THE EFFECT OF UNLABELED SAMPLES IN REDUCING THE SMALL SAMPLE SIZE PROBLEM AND MITIGATING THE HUGHES PHENOMENON[1]

Behzad M. Shahshahani[2] and David A. Landgrebe
School of Electrical Engineering
Purdue University
West Lafayette, IN 47906-1285
shahshahani@vnet.ibm.com landgreb@ecn.purdue.edu

[1]This work was supported in part by NASA under Grant NAGW-925.
[2]Dr. Shahshahani is now with IBM at Boca Raton, Florida.

THE EFFECT OF UNLABELED SAMPLES IN REDUCING THE SMALL SAMPLE SIZE PROBLEM AND MITIGATING THE HUGHES PHENOMENON

ABSTRACT

In this paper, we study the use of unlabeled samples in reducing the problem of small training sample size that can severely affect the recognition rate of classifiers when the dimensionality of the multispectral data is high. We show that by using additional unlabeled samples that are available at no extra cost, the performance may be improved, and therefore the Hughes phenomenon can be mitigated. Furthermore, by experiments, we show that by using additional unlabeled samples more representative estimates can be obtained. We also propose a semi-parametric method for incorporating the training (i.e., labeled) and unlabeled samples simultaneously into the parameter estimation process.

I. Introduction

An important problem in pattern recognition is the effect of small training sample size in classification performance. It is well known that when the ratio of the number of training samples to the number of feature measurements is small, the estimates of the discriminant functions are not accurate, and therefore the classification results may not be satisfactory. This problem is becoming increasingly significant in remote sensing, as the number of spectral bands in sensors becomes larger. The new generation of the remote sensing sensors that are proposed for the Earth Observing System (EOS) can produce data in large number of spectral bands. The MODIS sensor produces data in about 50 bands [1], whereas the AVIRIS sensor produces as many as 200 spectral bands [2]. One objective of using such high resolution sensors is to discriminate among more ground cover classes and hence obtain a better understanding about the nature of the materials that cover the surface of the Earth. Details such as differences among various types or conditions of the same species that were not possible to observe using the older generations of

sensors such as Thematic-Mapper of Landsat, should be apparent by using the higher resolution sensors.

To fully use the information contained in the new feature measurements, training samples are needed from all the classes of interest. A large number of classes of interest, and a large number of spectral bands to be used, require a large number of training samples. Such training samples are usually very expensive and time consuming to acquire. For remote sensing applications, ground truth information must be gathered by visual inspection of the scene near the same time that the data is being taken, by using an experienced analyst for identifying the class labels of data based on their spectral responses, or by other means. In any case, usually only a limited number of training samples can be obtained. These training samples are often used for deciding what features in data are useful for discriminating among the classes of interest, and for designing classifiers based on these derived features. Figure 1 illustrates a typical scenario for analyzing remote sensing data.

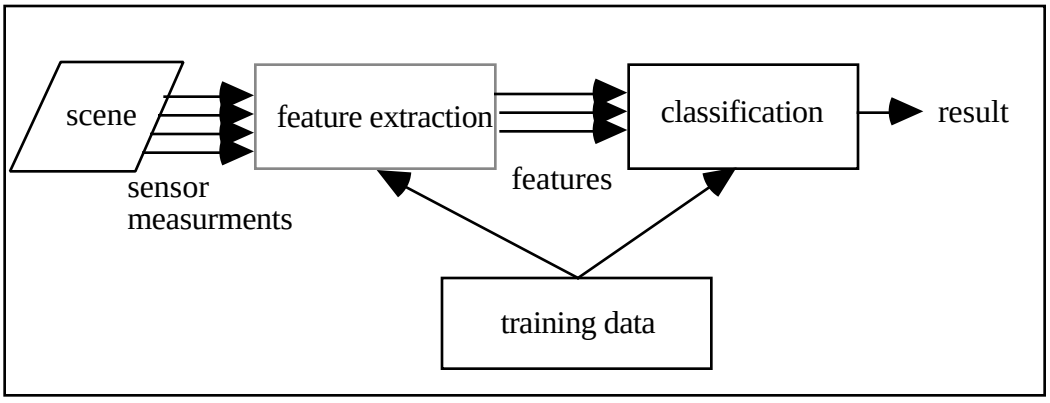

Figure 1: Typical steps in the analysis of remote sensing data

Usually, both the feature extraction and the classification stages of the analysis are based on optimizing a criterion that must be estimated using the training samples. If the number of training samples is small compared to the dimensionality of the data, both of these stages may suffer from

bad estimates. Therefore, the resulting performance of the whole analysis may be less than satisfactory.

An additional problem that usually exists in remote sensing applications is the unrepresentative training samples problem. The training samples that are obtained from spatially adjacent regions may not be good representatives of the samples of the same class that might exist in other regions in the scene. This problem further aggravates the difficulties in analyzing remote sensing data.

The purpose of this work is to study some techniques for reducing the small sample size problems by using unlabeled observations that may be available in large number and with no extra cost. Including the unlabeled data in the process of designing classifiers can have the following potential advantages: 1) The large number of unlabeled samples can enhance the estimates of the parameters and therefore reduce the effect of the small sample size problem. 2) The estimates of the parameters that are obtained by using the training samples may be updated by using additional unlabeled samples to obtain statistics that are more representative of the true sampling distributions. 3) The prior probabilities of the classes that can not be found by training samples alone may be estimated by using unlabeled samples.

The organization of the paper is as follows. In section II, the Hughes phenomenon, which is the loss of classifiability that is observed when the dimensionality of the data increases while the training sample size remains fixed, is briefly discussed. The effect of the classifier type on classification accuracy is discussed in section III. In section IV, the problem of small training sets is discussed with the aid of an experiment. The effect of additional unlabeled samples in improving the performance is studied in section V. Some methods for incorporating the unlabeled samples into the classifier design process are studied in section VI. In section VII, some experimental results are presented. In section VIII, the effect of unlabeled samples in

obtaining more representative estimates is demonstrated by an experiment. Final remarks are presented in section IX.

II. The Hughes Phenomenon

In a typical classification problem, the objective is to assign a class label, from a set of such labels, to an incoming observation. The minimum expected error that can be achieved in performing the classification process is referred to as the Bayes' error. A decision rule that assigns a sample to the class with highest a posteriori probability (the MAP classifier), achieves the Bayes' error [3]. To design such a classifier, knowledge of the a posteriori probabilities and thus, the class conditional probability density functions is required. If such knowledge is available then by increasing the dimensionality of data one would expect to enhance the performance. In other words, the Bayes error is a decreasing function of the dimensionality of the data. After all, a new feature can only add information about a sample and thus, one would expect to do at least as well as if such information was not available.

In practice, however, class conditional probability density functions (pdf's) need to be estimated from a set of training samples. When these estimates are used in place of the true values of the pdf's, the resulting decision rule is sub-optimal and hence has a higher probability of error. The expected value of the probability of error taken over all training sample sets of a particular size is, therefore, larger than the Bayes error. When a new feature is added to the data the Bayes error decreases, but at the same time the bias of the classification error increases. This increase is due to the fact that more parameters need to be estimated from the same number of samples. If the increase in the bias of the classification error is more than the decrease in the Bayes error, then the use of the additional feature degrades the performance of the decision rule. This phenomenon is called the Hughes effect [4]. The larger the number of the parameters that need to be estimated, the more severe the Hughes phenomenon can become. Therefore, when the

dimensionality of data and the complexity of the decision rule increase, the Hughes effect can become more severe.

III. Classification Performance Versus Classifier Type

The functional form of a classifier determines the shape of the decision boundaries that it can produce. Linear classifiers, such as the Minimum Euclidean Distance (MED classifier) classifier, which is optimal when classes are Gaussian with identity covariance matrices, can produce hyper-plane boundaries, whereas quadratic classifiers, such as the Gaussian Maximum Likelihood (GML classifier) classifier, which is optimal when the classes are Gaussian with different covariance matrices, can produce quadratic boundaries. More complex classifiers can create even more complex boundaries. Obviously, the more complex the classifier is, the more powerful it is in terms of its ability to discriminate among various classes of different shapes. In remote sensing, it has been observed that quadratic classifiers that take advantage of the second order statistics of the classes, e.g., GML classifiers, are very powerful for discrimination [5]. The value of the second order statistics is evidently more prominent when the dimensionality of the data is high. In high dimensions it seems that the second order variations of the classes contain more information than the first order variations [5]. To demonstrate this fact, the following experiments were performed (additional similar experiments are reported in [5]):

Experiment 1 (AVIRIS data):

A portion of an AVIRIS frame (consisting of 210 bands) taken over Tippecanoe county in Indiana was used in this experiment. Four ground cover classes were determined by consulting the ground truth map. The classes were bare soil (380 pixels), wheat (513 pixels), soybean (741 pixels), and corn (836 pixels). The average pair-wise Bhattacharyya distance between the classes was computed for every fifth band of the AVIRIS data. The bands were then ranked according to the average Bhattacharyya distance. The dimensionality of the data was incremented from 1 to 18

by sequentially adding more bands, i.e., for dimension 1 the first ranked band was used, for dimension two the first two ranked bands were used and so on. In this way at each dimension all the information in the previous dimensions was present. One hundred training samples were drawn randomly from each class. The statistics of each class (mean vector and covariance matrix) were estimated by the maximum likelihood (ML) estimators. The rest of the samples were classified using the MED and GML classifiers, and the total classification accuracy (the ratio of the number of correctly classified samples to the total number of samples) was computed. Each experiment was repeated ten times independently and the average of the ten trials was obtained. The results are shown in Figure 2.

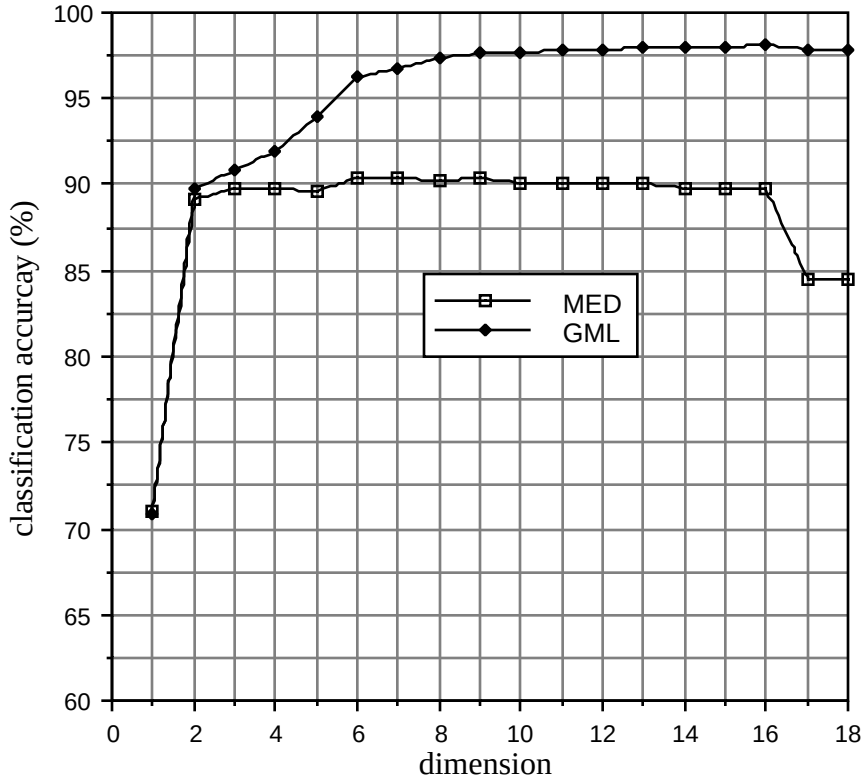

Figure 2: Classification accuracies of the MED and GML classifiers versus dimensionality for the AVIRIS data set based on 100 training samples per class.

From Figure 2 it is seen that the GML classifier, that takes advantage of the second order statistics of the classes and creates quadratic boundaries, is more powerful in discriminating among the classes, especially when the dimensionality of the data increases. However, the number of the parameters in the GML classifier is more than that in the MED classifier. As the dimensionality grows, the number of entries in the covariance matrices of the classes increases rapidly. Therefore, when the dimensionality of the data begins to approach the number of training samples, one would expect the Hughes phenomenon to affect the GML classifier more severely. This point will be discussed in more detail in the next section.

IV. Effect of Small Training Sample Size

Consider a classification problem involving m classes with prior probabilities α_i and probability density functions $f_i(x)$. By e^* we denote the Bayes' error achieved by using the MAP classifier when α_i and $f_i(x)$ are known. Let θ denote the vector of parameters of the MAP classifier. If the pdf's are parametric (such as multivariate Gaussian), θ usually includes the parameters of each class (e.g., mean vectors and covariance matrices) and the associated prior probabilities. On the other hand, if $f_i(x)$ is not considered to be parametric, θ is assumed to contain the value of $f_i(x)$ at each particular sample x under consideration. Let θ^* denote the true value of θ. The error achieved by using θ in the decision rule is e^*, the Bayes error. Now, assume that $\hat{\theta}$ is an estimate of θ^*. If the deviation of $\hat{\theta}$ from θ is not large, one can approximate the error corresponding to the decision rule obtained using $\hat{\theta}$ by using a Taylor series expansion of up to the second term:

$$\hat{e} = e(\hat{\theta}) \approx e^* + \left.\frac{\partial e^T(\theta)}{\partial \theta}\right|_{\theta=\theta^*} (\hat{\theta} - \theta^*) + \frac{1}{2} tr\{\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^*} (\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T\} \qquad (1)$$

where tr(A) denotes the trace of matrix A. The term $\left.\frac{\partial e^T(\theta)}{\partial \theta}\right|_{\theta=\theta^*}$ is zero since θ^* is an extreme point of $e(\theta)$. If the bias of $\hat{\theta}$ is zero or negligible ($E\{\hat{\theta}\} \approx \theta^*$), then the expected value of $\hat{e}$ can be approximated as follows:

$$E\{\hat{e}\} \approx e^{*} + \frac{1}{2}\,\mathrm{tr}\{\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^{*}} \mathrm{cov}(\hat{\theta})\} \qquad (2)$$

Notice that the bias term on the right hand of equation (2) is non-negative, because it is the trace of the product of two positive semi-definite matrices [6]. As the number of the parameters (θ) increases, the number of terms in the bias increases and hence the expected value of the error increases, too. If this increase is not canceled by the decrease in the Bayes' error that the additional parameters may provide, then the Hughes phenomenon occurs. To demonstrate this fact, experiment 1 is repeated here, but instead of 100 training samples, 20 training samples per class are used. The effect of small training sample size is evident in the results shown in Figure 3.

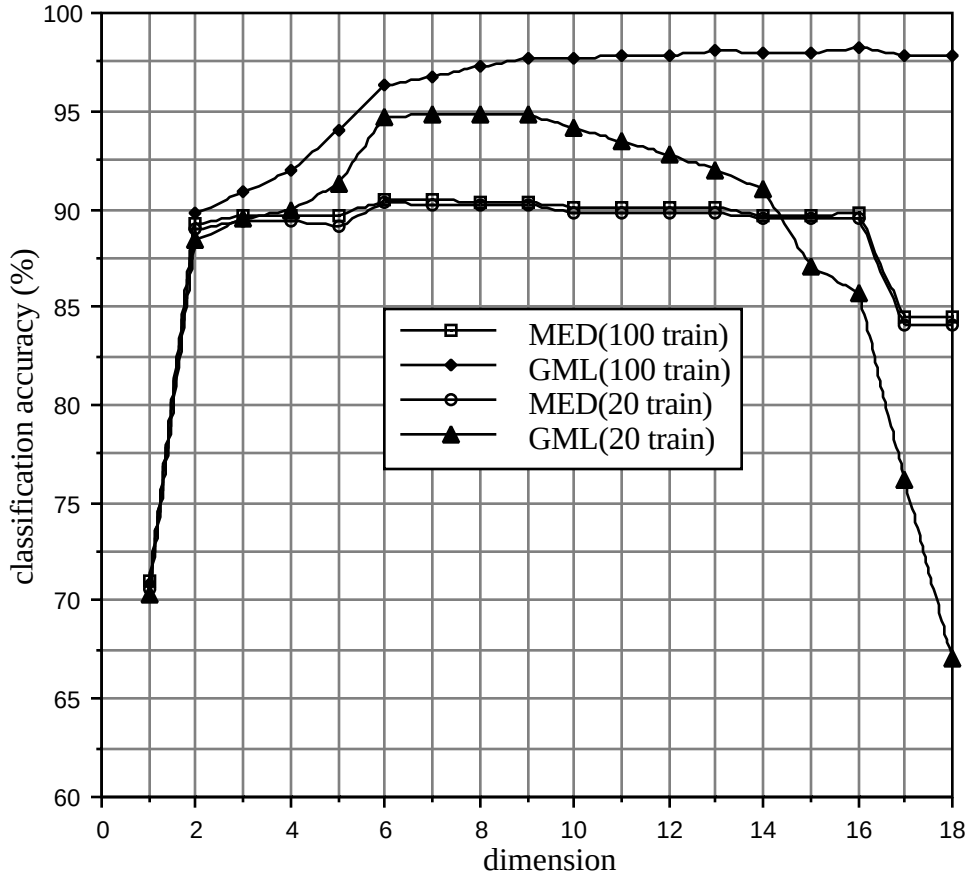

Figure 3: Effect of small sample size in the performance of the MED and GML Classifiers for experiment 1 (AVIRIS data).

From Figure 3, it can be seen that when the number of training samples is small the GML classifier is more severely affected by the Hughes phenomenon than the MED classifier. The behavior of the MED classifier was not significantly changed when the numbers of training samples was reduced but the accuracy of the GML classifier started to decrease after the dimensionality passed beyond a certain point. Therefore, although the second order statistics can be invaluable for discrimination in high dimensional spaces, if not properly estimated they can also significantly reduce the performance. As can be seen from equation 2, what causes the increase in the expected error is the covariance of the estimates of the parameters. Since, the sample average and sample covariance are the minimum variance unbiased estimators for the mean and covariance matrix, it appears that not much improvement can be hoped for if only training samples are used in the estimation process. However, if by using additional information, such as the information contained in unlabeled samples, estimates with lower covariance matrices can be found, then the bias in the classification error may be reduced and therefore the Hughes phenomenon may be mitigated.

V. Effect of Additional Unlabeled Samples

Let's consider the bias term in the right hand side of equation 2. Consider two different estimators, $\tilde{\theta}$ and $\hat{\theta}$, which both have negligible bias, and assume that $\mathrm{cov}(\tilde{\theta}) \leq \mathrm{cov}(\hat{\theta})$ (i.e., $\mathrm{cov}(\hat{\theta}) - \mathrm{cov}(\tilde{\theta})$ is positive semi-definite). Then one can show that:

$$\mathrm{tr}\{\frac{\partial^2 e(\theta)}{\partial \theta^2}\Big|_{\theta=\theta^*} \mathrm{cov}(\tilde{\theta})\} \leq \mathrm{tr}\{\frac{\partial^2 e(\theta)}{\partial \theta^2}\Big|_{\theta=\theta^*} \mathrm{cov}(\hat{\theta})\}$$

The above inequality is true because both the covariance matrix and the Hessian matrix at θ^* are positive semi-definite (the Hessian is positive semi-definite at θ^* since θ^* is a minimum of $e(\theta)$, so $e(\theta)$ is convex around θ^*). Therefore one can write:

$$\mathrm{tr}\{\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^*} \mathrm{cov}(\hat{\theta})\} - \mathrm{tr}\{\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^*} \mathrm{cov}(\tilde{\theta})\}$$

$$= \mathrm{tr}\{\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^*} [\mathrm{cov}(\hat{\theta}) - \mathrm{cov}(\tilde{\theta})]\} \geq 0$$

where the last inequality is obtained because the trace of the product of two positive semi-definite matrices is non-negative [6]. Therefore, the expected error due to using $\tilde{\theta}$ in the decision rule is less than the expected error due to using $\hat{\theta}$:

$$E\{\tilde{e}\} \leq E\{\hat{e}\}$$

It is possible to show that, by using additional unlabeled samples, estimates with smaller covariance matrices can be found. Therefore, better performance can be obtained without the additional cost of obtaining more training samples.

Let us assume that $\hat{\theta}$ is an estimate of θ^* obtained by using the training samples. Furthermore, assume that $\hat{\theta}$ is asymptotically unbiased and efficient (for example, maximum likelihood estimates always posses these properties [7]). In other words, for moderately large sample sizes $E\{\hat{\theta}\} \approx \theta^*$ and $\mathrm{cov}(\hat{\theta}) \approx I_s^{-1}$, where I_s is the Fisher information matrix [7]. The subscript "s" denotes that the Fisher information matrix corresponds to an estimate obtained by using training samples that are drawn from each class separately. The Fisher information matrix is positive semi-definite and is defined as follows:

$$I = E\{[\frac{\partial}{\partial \theta} \log f(x)][\frac{\partial}{\partial \theta} \log f(x)]^T\}$$

Now, let us assume that $\tilde{\theta}$ is another estimate of θ^* obtained by using some unlabeled samples in addition to the training samples. The unlabeled samples are drawn randomly from the mixture of the m classes. If $\tilde{\theta}$ possesses the same properties of asymptotic unbiasedness and efficiency, one can approximate $\text{cov}(\tilde{\theta})$ by I_c^{-1} where I_c is the Fisher information matrix corresponding to the estimate that is obtained by combining training and unlabeled samples. Provided that the unlabeled and training samples are independent, one can write:

$$I_c = I_s + I_u$$

where I_u is another information matrix corresponding to the information contained in the unlabeled samples for estimating θ^*. Since all the information matrices are positive semi-definite one can write $I_c \geq I_s$. Therefore, $\text{cov}(\tilde{\theta}) \leq \text{cov}(\hat{\theta})$. Thus, one can conclude that the expected error of the decision rule that uses $\tilde{\theta}$ is less than the one that is obtained by using $\hat{\theta}$.

The implication of the above statement is that, if reasonable estimates for the required parameters can be found that use both the training and unlabeled samples, then they should be used in the decision rule. In particular, the benefit of using such estimates over the ones obtained by training samples alone is that the Hughes phenomenon will occur at a higher dimension because the estimates obtained using both training and unlabeled samples provide lower expected classification error. Therefore, more features can be used without sacrificing the performance and in fact, the additional information in the new features may cause an improvement in the classification accuracy. In the next section some methods for estimating the probability density functions using both training and unlabeled samples are studied.

VI. Methods of Incorporating Unlabeled Samples

A. Parametric Case

A particular case of interest is when individual classes are multivariate Gaussian. In this case, the ML estimates of the parameters of the mixture density consisting of the m normal classes can be found by the Expectation-Maximization (EM) algorithm [8]. Assume that there are m Gaussian classes and from the i^{th} class N_i training samples are available. Denote these training samples by z_{ik} where i indicates the class (i=1,...,m), and k is the index of each particular sample. In addition, assume that N unlabeled samples denoted by x_k are available from the mixture density $f(x|\theta) = \sum_{i=1}^{m} \alpha_i f_i(x)$. The EM equations for approximating the ML estimates of the parameters of the mixture density are the following[3] [9]:

$$\alpha_i^+ = \frac{\sum_{k=1}^{N} \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)}}{N} \tag{3}$$

$$\mu_i^+ = \frac{\sum_{k=1}^{N} \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} x_k + \sum_{k=1}^{N_i} z_{ik}}{\sum_{k=1}^{N} \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} + N_i} \tag{4}$$

$$\Sigma_i^+ = \frac{\sum_{k=1}^{N} \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} (x_k - \mu_i^+)(x_k - \mu_i^+)^T + \sum_{k=1}^{N_i} (z_{ik} - \mu_i^+)(z_{ik} - \mu_i^+)^T}{\sum_{k=1}^{N} \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} + N_i} \tag{5}$$

[3]Here we assume that the training samples are drawn separately from each class and therefore contain no information regarding the prior probabilities of the classes. If training samples are randomly drawn from the data set, equation (3) must be modified [9].

where μ_i and Σ_i are the mean vector and the covariance matrix of class i, and superscripts "c" and "+" denote the current and next values of the parameters respectively. The parameter set θ contains all the prior probabilities, mean vectors and covariance matrices. The ML estimates are obtained by starting from an initial point in the parameter space and iterating through the above equations. A reasonable starting point is the estimates obtained by using the training samples alone.

B. Nonparametric Case

The form of the EM equations usually resembles the regular ML estimates with the distinction that to each unlabeled sample a set of weights is attached that shows the "degree of membership" of that sample to each component of the mixture. These weights are equal to the posterior probability of each component given the unlabeled sample and the current values of the parameters. Based on this, in [10] a nonparametric approach to mixture density identification is proposed that uses both training and unlabeled samples. First training samples are used to obtain "weights" for the unlabeled samples, and then these weights are used with the unlabeled samples to obtain better estimates of the component densities of the mixture. It is shown that the estimates obtained in this way have smaller variance than the nonparametric estimates that are based on training samples alone.

C. Semi-Parametric Case

The assumption of normality is often prohibitive in practice. Usually, variations in soil type and moisture, plantation time, etc., cause the class conditional pdf's to be multimodal. On the other hand, nonparametric methods are usually too sensitive to the shape and size of the kernels that are used for approximating the density functions. Here, we consider a semi-parametric approach for density estimation. The individual classes are allowed to have multiple normal components. The pdf of each class is therefore modeled by a normal mixture density. By varying the number of components, various models for each class can be obtained. Theoretically, every smooth

density function can be approximated to within any accuracy by a mixture of normals. Therefore, the presented method is justified. We use the EM algorithm to find the ML estimates of the parameters when both training and unlabeled samples are present.

Let us assume that there are J classes in the feature space denoted by $S_1,...,S_J$. Each class can have several Gaussian components. Let m denote the total number of the Gaussian components. We write $i \in S_j$ to indicate that component i belongs to class S_j. The pdf of the feature can then be written as a mixture of m Gaussian components where the set of components can be partitioned into m classes:

$$f(x|\theta) = \sum_{i=1}^{m} \alpha_i f_i(x|\varphi_i)$$

where $\varphi_i = (\mu_i, \Sigma_i)$, $\theta = (\alpha_1,..., \alpha_m, \mu_1,..., \mu_m, \Sigma_1,..., \Sigma_m)$.

From each class S_j, N_j training samples are assumed to be available. We denote these samples by z_{jk} where j=1,...,J indicates the class of origin and k=1,...,N_j is the index of each particular sample. The training samples here are known to come from a particular class without any reference to the exact component within that class. In addition to the training samples, N unlabeled samples denoted by x_k, k=1,...,N, are also assumed to be available from the mixture. The log likelihood to be maximized for obtaining the ML estimates can be written in the following form:

$$L(\theta) = \sum_{k=1}^{N} \log f(x_k|\theta) + \sum_{j=1}^{J} \sum_{k=1}^{N_j} \log \left(\frac{1}{\sum_{t \in S_j} \alpha_t} \sum_{l \in S_j} \alpha_l f_l(z_{jk}|\varphi_l) \right)$$

The first term in the above log likelihood function is the likelihood of the unlabeled samples with respect to the mixture density, and the second term indicates the likelihood of the training

samples with respect to their corresponding classes of origin. The EM equations for obtaining the ML estimates are the following [11] (see [12] for the details):

$$\alpha_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})}{N(1 + \frac{N_j}{\sum_{r \in S_j} \sum_{k=1}^{N} P^c(r|x_k)})} \tag{6}$$

$$\mu_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k)x_k + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})z_{jk}}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})} \tag{7}$$

$$\Sigma_i^+ = \frac{\sum_{k=1}^{N} P^c(i|x_k)(x_k - \mu_i^+)(x_k - \mu_i^+)^T + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})(z_{jk} - \mu_i^+)(z_{jk} - \mu_i^+)^T}{\sum_{k=1}^{N} P^c(i|x_k) + \sum_{k=1}^{N_j} P_j^c(i|z_{jk})} \tag{8}$$

where $i \in S_j$, and $P^c(.|.)$ and $P^c_j(.|.)$ are the current values of the posterior probabilities:

$$P^c(i|x_k) = \frac{\alpha_i^c f_i(x_k|\mu_i^c, \Sigma_i^c)}{f(x_k|\theta^c)} \qquad P_j^c(i|z_{jk}) = \frac{\alpha_i^c f_i(z_{jk}|\mu_i^c, \Sigma_i^c)}{\sum_{t \in S_j} \alpha_t^c f_t(z_{jk}|\mu_t^c, \Sigma_t^c)}$$

VII. Experimental Results

The equation 3, 4, and 5 were used with the data and training set of experiment 1 to demonstrate the effect of unlabeled samples in enhancing the performance. Experiment 1 was repeated but with 20 training samples and an additional number of unlabeled samples used via the above equations for estimating the parameters. Subsequently the rest of the samples were classified according to the MAP decision rule (which also incorporates the second order statistics). The experiment was performed once with 500 unlabeled samples and once with 1000 unlabeled samples.

Notice also that an additional benefit of using unlabeled samples is that since the prior probabilities of the classes can be obtained, instead of the ML classifier, the MAP classifier can be constructed. Without the unlabeled samples, generally the prior probabilities can not be estimated, because the training samples are usually obtained separately from each class. Figure 4 shows the results. In this figure, the curve for the case where only training samples are used is also shown for comparison and is labeled "supervised." [4]

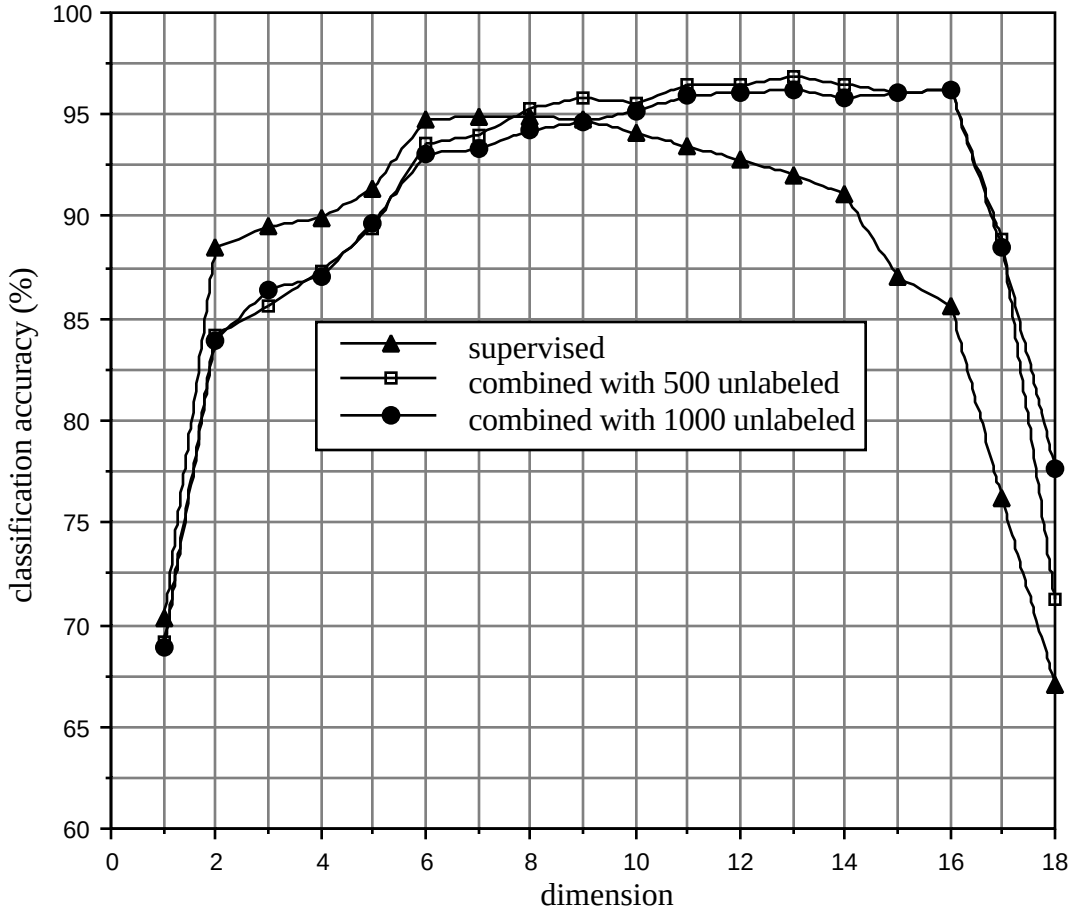

Figure 4: Effect of additional unlabeled samples in the classification performance for experiment 1 (AVIRIS data) with 20 training samples/class.

[4]The graphs published in the proceedings of IGARSS'93 paper are corrected here. In the former paper, the tails of the curves were shown incorrectly to decay too rapidly, hence the Hughes phenomenon was exaggerated.

From Figure 4 it can be seen that the use of additional unlabeled samples in the learning process can enhance the classification performance when the dimensionality of data begins to approach the number of training samples. In experiment 1, the Hughes phenomenon that began around dimension 8 when supervised learning is used, is delayed to dimension 16 when 500 or 1000 additional unlabeled samples are incorporated. Meanwhile, the minimum error for the supervised learning case was 5.42% and was achieved at dimension 7. For the cases with additional 500 and 1000 unlabeled samples, the minimum errors were 3.11% and 3.78% at dimensions 13, and 16 respectively. Therefore, the use of additional unlabeled samples not only delayed the occurrence of the Hughes phenomenon but also made the information in the new features usable for decreasing the error further.

VIII. Effect Of Unlabeled Samples In Reducing The Unrepresentative Training Samples Problem

In remote sensing, the training samples are usually selected from spatially adjacent regions. Often, the spatial correlation among the neighboring pixels is high. This correlation is usually reduced rapidly as the distance between the pixels increases. This phenomenon, causes a problem when training samples are used alone for estimating the class parameters. Usually, the parameters estimated in this way are only representative of the training fields and their nearby area. The rest of the multi-spectral data is, therefore, not represented well. Thus, the classification results that are based on such training fields are not robust in the sense that by changing the training fields, the results might change significantly. Consequently, the selection of "good" training fields becomes a burden on the user's shoulders. Often, training fields are added and eliminated empirically. It is, therefore, desirable to be able to update the parameter estimates in a way to make them more representative of the whole image. When the unlabeled samples are added to the learning process, the parameter estimates get updated and become more representative of the whole data.

In Figure 5, the part of the AVIRIS data set that is used in experiments 1 is shown. Here, experiments 1 is repeated with the distinction that 20 adjacent training samples from each class are selected. The training fields are high lighted in Figure 5. We classify the data once by using only the training samples for estimating the parameters of the GML classifier, and once by adding 500 and 1000 randomly drawn unlabeled samples from the scene.

Figure 5: The AVIRIS site with training fields high lighted.

The classification accuracy is shown in Figure 6. To show how representative the estimated parameters were, the probability map [13] associated with the classification was obtained. The probability map is obtained by gray coding the Mahalanobis distance of each pixel for the class to which it was classified. Dark pixels are the ones that are classified with low conditional probabilities. Light pixels are the ones that are classified with high conditional probabilities. Figure 7 shows the probability map for the experiment. It is seen from this figure that when supervised learning was used the only bright spots were near the training fields. In order words, the rest of the data were not represented well. By adding unlabeled samples to the estimation process, more representative estimates are obtained, and thus the probability maps are brighter.

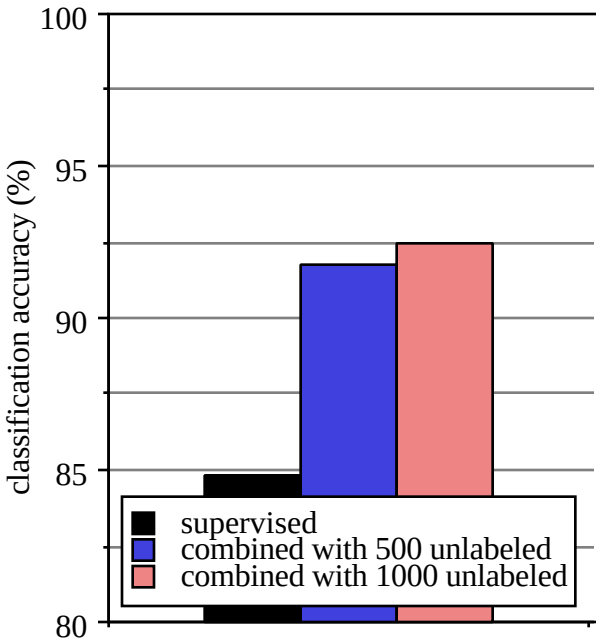

Figure 6: Classification results based on adjacent training samples.

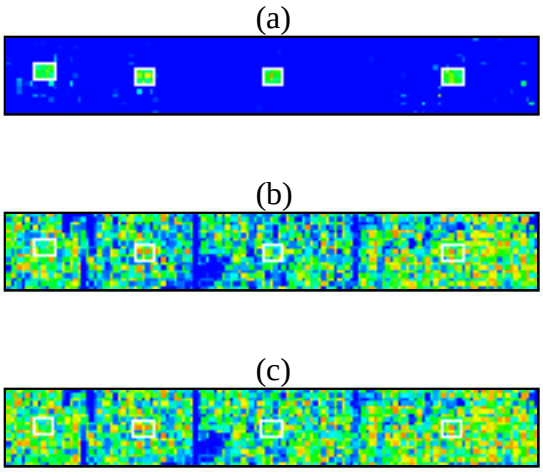

Figure 7: Probability maps for AVIRIS data set. (a) supervised learning, (b) combined with 500 unlabeled samples, (c) combined with 1000 unlabeled samples.

IX. Discussion and Concluding Remarks

In this paper, the effect of additional unlabeled samples in enhancing the classification performance was studied. It was observed that by incorporating unlabeled samples into the estimation process the Hughes phenomenon might be mitigated and the peak performance can be increased and shifted to a higher dimension. This phenomenon has several advantages. First, as it

was shown in section VII, when unlabeled samples are used, the peak performance was enhanced. In other words, the information in the new feature measurements can be used to further reduce the error. Without the unlabeled samples, the peak performance might occur at a lower dimension after which no further improvement can be obtained, and hence the new feature measurements are useless.

Second, the mitigation of the Hughes phenomenon is important in the feature extraction process. The feature extraction process is usually based on finding features that optimize a particular criterion. For example, in discriminant analysis within class and between class scatter matrices are estimated by using the training samples, and then features that optimize a function of these matrices are obtained. The purpose is, of course, to eliminate the less informative features and thereby speed up the classification process. However, if the estimates of the within and between class scatter matrices are not reliable (due to limited numbers of training samples), then the features obtained are not suitable. Using additional unlabeled samples can help obtaining better estimates of these matrices. Similarly, in the Decision Boundary Feature Extraction method [14], training samples are used to obtain a decision boundary in the original high dimensional space and then features that are relevant to this boundary are kept. Again, if training samples are limited, then the decision boundary in the original space is not suitable. Third, when the training samples are not good representatives of the true sampling distributions of the classes, the additional unlabeled samples may help update the class statistics and make them more representative.

An important practical point that needs to be kept in mind is that although in theory the additional unlabeled samples should always improve the performance, in practice this might not always be the case. For example, in Figures 4 it can be seen that when the dimensionality is small compared to the number of training samples the supervised learning process showed a slightly better performance than when unlabeled samples are added. The reason for this behavior is the deviation of the real world situations from the models that are assumed. For example, the

unlabeled samples that are drawn from the scene might contain outliers, boundary pixels, mixels, or samples of unknown classes. Such samples can hurt the performance. Therefore, care must be taken when combined supervised-unsupervised learning is used in practice[5]. Based on these issues the following steps for designing classifiers are suggested:

1) Estimate the Bayes error in order to have an understanding of the difficulty of the problem. Unlabeled samples can also be used for Bayes error estimation [15].

2) Design a classifier using the training samples alone.

3) Test the performance of the designed classifier (test samples, resubstitution, leave-one-out, etc.). Unlabeled samples can also be used for estimating the classification error of a classifier [16].

4) If the performance of the classifier was not satisfactory, draw a set of unlabeled samples and design a new classifier using both training and unlabeled samples. Test the classifier again and if necessary use more unlabeled samples.

References

[1] V.V. Salomonson, W.L. Barnes, P.W. Maymon, H.E. Montgomery, H. Ostrow, "MODIS: Advanced Facility Instrument for Studies of the Earth as a System," *IEEE Trans. Geoscience & Remote Sensing*, Vol. 27, No. 2, March 1989, pp 145-153.

[2] G. Vane, R.O. Green, T.G. Chrien, H.T. Enmark, E.G. Hansen, W.M. Porter, "The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sens. Environ*, 44, 1993, pp 127-143.

5 A version of the iterative algorithm discussed in this paper has been implemented in a general purpose multispectral image data analysis sytem called MultiSpec (© Purdue Research Foundation). Information about MultiSpec, which runs under the Macintosh operating system, may be obtained via electronic mail from landgreb@ecn.purdue.edu.

[3] K. Fukunaga, *Intro. Statistical Pattern Recognition*, San Diego: Academic Press Inc., 1990.

[4] G.F. Hughes, "On The Mean Accuracy Of Statistical Pattern Recognizers," *IEEE Trans. Infor. Theory*, Vol. IT-14, No. 1, pp 55-63, 1968.

[5] C. Lee, D.A. Landgrebe, "Analyzing High Dimensional Multispectral Data," *IEEE Trans. Geoscience & Remote Sensing*, Vol. 31, No. 4, July 1993, pp 792-800.

[6] F. A. Graybill, *Matrices With Applications In Statistics*, Belmont: Wadsworth Inc., 1983.

[7] H.W. Sorenson, *Parameter Estimation: Principles And Problems*, New York: M. Dekker, 1980.

[8] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood Estimation from Incomplete Data via EM Algorithm," *J. R. Statist. Soc.*, B 39, pp 1-38, 1977.

[9] R.A. Redner, H.F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, Vol. 26, No. 2, pp 195-239, 1984.

[10] P. Hall, D.M. Titterington, "The Use of Uncategorized Data to improve the Performance of a Nonparametric Estimator of a Mixture Density," *J.R. Statist. Soc.* B 47, pp 155-163, 1985.

[11] B.M. Shahshahani , D.A. Landgrebe, "Using Partially Labeled Data for Normal Mixture Identification with Application to Class Definition," *in Proc. IEEE International Geoscience & Remote Sensing Symposium*, pp 1603-1605, 1992.

[12] B.M. Shahshahani, *PhD Dissertation*, Purdue University, December 1993.

[13] D.A. Landgrebe and L. Biehl, *An Introduction to MultiSpec*, School of Electrical Engineering, Purdue University, IN 47907-1285.

[14] C. Lee, D.A. Landgrebe, "Feature Extraction Based on Decision Boundaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol 15, No 4, pp388-400, 1993.

[15] K. Fukunaga, D. Kessell, "Nonparametric Bayes Error Estimation Using Unclassified Samples," *IEEE Trans. Info. Theory*, Vol. IT-19, pp 434-440, 1973.

[16] D.S. Moore, S.J. Whitsitt, D.A. Landgrebe, "Variance Comparisons for Unbiased Estimators of Probability of Correct Classification," *IEEE Trans. Info. Theory*, 22, pp 102-105, 1976.