Lowpass Filter For Increasing Class Separability

Pi-Fuei Hsieh and David Landgrebe
School of Electrical & Computer Engineering, Purdue University
West Lafayette IN 47907-1285
hsieh@ecn.purdue.edu

Abstract. In remote sensing, the number of training samples is often limited. For hyperspectral data, it becomes more difficult to obtain accurate estimates of class statistics because of the small ratio of the training sample size to dimensionality. Generally speaking, classification performance depends on four factors: class separability, the training sample size, dimensionality, and classifier type (or discriminant function). To improve classification performance, attention is often focused on seeking improvements on the factors other than class separability because class separability is usually considered inherent and predetermined. The objective of this paper is to call attention to the fact that class separability can be increased. The lowpass filter is proposed as a means for increasing class separability if a data set consists of multi-pixel objects. In addition, an analysis procedure is proposed in the following order: the lowpass filter, the EM algorithm, feature extraction, and a maximum likelihood classifier. Experiments with hyperspectral data show that increasing class separability compensates for the loss of the classification accuracy caused by the poor statistics estimation due to the small ratio of the training sample size to dimensionality.

INTRODUCTION

When the number of training samples is relatively small compared to the dimensionality, maximum likelihood estimates of parameters have large variances, leading to a large classification error [1]. Quite often, the small training sample size problem is encountered in hyperspectral data analysis. Although class separability usually increases as dimensionality increases, the growth of classification accuracy due to high class separability often fail to compensate for the loss of the accuracy of parameter estimation. As a result, a peaking phenomenon appears in the relation of classification accuracy versus dimensionality. This is often referred to as the Hughes phenomenon [2]. Several methods have been proposed for mitigating the Hughes phenomenon. Examples include Leave-One-Out Covariance Estimation [3], and the EM algorithm [4]. Each has provided a certain degree of improvement by reducing dimensionality, selecting classifier types, and increasing the effective number of training samples, respectively. In this paper, a new approach is proposed, from the aspect of class separability.

EFFECT OF CLASS SEPARABILITY ON THE HUGHES PHENOMENON

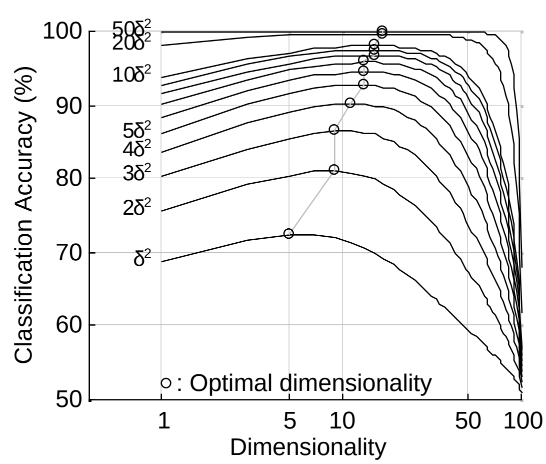
The classification performance is usually evaluated by classification accuracy. Consider two equally likely classes that are characterized by normal distributions. The number of training samples for each class is assumed to be finite and fixed. By means of simulation [5], an asymptotic expression for classification errors is given

$$\varepsilon \approx \Phi\left(-\frac{\delta}{2}\frac{1}{\sqrt{\alpha}}\right)$$

where Φ is a cumulative function of the standard normal distribution, δ^2 is the Mahalanobis distance and

$$\alpha = 1 + \frac{(\delta^2 + n)^2 + n^2}{2\delta^2 (N - n)}$$

for the quadratic discriminant function. To illustrate the relation of classification accuracy versus dimensionality, let us consider a model [6] for class separation: $\delta_n^2 = \delta_1^2 n^{1/R}$, where the degree of class separation is expressed in terms of the Mahalanobis distance and increases monotonically with dimensionality. R=3 is chosen for the study of the Hughes phenomenon. The figure below shows the effect of class separability on the Hughes phenomenon. The parameter δ refers to the case in which the degree of class separability is increased by a positive scalar δ^2 (δ>1). As class separability increases, the peak shifts upwards and to the right, indicating that the classification accuracy at all dimensionality improves.



Presented at the International Geoscience and Remote Sensing Symposium, Seattle, WA July 6-10, 1998.

USE OF THE LOWPASS FILTER FOR INCREASING CLASS SEPARABILITY

The lowpass filter has long been used as a smoothing technique in image processing [7,8]. From the standpoint of pattern recognition, the lowpass filter can be considered, from a different angle, that of a means for increasing class separability. The key requirement for this approach is that a data set contains a wealth of information about class-dependent spatial correlation. By reducing the variation of samples within each class, the lowpass filter widens the gap between classes in the feature space and reduces the demand on the precision of decision boundary, leading to higher classification accuracy. Also, the lowpass filter is easy to implement. Each sample is replaced by the weighted average of its neighboring samples within a user specified "window". Equal weighting is used in this study for the sake of simplicity and efficiency. Let X_1, X_2, …, X_w represent the samples within a "window" of window size w. Assume that the samples in the window are independent and identically distributed random vectors of the normal density $N(\mu, \Sigma)$, then the lowpass filtered sample, $Y=(1/w)(X_1+X_2+\ldots+X_w)$, possesses a normal density $N(\mu, \Sigma/w)$, where the covariance matrix is scaled down by w and the mean vector remains the same. Use of the lowpass filter reduces the variation within a class but does not change the location of the class mean in the feature space. To have the maximum amount of reduction in variation, it is desirable to have equal weighting. To relate the effect of the lowpass filter to class separability, let us consider the Bhattacharyya distance between two normal distributions. The Bhattacharyya distance is widely used as a measure of class separability because of its analytical form and its relation to the Bayes error.

$$B_Y = \frac{1}{8}(\mu_{1Y}-\mu_{2Y})^T\left(\frac{\Sigma_{1Y}+\Sigma_{2Y}}{2}\right)^{-1}(\mu_{1Y}-\mu_{2Y}) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma_{1Y}+\Sigma_{2Y}}{2}\right|}{\sqrt{|\Sigma_{1Y}||\Sigma_{2Y}|}} \triangleq B_{1Y}+B_{2Y} = wB_{1X}+B_{2X}.$$

The first term and the second term represent the class separability due to the mean difference and covariance difference, respectively. Note that the "mean difference" used here is in the sense of the Mahalanobis "distance" rather than the Euclidean distance. When a lowpass filter is applied to each class, the first term increases by w times whereas the second term remains the same. In contrast, the mean difference in Euclidean terms remains the same because use of the lowpass filter does not change the locations of class means. If two classes have equal covariance matrices, the Bhattacharyya distance becomes a form of the Mahalanobis distance. The optimal classification performance is usually expressed by the Bayes error, which can be bounded by the Bhattacharyya distance [9].

$$\varepsilon_{Bayes,Y} \le \sqrt{P_1 P_2}\, e^{-(wB_{1X}+B_{2X})}$$

After a lowpass filter of window size w is used, the first term of the Bhattacharyya distance increases by w. Thus, the bound on the Bayes error decreases exponentially with w. If the original class separability due to the mean-difference is large, the Bayes error declines rapidly. However, if two classes have common mean vectors, use of the lowpass filter does not reduce the Bayes error.

USE OF THE LP FILTER IN COMBINED SUPERVISED-UNSUPERVISED LEARNING

It has been noted [4] that class separability is of importance to the performance of combined supervised-unsupervised learning. When classes are well separated, combined supervised-unsupervised learning can perform comparably to supervised learning. The Expectation Maximization (EM) algorithm has been proposed as a method of the combined supervised-unsupervised learning [4]. The EM algorithm often converges to a local maximum of the likelihood function if classes are highly overlapped or the number of samples is small. Increasing class separability is effective for alleviating the convergence problem. The EM algorithm aims at increasing the effective number of training samples while the lowpass filter is used for increasing class separability. Since increasing class separability may improve the performance of the EM algorithm, the LP filter is proposed to be used before the EM algorithm.

EXPERIMENTS WITH REAL DATA

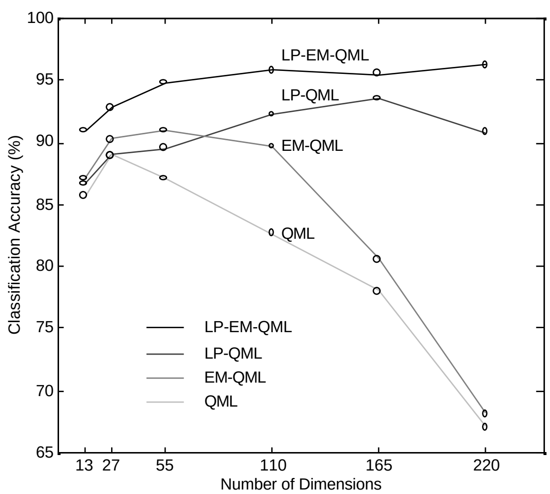
Test-1: The Hughes Phenomenon. The objective of this experiment is to test the performance of the methods that have been proposed for mitigating the Hughes phenomenon. A portion (size 85x68) of AVIRIS 1992 Indian Pine Test Site 3 was used. The full set of 220 channels was used. For dimensionality of 110, 55, 27, and 13, every 2nd, 4th, 8th, and 16th channels were selected, respectively. For dimensionality of 165, every fourth channel was removed. Four classes were defined: Corn-notill, Soybean-notill, Soybean-min, and Grass. There were 230 training samples used for each class, and 910, 638, 1421, and 618 test samples used for these classes, respectively. Four procedures, denoted by QML, EM-QML, LP-QML, and LP-EM-QML, were compared. LP-EM-QML refers to the procedure where the lowpass filter (LP) and the EM algorithm (EM) were followed by the quadratic ML classifier (QML). This data set is considered suitable for use of LP because it consists of multipixel homogeneous objects and there is a difference in class means. This data set is also suitable for use of the EM algorithm since the list of classes is exhaustive and the assumption of normal distributions seems appropriate. A lowpass filter of window size 3x3 was used. Sample means and sample covariances were used as the starting point for the EM algorithm. The new estimates of statistics generated by the EM algorithm were used to design a maximum likelihood classifier for classifying test samples. The EM iteration stopped after 20 iterations. All samples other than training samples were used as "unlabeled" samples that were mentioned in the EM algorithm. That is, the "unlabeled" samples used in the EM algorithm included "test" samples.

The following figure shows, comparing the QML and EM-QML cases, that the EM algorithm mitigated the Hughes phenomenon when the number of dimensions was not large. However, due to a finite number of unlabeled samples, the performance of the EM algorithm was poor at high dimensionality. As class separability was increased by using the lowpass filter, the Hughes phenomenon was alleviated.

Presented at the International Geoscience and Remote Sensing Symposium, Seattle, WA July 6-10, 1998.

When lowpass filtering was subsequently used, the overall accuracy increased, and consequently the peaking phenomenon disappeared. This implies that the optimal number of dimensions is the full dimensionality. Feature extraction methods can be used subsequently to find the subspace that retains discriminant information, as shown in Test-2.

Test-2: The complete analysis procedure. The data set was the same as in Test-1. Feature extraction (FE) was incorporated into the analysis procedures. In analyzing hyperspectral data, the information about discriminating among classes is often contained primarily in a smaller number of features. Feature extraction is used to remove redundant features in order to speed up classification. The objective of this experiment is to compare LP-EM-FE-QML with EM-FE-QML. Since feature extraction methods are based on class statistics, it is desirable to perform the EM algorithm ahead of feature extraction in order to provide better class statistics for feature extraction. Since the performance of the EM algorithm depends on class separability, the EM algorithm is preceded by the lowpass filter so as to obtain good class separability beforehand.



In this experiment, two feature extraction methods were used: the Decision Boundary Feature Extraction (DBFE) [11] and the Discriminant Analysis Feature Extraction (DAFE) [9]. DAFE generates at most L-1 discriminant features in order of significance (L is the number of classes) while DBFE generates features all sorted. To reduce the dimensionality, the first three features from DAFE and the best features from DBFE that had achieved a significance level of 99% were selected. There were 36 features selected for EM-DBFE-QML and seven features for LP-EM-DBFE-QML. The classification results are summarized in the table below. Results show that procedures with LP incorporated outperform procedures without LP.

Procedure without LP	Accuracy (%)	Procedure with LP	Accuracy (%)
QML	67.2 %	LP---------------QML	90.8 %
EM---------QML	68.2 %	LP-EM----------QML	96.0 %
EM-DBFE-QML	70.3 %	LP-EM-DBFE-QML	96.8 %
EM-DAFE-QML	81.2 %	LP-EM-DAFE-QML	97.0 %

It should be noted that the blurring effect of the lowpass filter on borders might have a serious impact on the performance of the EM algorithm. The blurred borders tend to become outliers or unknown classes. This may harm the performance of the EM algorithm. To avoid this EM outlier problem, it is wise to remove border samples by using a border mask. For detecting borders, scores of image segmentation or edge detection algorithms have been developed.

CONCLUSIONS

In this paper, the lowpass filter is proposed for increasing the class separability of a data set consisting of multipixel homogeneous objects. The effects of class separability on classification errors in supervised learning and in combined supervised-unsupervised learning were considered. It was shown that, by using a lowpass filter, the Hughes phenomenon was mitigated in both supervised learning and combined supervised-unsupervised learning. The combination of the LP filter and the EM algorithm has achieved the best performance for mitigating the Hughes phenomenon. Additional details about this method may be found in [12]. The work leading to this paper was supported in part by NASA Grant NAGW5-3975.

[1] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-11, No. 8, pp. 873-885, Aug. 1989.

[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, vol. IT-14, No. 1, pp. 55-63, 1968.

[3] J.P. Hoffbeck and D. A. Landgrebe, "Classification of high dimensional multispectral data," TR-EE 95-14, Purdue University, May 1995.

[4] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," IEEE Transactions on Geoscience and Remote Sensing, Vol. 32, No. 5, pp. 1087-1095, September 1994.

[5] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-2, no.3, pp. 242-252, May 1980.

[6] T.S. El-Sheikh and A.G. Wacker, "Effect of dimensionality and estimation on the performance of Gaussian classifiers," Pattern Recognition, vol. 12, pp. 115-126, 1980.

[7] A. K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Inc., 1989.

[8] A. Rosenfeld and A. C. Kak, Digital Picture Processing, second edition, Academic Press, 1982.

[9] K. Fukunaga, Introduction to Statistical Pattern

Recognition, second ed., San Diego: Academic Press Inc., 1990.
[10] B. M. Shahshahani and D. A. Landgrebe, "Classification of multi-spectral data by joint supervised-unsupervised learning," TR-EE 94-1, Purdue University, January 1994.[12]
[11] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," IEEE Trans. Pattern Anal. Machine Intell., vol. 15, no. 4, pp. 321-325, April 1993.
[12] Pi-Fuei Hsieh and David Landgrebe, "Classification Of High Dimensional Data," PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 98-4, May 1998 (121 pages)