on first, the noise level versus the number of spectral bands for a fixed number of samples, and second, the number of samples versus the number of spectral bands for a fixed noise level. The overfitting trends are clear. Overfitting is worse for a high number of spectral bands and a low number of samples. Also, it is worse for data with high noise levels.

4. HOW TO RECOGNIZE OVERFITTING

In the model exercise, overfitting is easily recognize – if SM_s/SM_{in} is significantly greater than unity we know that overfitting is a problem. With real data, SM_{in} may not be known. Nevertheless, it may be possible to recognize overfitting.

The first step is to compare the vectors, LDV and Δ . Generally, Δ has a smoothness similar to the individual spectra. If this is also true for LDV then there is little chance of overfitting. In addition, the SM values for the LDV and Δ should be compared. If they are about the same there is probably no overfitting. Of course, in that case there is little advantage to adding the complexity of calculating Σ . One should just use Δ as the discriminant vector.

The classic way to discover overfitting is to compare SM as a function of degrees of freedom for a training and test data set. The expected curves are displayed in figure 11. SM for the training data improves monotonically as the degrees of freedom increases. SM for the test data has a maximum at the appropriate choice for the degrees of freedom. If there is no maximum in the test data, then one has to question whether it is sufficiently independent from the training data. Unfortunately, the complete curves of figure 11 can be constructed only for calibration models where the degrees of freedom can be varied. One example is the principal component (PC) dimension reduction method described below, where the number of PCs retained is the degree of freedom. Even if the degrees of freedom cannot be conveniently varied, SM for training and test sample data sets should always be compared at a single point. If the former is more than a bit larger than the latter, the LDV is certainly overfit. On the other hand, if they are about equal one still has to be cautious of overfitting, since it is very common to have training and test data sets that are not truly independent.

Another tactic is to try some of the methods to alleviate overfitting that are discussed below. If a smooth version of LDV (calculated by one of the methods below) decreases SM only marginally, then overfitting is not present.

5. HOW TO ALLEVIATE OVERFITTING

The correct solution to the problem of overfitting is to insure that there are sufficient samples in the data set. We believe that a good rule of thumb may be 100 times more samples than the dimension of the spectra. Even with apparently sufficient data one should check for overfitting using the methods of the last section. If overfitting is a problem it can be attacked at several points: 1) the individual spectra, 2) the discriminant vector, or 3) the covariance matrix.

A naive approach to overfitting may be to smooth the spectral data. This, however, proves disastrous. The roughness in Σ^{-1} is caused by correlation among the spectral bands, thereby producing an erratic matrix inversion. The smoothed spectra have an even higher correlation among the bands and produce an even rougher Σ^{-1} . If smoothing the spectra worsens the situation, perhaps adding noise will help. But that just leads to Δ , as seen in figure 5, so is not of much interest.

It is evident from figure 6 that the correct LDV_{in} has the appearance of a smoothed version of the simulated LDV_s. Even though there is no mathematical justification to smooth the LDV, our opinion is that this approach may work well for remote spectral imaging where the spectra are relatively nondescript. Nevertheless, we are not comfortable recommending this approach until we have had the chance to do further studies. In particular we wonder if the approach would work well when the spectra are more complex than those in the present model.

Since the covariance matrix is at the heart of the overfitting problem,⁴ this may be the most fruitful point to attack. We can think of three paths to follow: 1) increase the number of samples through simulation, 2) reduce the dimension of the problem, and 3) modify the covariance matrix (or its inverse) directly.

If there are sufficient samples, the overfitting problem is solved. So why not simulate more samples. However, the obvious way to simulate more spectra is to build them from the actual samples by introducing a randomness at each wavelength. Once again we have the path described by equation 3 and figure 5. This holds little promise.

We recognize that, for a fixed number of samples, the susceptibility to overfitting is decreased when the dimension of the problem is decreased. A trivial way to reduce the dimension is to deresolve the spectra by binning neighboring bands. We have seen in figure 10 that this can indeed reduce the overfitting. But it is possible that statistically valid information may be lost in deresolving the spectra. A more interesting idea is to reduce the dimension of the spectra by replacing the spectra with their scores on a few of the principal components. This can avoid overfitting, but once again there is no assurance that information crucial to class discrimination is retained after the truncation. It is possible that the best classification may be had with PCs 2, 3, and 7; rather than 1, 2, and 3. This suggests a search through the PCs for the few that may be relevant. Other dimension reduction methods have been suggested such as projection pursuit and canonical discriminant analysis. Our effort in this area is ongoing.

We are very interested in modifications that may be made directly to the covariance matrix itself. Figure 7 suggest that a smooth version of the inverse covariance matrix may approximate the ideal matrix. Also, we have mentioned that adding a scaled identity matrix to the experimental matrix (equivalent to adding noise to the spectra) has a smoothing effect on the LDV. But there are more interesting and effective smoothing methods. The method of Hastie⁵ and others known as penalized discriminant analysis involves

regularization of the covariance matrix which has the effect of smoothing it. Again, we intend to use the model to test the efficacy of this approach in future work.

6. CONCLUSIONS

This paper illustrates the basic issues regarding overfitting in multivariate classification. The model we employ gives the ideal Σ , LDV, and SM, and allows us to study the affect of experimental choices on these fundamental statistical quantities. Also, we use the ratio, SM_s/SM_{in} , to directly quantify overfitting. As expected, we see that overfitting is a problem when too few samples are collected. In addition, overfitting is worse when the data has high spectral dimension and high noise levels. The effect of noise level is a new finding, as far as we know. However, we have not yet proven it to be a general result. In sections 4 and 5 we present several methods to recognize and alleviate overfitting. In future work, the modeling approach of this paper will be used to compare and contrast these methods. Once a few preferred methods have been identified we plan to apply them to actual data.

In combing the literature for examples of multivariate classification applied to remotely sensed spectral imagery, we see an alarming rate of possible overfitting. Particularly sensitive to overfitting, are comparisons of sensor performance, where the number of spectral bands changes, and comparisons of algorithm performance, where the number of degrees of freedom changes. Overfitting leads to erroneous conclusions in these types of studies. It also leads to an optimistic projection of performance in general. For instance, suppose that the correct SM is 3.0 and the overfit prediction of SM is 6.0. For a false alarm rate of 0.001, the correct success rate is only 0.58, while the overfit projection is 1.00. This is a huge difference to most practical end users. To avoid unduly optimistic performance projections, one has to work in a dimension that is supported by the size of the training data set and/or one has to be sure of the independence of the test data set. If this is not possible we recommend reporting the classification performance (at least for comparison purposes) when Δ is used as the discriminant vector. Also, we recommend that the LDV or its equivalent be graphically displayed in published work. If LDV is much more ragged than a typical spectrum the author should be careful to consider the topic of overfitting.

ACKNOWLEDGEMENTS

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract number DE-ACO4-94AL85000. Summer students Brian Post and Kasey Hutt contributed to this work.

REFERENCES

1. P. E. Green and J. D. Carroll, *Mathematical Tools For Applied Multivariate Analysis*, Academic Press, New York, 1976.
2. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, 1990.
3. K. Fukunaga and R. R. Hayes, "Effects of Sample Size in Classifier Design," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, **11**, pp. 873-885, 1989.
4. J. P. Hoffbeck and D. A. Landgrebe, "Covariance Matrix Estimation and Classification with Limited Training Data," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, **32**, pp. 763-767, 1996.
5. T. Hastie, A. Buja, and R. Tibshirani, "Penalized Discriminant Analysis," *Annals of Statistics*, pp. 1-31, 1995.

Figure 1. An example of overfitting using fabricated data. The axes are arbitrary. The circles are the data, the straight line is a fit with two degrees of freedom, and the polynomial is a fit with five degrees of freedom.

Figure 2. a) Two Gaussian distributions illustrating the definition of the separation metric (SM). b) ROC curve (explained in the text) showing its relationship to SM.

Figure 3. a) Basis vectors. b) Mean vectors. c) Spectra illustrating the within-class variance of class 1. d) Spectra illustrating the within-class variance of class 2.

Figure 4. An illustration of separating two classes. Data clouds in two spectral dimensions, represented by the ellipses, can be projected onto either the difference in class means, Δ , or the linear discriminant vector, LDV. Considering also equation 8, it is clear that the LDV uses the shape information in the covariance matrix to improve the separation.

Figure 5. Comparison of a) ideal LDV_i (no noise), b) ideal-noise LDV_{in} , and c) difference between the class means, Δ_i . We see that noise causes LDV_{in} to become more like Δ_i . Noise level, σ_n , is 0.1.

Figure 6. Comparison of the simulated LDV_s (rough) and the ideal-noise LDV_{in} (smooth) for a) 100 samples and b) 1000 samples. The noise fraction is 0.03 and the number of spectral bands is 31.

Figure 7. Comparing three versions of the inverse covariance matrix that pertain to figure 6. a) ideal-noise, b) simulated with 100 samples, and c) simulated with 1000 samples.

Figure 8. Comparison of the histograms pertaining to figure 6 and 7. When the ideal-noise LDV_{in} is used, SM does not depend on the number of samples. When the simulated LDV_s is used, SM decreases with large numbers of samples and approaches the true value.

Figure 9. Overfitting as a function of noise fraction and number of spectral bands. The number of samples is fixed at 400.

Figure 10. Overfitting as a function of number of samples and number of spectral bands. The noise is fixed at 0.05.

Figure 11. The dependence of SM on the degrees of freedom in the calibration model. SM for the training data increases monotonically while SM for the test data reaches a maximum.