Covariance Estimation For Limited Training Samples

Saldju Tadjudin
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
Phone (765) 494-9217
Fax (765) 494-3358
tadjudin@ecn.purdue.edu

David A. Landgrebe
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285
Phone (765) 494-3486
Fax (765) 494-3358
landgreb@ecn.purdue.edu

Abstract -- In Gaussian maximum likelihood classification, the mean vector and covariance matrix are usually estimated from training samples. When the training sample size is small compared to dimensionality, the sample estimates, especially the covariance matrix becomes highly variable and consequently, the classifier performs poorly. In particular, if the number of training samples is less than dimensionality, the sample covariance estimate becomes singular so the quadratic classifier cannot be applied. Unfortunately, the problem of limited training samples is prevalent in remote sensing applications. While the recent progress in sensor technology has increased the number of spectral features making possible more classes to be identified, the training data remain expensive and difficult to acquire. In this work, the problem of small training set size on the classification performance is addressed by introducing a covariance estimation method for limited training samples. The proposed approach can be viewed as an intermediate method between linear and quadratic classifiers by selecting an appropriate mixture of covariance matrices. The mixture of covariance matrices is formulated under an empirical Bayesian setting which is advantageous when the training sample size reflects the prior of each class. The experimental results show that the proposed method improves the classification performance when training sample sizes are limited.

INTRODUCTION

In the conventional Gaussian maximum likelihood (ML) classifier, the sample estimates are computed from training samples and are used as the ML estimates of the mean vector and covariance matrix. The quadratic classifier's performance can be degraded when the number of dimensions is large compared to the training set size due to the instability of sample estimates. In particular, the sample covariance estimate becomes highly variable and may even be singular. One way to deal with the instability of covariance estimate is to employ the linear classifier which is obtained by replacing each class covariance estimate with their average. Although a linear classifier often performs better than a quadratic classifier for small training set size, the choice between linear and quadratic classifiers is rather restrictive. Several methods[1][2][3] have been proposed where the sample covariance estimate is replaced by partially pooled covariance matrices of various forms. In this formulation, some degree of regularization is applied to reduce the number of parameters to be estimated, thus improving classification performance with small training set size. Therefore, regularization techniques can also be viewed as choosing an intermediate classifier between the linear and quadratic classifiers. In general, regularization procedures can be divided into two tasks: 1) the choice of covariance mixture models, and 2) model selection. To perform regularization, one must first decide upon a set of appropriate covariance mixture models that represent a "plausible" set of covariance estimates. Normally, a covariance mixture of the following form is assumed:

$$\hat{\Sigma}_i = (1 - w_i) S_i + w_i S_p \qquad 0 \le w_i \le 1$$

The regularization or mixing parameter w_i then controls the biasing of individual class covariance sample estimate S_i to a pooled covariance matrix S_p. However, this partially pooled covariance estimate may not provide enough regularization even for a linear classifier. In the case when the total number of training samples is comparable to or is less than the dimension, even the linear classifier becomes ill- or poorly-posed. Therefore, an alternative covariance mixture is provided by biasing the sample covariance toward some non-singular diagonal matrix Λ:

$$\hat{\Sigma}_i = (1 - w_i) S_i + w_i \Lambda \qquad 0 \le w_i \le 1$$

For given value(s) of the mixing parameter(s), the amount of bias will depend on how closely the estimates $\hat{\Sigma}_i$ actually represent those true parameters Σ_i. Therefore, the goal of model selection is to select appropriate values for the mixing

Work leading to this paper was supported in part by NASA under Grant NAG5-3975 and the Army Research Office under Grant DAAH04-96-1-0444.

parameters which can be estimated from minimizing a loss function based on the training samples. A popular minimization criterion is based on cross-validated estimation of classification error. This criterion has the benefit of being directly related to classification accuracy even though it is computationally intensive. However, the process of estimating each class covariance matrix involves the covariance estimates of all classes, which implies that the same mixing parameter has to be used for all classes. The same choice of mixing parameter might not be optimal for all classes. Furthermore, the same classification error rate might occur along a wide range of parameter values and hence the optimal value of mixing parameter is non-unique. Therefore, a tie-breaking technique is needed. Another maximization criterion which has been applied is the sum of the average leave-one-out likelihood values. This criterion requires less computation than the leave-one-out classification error procedure. It also has the advantage that each class covariance matrix can be estimated independently of the others. Therefore, the mixing parameter can be different for each class. Moreover, not all classes need to be subjected to regularization, especially those with sufficient training samples. However, a major drawback of this criterion is the lack of direct relationship with classification accuracy. In this work, we propose a new covariance estimator based on a Bayesian formulation. The proposed estimator is essentially an extension of previous works in [1][2][3].

PROPOSED COVARIANCE ESTIMATOR

The first form of covariance mixtures is derived by assuming that the total number of training samples is greater than the dimensionality. In this case, the common covariance matrix is non-singular. The assumption of normally distributed samples implies that the sample covariance matrices S_i are mutually independent with

$$S_i \sim W\left(\frac{1}{f_i}\Sigma_i, f_i\right)$$

where $f_i = N_i - 1$, N_i is the number of training samples for class i and W denotes the central Wishart distribution with f_i degrees of freedom and parameter matrix Σ_i. Then the family of inverted Wishart distributions provides a convenient family of prior distributions for the true covariance Σ_i. Assume that each Σ_i has an inverted Wishart prior distribution so that the Σ_i are mutually independent with

$$\Sigma_i \sim W^{-1}\left((t-p-1)\Psi, t\right) \qquad t > p+1$$

where W^{-1} is an inverted Wishart distribution with parameters Ψ and t for p dimensions. Then the prior mean Ψ represents the central location of the prior distribution of the Σ_i, and t controls the concentration of the Σ_i around Ψ. Under squared error loss, the Bayes estimator of Σ_i is given by[2]

$$\hat{\Sigma}_i(\Psi, t) = \frac{f_i}{f_i + t - p - 1} S_i + \frac{t-p-1}{f_i + t - p - 1}\Psi$$

By letting $w_i = \frac{t-p-1}{f_i + t - p - 1}$, and Ψ be a pooled covariance estimate S_p, the Σ_i can then be replaced by partially pooled estimates of the form :

$$\hat{\Sigma}_i = (1 - w_i)S_i + w_i S_p \qquad 0 \le w_i \le 1$$

The value of t can in turn be expressed in terms of w_i :

$$t = \frac{w_i(f_i - p - 1) + p + 1}{1 - w_i} \qquad 0 \le w_i < 1.$$

The pooled covariance estimate is then defined by the generalized least squared estimator of Ψ, designated as $S_p(t)$, for L classes and a given t:

$$S_p(t) = \left(\sum_{i=1}^{L} \frac{f_i}{f_i + t - p - 1}\right)^{-1} \sum_{i=1}^{L} \frac{f_i}{f_i + t - p - 1} S_i$$

When the total number of training samples is close to or less than the number of features, even the pooled covariance matrix becomes unstable. In this case, biasing the sample and common covariance estimates towards some form of diagonal matrix can avoid the problem of singularity. We bias the sample and common covariance estimates towards their own diagonal elements which are advantageous when the class covariance matrix is ellipsoidal. The proposed covariance estimator then has the following form:

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & 0 \le \alpha_i < 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p(t) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & 2 \le \alpha_i \le 3 \end{cases}.$$

where $S = \frac{1}{L}\sum_{i=1}^{L} S_i$. The maximization of leave-one-out average log likelihood is used as the criterion to select the appropriate mixture model. Therefore, to select an appropriate mixture, the value of α_i is fixed and the leave-one-out average likelihood is computed and compared for each α_i. The direct implementation of the leave-one-out likelihood function for each class with N_i training samples would require the computation of N_i matrix inverses and determinants at each value of α_i. Fortunately, a more efficient implementation can be derived using the rank-one down-date of the covariance matrix.

EXPERIMENTAL RESULTS

For the experiment, we use a 145X145 pixel AVIRIS image. The water absorption bands have been discarded, leaving a total of 200 channels. This data contains 17 classes of varying sizes. The purpose of this experiment is to demonstrate the effect of covariance estimation on classes with varying covariance structures and different training sample size. The training samples are selected to be 1% of the number of labeled samples for each class. The labeled samples, excluding the training samples are then used as test samples. The classes, and the numbers of labeled samples are listed in Table 1. This data was obtained in June 1992 so most of the row crops in the agricultural portion of the test site had not reached their maximum ground cover. Therefore, the classification of these crops becomes challenging since the spectral information comes from a mixture of the crops, the soil variations and previous crop residues. These crops are listed as the first seven classes and their mean classification accuracy is computed separately. The classification procedures for testing the data are shown in Table 2. Since the Euclidean distance classifier does not utilize the covariance information, its performance would indicate whether the second order statistics are useful for the classification of high dimensional data with limited training samples. The use of the common covariance estimate for all classes is equivalent to a linear classifier. The leave-one-out covariance estimator[2] (LOOC) is implemented to compare with the proposed Bayesian leave-one-out covariance estimator (bLOOC). The mixing parameter α_i is set at 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75 and 3. In addition to using the covariance estimator to help increase the stability of covariance estimate, the use of feature extraction can also help reduce the number of features to cope with small training set sizes. We perform Discriminant Analysis Feature Extraction[4] (DAFE) which only utilizes mean information and is therefore less sensitive to small training sample size. The sample covariance estimate is not tested in this experiment since the numbers of training samples for some classes are extremely small. Two types of classifiers, namely, the quadratic classifier (QC) and the spatial-spectral classifier ECHO[5] (Extraction and Classification of Homogeneous Objects) are then applied and compared. While the quadratic classifier assign individual pixels to one of the classes, the ECHO classifier first divides the image into groups of contiguous pixels and classifies each group to one of the classes. The results of classification are shown in Table . The highest accuracy is highlighted in bold letters.

DISCUSSION AND CONCLUSION

The performance of the Euclidean distance classifier is significantly lower than the other classifiers. This shows that the second order statistics are useful for classifying high dimensional data even though the training samples are limited. Although the class covariance matrices differ substantially, the use of common covariance matrix and hence the linear classifier improves the performance substantially compared to the Euclidean distance classifier. The table shows that the best performance is achieved by using bLOOC, DAFE and the ECHO classifier. The classification accuracy increases substantially for the row crops 1-7. Compared with the second best result obtained from the classifier LOOC+DAFE+ECHO, the accuracy increases from 82.72% to 89.06%. The mean accuracy for all classes improves from 80.35% to 82.90% as well. It should be mentioned that when all classes have equal number of training samples, bLOOC has the same form as LOOC. Therefore, the proposed Bayesian estimator is beneficial when the sample sizes are unequal and the training set size reflects the true priors. Further details of this method can be found in [6].

REFERENCES

[1] J.F. Friedman, "Regularized discriminant analysis," J. R. Statist. Soc., Vol 84, pp. 17-42, 1989.

[2] J.P. Hoffbeck and D.A. Landgrebe, "Covariance matrix estimation and classification with limited training data", IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol 18, No. 7, pp. 763-767, 1996.

[3] W. Rayens and T. Greene, "Covariance pooling and stabilization for classification," Computational Statistics and Data Analysis, Vol 11 pp. 17-42, 1991.

[4] K. Fukunaga, Introduction to Statistical Pattern Recognition. 2nd Ed., Boston: Academic Press, 1990.

[5] R.L. Kettig and D.A. Landgrebe, "Classification of multispectral image data by extraction and classification of homogeneous objects," IEEE Trans. Geosci. Electro., Vol GE-14, No. 1, pp. 19-26, 1976.

[6] S. Tadjudin, Classification of High Dimensional Data with Limited Training Samples, Ph.D. Thesis, School of Electrical and Computer Engineering, Purdue University, 1998.

Table 1 Class description of the AVIRIS data set

Class Name	No. of Labeled Samples
1. Corn-no till	1423
2. Corn-min till	834
3. Corn	234
4. Soybeans-no till	797
5. Soybeans-no till2	171
6. Soybeans-min till	2468
7. Soybeans-clean till	614
8. Alfalfa	54
9. Grass/Pasture	497
10. Grass/Trees	747
11. Grass/pasture-mowed	26
12. Hay-windrowed	489
13. Oats	20
14. Wheat	212
15. Woods	1294
16. Bldg-Grass-Tree-Drives	380
17. Stone-steel towers	95

Table 2 Classification accuracy for the AVIRIS data

Procedures	Class 1-17 (%)	Class 1-7 (%)
Euclidean Distance	48.23	31.79
Common Cov+DAFE+QC	74.81	70.18
Common Cov+DAFE+ECHO	76.78	74.94
LOOC+DAFE+QC	75.29	70.71
LOOC+DAFE+ECHO	80.35	82.72
bLOOC+DAFE+QC	75.53	72.61
bLOOC+DAFE+ECHO	**82.91**	**89.06**