NASA-CR-204756

USE OF UNLABELED SAMPLES FOR MITIGATING THE HUGHES PHENOMENON[1] *

Behzad M. Shahshahani and David A. Landgrebe

School of Electrical Engineering
Purdue University
West Lafayette, IN 47906
Tel: (317) 494-1743; Fax: (317) 494-6440
behzad@ecn.purdue.edu landgreb@ecn.purdue.edu

Abstract

The use of unlabeled samples in improving the performance of classifiers is studied. When the number of training samples is fixed and small, additional feature measurements may reduce the performance of a statistical classifier. It is shown that by using unlabeled samples estimates of the parameters can be improved and therefore this phenomenon may be mitigated. Various methods for using unlabeled samples are reviewed and experimental results are provided.
Key Words: Hughes Phenomenon, Small Sample Size, Classifier Design

I. Introduction

Multispectral sensor technology is advancing towards the production of sensors with larger numbers of spectral bands. The hope is that the detailed spectral responses produced by these sensors could be used for discriminating between more classes thus, providing better understanding of the nature of the materials populating the scene. For example, high dimensional spectral responses are considered very informative for distinguishing among sub-classes of a particular ground cover class. In order to design a classifier, training samples from each sub-class are required. It becomes difficult and expensive to obtain large training sample sets for each sub-class. Consequently, the conventional problem of small sample size versus high dimensionality arises. This problem, often referred to as the Hughes phenomenon [1], is the loss of classifiability that is observed when the dimensionality of the data increases while the training sample size remains fixed.

II. Hughes Phenomenon

The minimum achievable error in a classification problem is the Bayes error. A decision rule that assigns a sample to the class that has the maximum *a posterior* probability (MAP classifier) achieves the Bayes error. In order to design such a classifier, knowledge of the posterior probabilities and thus, the class conditional probability density functions is required. If such knowledge is available then by increasing the dimensionality one would expect to enhance the performance. In other words, the Bayes error is a decreasing function of the dimensionality of the data. 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.

III. Classification Error

Consider a classification problem involving m classes with prior probabilities P_i and probability density functions $f_i(x)$. By e^* we denote the Bayes error achieved by using the MAP classifier when P_i and $f_i(x)$ are known. Let θ denote the vector of parameters of the MAP classifier. Let θ^* denote the true value of θ. The error achieved by using θ^* in the decision rule is e^*, the Bayes error. Now, let's assume that $\hat{\theta}$ is an estimate of θ^*. If 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:

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

where $tr(A)$ denotes the trace of matrix A and $H(\theta^*)$ is the Hessian evaluated at θ^* ($\left.\frac{\partial^2 e(\theta)}{\partial \theta^2}\right|_{\theta=\theta^*}$). 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:

$$E\{\hat{e}\} \approx e^* + \frac{1}{2} tr\{H(\theta^*) cov(\hat{\theta})\}$$

The second term in the right hand side is positive since both the covariance matrix and the Hessian 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^*). Now, consider another unbiased estimate $\bar{\theta}$. If $cov(\hat{\theta}) \leq cov(\bar{\theta})$ (in other words $cov(\bar{\theta}) - cov(\hat{\theta})$ is positive semi-definite) then one can easily show that :

$$tr\{H(\theta^*) cov(\hat{\theta})\} \leq tr\{H(\theta^*) cov(\bar{\theta})\}$$

IV. Effect of Additional Unlabeled 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). In other words, for large sample sizes, $E\{\hat{\theta}\} \approx \theta^*$ and $cov(\hat{\theta}) \approx I_s^{-1}$, where I_s is the Fisher information matrix. The subscript s denotes that the Fisher information matrix corresponds to a supervised estimate obtained by using training samples that are drawn from each class separately. The Fisher information matrix is positive definite and is defined as follows:

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

[1] This work was supported in part by NASA under Grant NAGW-925.

*Proceedings of the International Geoscience and Remote Sensing Symposium(IGARSS'93), Tokyo, pp 1535-7, August 1993.

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 $cov(\tilde{\theta})$ 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 of the information matrices are positive definite one can write $I_c \geq I_s$. Therefore, $cov(\tilde{\theta}) \leq cov(\hat{\theta})$. From the developments of section III, 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 $\hat{\theta}$:

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

Therefore a decision rule that is obtained by incorporating the unlabeled samples in the estimation process achieves a lower error rate. Thus, by using such a decision rule the Hughes phenomenon may be delayed to a higher dimension. Consequently, 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.

V. Methods of Incorporating Unlabeled Samples

Parametric Case

A particular case of interest is when individual classes are multivariate Gaussian. In this case, the maximum likelihood (ML) estimates of the parameters of the mixture density consisting of the m normal classes can be found by the EM (Expectation-Maximization) algorithm [2]. Assuming that n unlabeled samples denoted by x_k are available from the mixture density $f(x/\theta) = \sum_{i=1}^{m} P_i f_i(x)$, and n_i training samples denoted by z_{ik} are available from each class i, the ML estimates of the parameters of the mixture can be found by using the following iterative equations [3]:

$$P_i^+ = \frac{\sum_{k=1}^{n} P^c(i/x_k)}{n} \quad , \quad \mu_i^+ = \frac{\sum_{k=1}^{n} P^c(i/x_k).x_k + \sum_{k=1}^{n_i} z_{ik}}{\sum_{k=1}^{n} P^c(i/x_k) + n_i}$$

$$\Sigma_i^+ = \frac{\sum_{k=1}^{n} P^c(i/x_k)(x_k - \mu_i^+)(x_k - \mu_i^+)^T + \sum_{k=1}^{n_i} (z_{ik} - \mu_i^+)(z_{ik} - \mu_i^+)^T}{\sum_{k=1}^{n} P^c(i/x_k) + n_i}$$

where $$P^c(i/x_k) = \frac{P_i^c f_i(x_k/\mu_i^c, \Sigma_i^c)}{f(x_k/\theta^c)}$$

and μ_i and Σ_i are the mean vector and the covariance matrix of class i, superscripts + and c denote the next and current values of the parameters respectively. The parameter set θ contains all the prior probabilities, mean vectors and covariance matrices.

Multi-Component Classes

Another interesting case that arises regularly in remote sensing is when the individual classes have multiple components. If each such component is assumed to be multivariate Gaussian, then the class conditional pdf's, $f_i(x)$, are themselves mixture densities. Often training samples are known to belong to a specific class without any reference to the particular component within that class. In this case again the EM algorithm can be used to obtain the ML estimates of the parameters. The EM equations are presented in [4].

Non Parametric Case

Unlabeled samples can also be utilized in the nonparametric estimation of the pdf's. One technique for doing so was proposed in [5]. Let $\hat{f}_i(x)$ be a nonparametric estimate of $f_i(x)$ obtained by using the training samples. Let $\hat{f}(x)$ be the nonparametric estimate of the mixture density $f(x)$ based on the training samples (obtained by a linear sum of $\hat{f}_i(x)$), and $f^*(x)$ be a nonparametric estimate of $f(x)$ obtained by using the unlabeled samples alone. Then a nonparametric estimate $\tilde{f}_i(x)$ of $f_i(x)$ based on both the training samples and unlabeled samples can be constructed by the following formula:

$$\tilde{f}_i(x) = \frac{\hat{f}_i(x)}{\hat{f}(x)} f^*(x)$$

The estimate $\tilde{f}_i(x)$ was shown to have a lower variance than $\hat{f}_i(x)$ [5].

VI. Experimental Results

Experiment 1 (AVIRIS data)

A portion of an AVIRIS frame (consisting of 200 bands) taken over Tippecanoe county in Indiana was considered in this experiment. Four ground cover classes were determined by using the ground truth map. The classes were bare soil (380 pixels), wheat (513 pixels), soybean (741 pixels), and corn (836 pixels). Dimensionality of the data was changed from 1 to 18 by sequentially adding more bands. Selection of a new band was based on the average pairwise Bhattacharrya distance measure. Twenty training samples were drawn randomly from each class. The statistics of each class were estimated once by using only the twenty training samples and once using both the twenty training samples and some unlabeled samples drawn from the total field. The parametric equations for the single component per class case were used to obtain the ML estimates. Consequently, the rest of the samples were classified and the total classification accuracy was computed. Each experiment was repeated ten times independently and the average of the ten trials was obtained. The results are shown in Figure 1. In Figure 1, the supervised curve refers to the case when training samples were used alone in the estimation process and maximum likelihood classification was performed consequently. The curves labeled combined 500 and combined 1000 refer to the cases when 500 and 1000 unlabeled samples were used in addition to the training samples respectively and MAP classification was subsequently performed.

It can be seen from Figure 1 that the Hughes phenomenon starts to appear around dimension 8 when supervised learning is used. It is delayed to dimension 14 when 500 unlabeled samples were used and to 16 when 1000 unlabeled samples were used. The minimum error for the supervised case was 5.93% and was achieved at dimension 8. For the cases with 500 and 1000 unlabeled samples, the minimum errors were 3.78% and 3.87% both at dimension 13. Therefore, the use of 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.

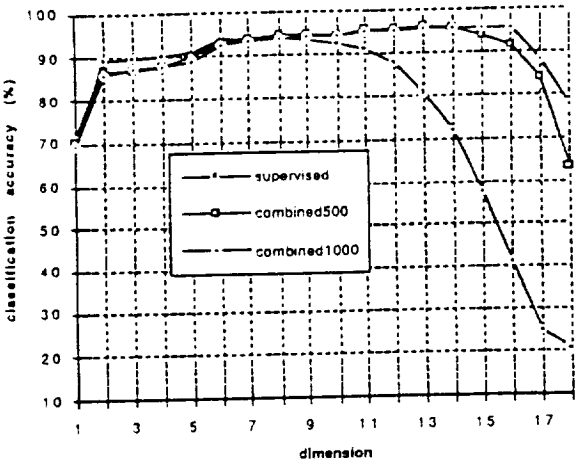

Figure 1: Classification performance versus dimensionality (AVIRIS data)

Experiment 2 (FLC1 Data)

The same kind of experiment was performed on a portion of the Flight Line C1 data set which is a 12 band multispectral image taken over Indiana. Four ground classes were determined using the ground truth map: corn (2436 pixels), soybean (2640 pixels), wheat (2365 pixels) and red clover (2793 pixels). dimensionality was changed from 1 to 12 by sequentially adding bands as described in experiment 1.

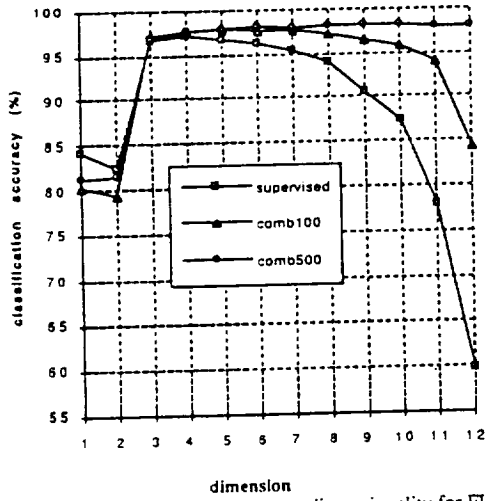

Figure 2: Classification performance versus dimensionality for FLC1 data.

From each class 15 training samples were drawn randomly and the statistics of each class were estimated once using these training samples alone, once using an additional 100 unlabeled samples, and once using 500 additional unlabeled samples. The rest of the samples were classified. Each experiment was performed 10 times independently and the average of the ten trials was obtained. The results are shown in Figure 2. In the supervised learning case, the Hughes phenomenon appeared at dimension 4, with 100 unlabeled sample it was delayed to approximately dimension 7, and with 500 unlabeled samples the Hughes phenomenon did not occur. The minimum error for the supervised learning case was 2.89% at dimension 4. For the combined100 and combined 500 cases it was 2.20% at dimension 5 and 1.86% at dimension 8, respectively.

VII. Concluding Remarks

The use of unlabeled samples for mitigating the Hughes phenomenon was investigated in this paper. Throughout the experiments it was seen that when the dimensionality is small (compared to the size of the training set) the estimates obtained by training samples alone are usually adequate for obtaining a reasonable decision rule. In this case, additional unlabeled samples did little to increase the accuracy, and in fact in our experiments they sometimes reduced the accuracy slightly. This result is most likely due to the fact that the unlabeled samples may contain outliers from unknown classes and/or from boundaries of the fields. However, when the dimensionality is high (compared to the number of training samples), unlabeled samples can significantly improve the accuracy. Since it is very difficult to guess the optimal number of features related to the training sample size in advance, it becomes important to reduce the effect of the Hughes phenomenon. For example, recently a new feature extraction method was proposed that is based on reducing the dimensionality of the data at the same time that the performance remains similar to the original space [6]. In such a case, one would like to ensure that the performance in the original higher dimensional space is not severely effected by the Hughes phenomenon. Based on the material presented in this paper, we suggest the following steps for designing classifiers when training samples are limited:

1) If possible, estimate the Bayes error in order to have an understanding of the difficulty of the problem.

2) Design a classifier using the training samples alone.

3) Test the performance of the designed classifier (test samples, resubstitution, leave-one-out, etc.).

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.

VIII. References

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

[2] 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.

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

[4] 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

[5] 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

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