NASA-CR-200334

High Dimensional Feature Reduction Via Projection Pursuit. *

Luis Jimenez and David Landgrebe
School of Electrical Engineering
Purdue University, West Lafayette IN 47907-1285
Phone: 317-494-1743 Fax: 317-494-6440
jimenez@ecn.purdue.edu landgreb@ecn.purdue.edu

Abstract

The recent development of more sophisticated remote sensing systems enables the measurement of radiation in many more spectral intervals than previous possible. An example of that technology is the AVIRIS system, which collects image data in 220 bands. As a result of this, new algorithms must be developed in order to analyze the more complex data effectively.

Data in a high dimensional space presents a substantial challenge, since intuitive concepts valid in a 2-3 dimensional space do not necessarily apply in higher dimensional spaces. For example, high dimensional space is mostly empty. This results from the concentration of data in the corners of hypercubes. Other examples may be cited.

Such observations suggest the need to project data to a subspace of a much lower dimension on a problem specific basis in such a manner that information is not lost. Projection Pursuit is a technique that will accomplish such a goal. Since it processes data in lower dimensions, it should avoid many of the difficulties of high dimensional spaces. In this paper, we begin the investigation of some of the properties of Projection Pursuit for this purpose.

Key Words: High Dimensional Space; Multispectral Data; Feature Reduction; Projection Pursuit; Projection Index; Classification.

I. Introduction

As the number of dimensions of high spectral resolution data increases, the capability to detect more detailed classes and to achieve increased classification accuracy should be possible. Supervised classification techniques must be trained with label samples, but the number of such samples is usually limited. Hughes proved that with a limited number of training samples there is a penalty in the classification accuracy as the number of features increases beyond some point (Hughes, 1968), thus this could become a problem in high dimensional cases. A number of techniques for feature extraction have been developed to reduce the dimensionality. Among those techniques are Principal Components, Discriminant Analysis, and Decision Boundary Feature Extraction (Lee & Landgrebe, 1993). These techniques estimate the statistics at full dimensionality in order to extract the relevant features for classification. If the number of training samples is not adequately large, the estimation of parameters in high dimensional data will not be accurate enough. As a result the estimated features may not be reliable. In order to avoid such difficulty, a pre-processing of the data that will produce a linear combination of features reducing the dimensionality is needed. Such reduction enables the estimation of parameters to be more accurate for feature extraction (see figure 1) (Lee & Landgrebe, 1993).

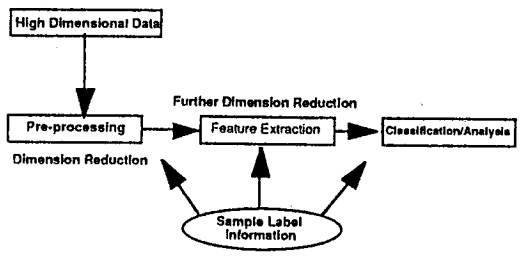

Figure 1. Pre-processing of high dimensional data.

There are a number of possibilities for pre-processing data in this fashion. A simple one might be to choose only every n^{th} channel, . Another might be averaging every n channels. None of these methods take into account a priori problem specific information. After considering the characteristics of the high dimensional space geometry of multispectral data, we propose the use of an algorithm that will linearly combine the features maximizing the distance between classes as a preprocess using labeled samples.

II. High Dimensional Space and the Curse of Dimensionality

The difficulty of dimensionality has been known for more than three decades, and its impact varies from one field to another. In combinatorial optimization over many dimensions it is seen as a growth of the computational effort exponentially with the number of dimensions. In statistics, it manifests itself as a problem of parameters or density estimation due to the sparcity of data. Such sparcity is produced by some geometrical properties of high dimensional feature space. Geometry characteristics exhibit surprising behavior of data in higher dimensions. For example, it has been shown that as the dimension increases (Scott 1992):

(1) The volume of a hypercube concentrates in the corners. The fraction of the volume of a hypersphere inscribed in a hypercube is:

$$f_{d1} = \frac{\text{volume}-\text{sphere}}{\text{volume}-\text{cube}} = \frac{\pi^{d/2}}{d\,2^{d-1}\Gamma\left(\frac{d}{2}\right)}$$

where d is the number of dimensions.

Note that $\lim_{d\to\infty} f_{d1} = 0$

(2) The volume of a hypersphere concentrates in the outside shell. The fraction of the volume of a sphere of radius $r-\varepsilon$ inscribed in another sphere of radius r is:

$$f_{d2} = \frac{V_d(r) - V_d(r-\varepsilon)}{V_d(r)} = \frac{r^d - (r-\varepsilon)^d}{r^d} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d$$

Note that : $\lim_{d\to\infty} f_{d2} = 1$

Both characteristics mentioned show that high dimensional space is mostly empty, which implies that multivariate data in R^d is almost always in a lower dimensional structure. A concrete consequence of the above discussion is that normally distributed data will have a tendency to concentrate on the tails; meanwhile uniform distributed data will be more likely to be collected in the corners, making density estimation extremely difficult.

(3) The diagonals are nearly orthogonal to all coordinate axis. The cosine of the angle between any diagonal vector and a Euclidean coordinate axis is:

$$\cos(\theta_d) = \pm\frac{1}{\sqrt{d}}\ , \text{ where } \lim_{d\to\infty}\cos(\theta_d) = 0$$

This piece of information is extremely important, because the projection of any cluster onto any diagonal, e.g., by averaging features, could destroy information of multispectral data.

In terms of parameter estimation, the number of samples required to make a given estimation is large in multispectral data., but in a nonparametric approach the number of samples required to estimate the density is even greater. As a consequence it is desirable to project the data to a lower dimensional space. Commonly used techniques, such as Principal Components or Discriminant Analysis have the disadvantage of requiring computations at full dimensionality, which results in the use of estimated statistics that are not necessarily accurate.

The goal is to make computations in a lower dimensional space

*Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'94), CD-ROM pp 1473-1479, Pasadena, CA, August 8-12, 1994.

where the projected data produce an "interesting" structure. A technique that will enable us to make such computations and to define what "interesting" means is Projection Pursuit.

III. Projection Pursuit

As has been defined (Scott 1992) "Projection pursuit is the numerical optimization of a criterion in search of the most interesting low-dimensional linear projection of a high dimensional data cloud." Projection Pursuit automatically picks an "interesting" lower dimensional projection of a high dimensional data maximizing or minimizing a function called the projection index. This technique is able to bypass many of the problems of high dimensionality by making the computations in a low dimensional space. A number of the classical methods that have been used for feature extraction are special cases of Projection Pursuit. With the use of suitable projection indexes, Principal Components and Discriminant Analysis were proved to be specific examples (Huber 1985).

The choice of the projection index is the most critical aspect of this technique. What "interesting" means depends on what function or projection index one uses. In remote sensing image analysis "interesting" would certainly be a projection where data separates into different meaningful clusters which are exhaustive, separable, and of information value (Swain & Davis, 1978).

For a mathematical interpretation let us define:
X: the initial multivariate data set (KxN).
A: an orthonormal matrix (MxK).
Y: the resulting projected data (MxN).
where $Y = A^TX$. A is the parameter matrix that optimizes the projection index $I(A^TX)$.

Many authors (Huber, 1985), (Jones & Sibson, 1987) use the computation of the negative of (Shannon) entropy as a projection index:

$$I(A^TX) = \int_{-\infty}^{\infty} f(Y)\log(Y)dY, \quad Y = A^TX$$

It is well known in Information Theory that entropy is maximized by the Normal distribution (Blahut 1987). Maximizing the negative entropy index will thus give the least normal projection. This type of linear projection would be expected to produce a multimodal density with the consequence of maximizing the separation between clusters. However, the negative entropy index has three main disadvantages. The first one is that one must center and spherize the data with the consequence of raising the contribution from noisy variables. The second disadvantage is that this index is suitable for a nonparametric approach which is wasteful of a priori information. The third disadvantage is that if there is not enough data the technique might divide a cluster of a class into different modes.

With that in mind we propose to use a parametric approach where maximizing a selected projection index will increase the distance among classes. The proposed projection index is the Bhattacharyya distance. For two classes the index will be:

$$I(A^TX) = \frac{1}{8}(M_{2Y} - M_{1Y})^T\left[\frac{\Sigma_{1Y}+\Sigma_{2Y}}{2}\right]^{-1}(M_{2Y} - M_{1Y}) + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma_{1Y}+\Sigma_{2Y}}{2}\right|}{\sqrt{|\Sigma_{1Y}||\Sigma_{2Y}|}}\right]$$

Here we will compute the matrix A that maximizes Bhattacharyya distance in the projected space. If there are more than two classes the average or the minimum of the combination of distances can be maximized. Let C be the number of classes. The index for average Bhattacharyya distance is:

$$I(A^TX) = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{8}(M_{2Y}^i - M_{1Y}^i)^T\left[\frac{\Sigma_{1Y}^i+\Sigma_{2Y}^i}{2}\right]^{-1}(M_{2Y}^i - M_{1Y}^i) + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma_{1Y}^i+\Sigma_{2Y}^i}{2}\right|}{\sqrt{|\Sigma_{1Y}^i||\Sigma_{2Y}^i|}}\right]$$

The index for minimum Bhattacharyya is:

$$I(A^TX) = \min_{i\in C}\left(\frac{1}{8}(M_{2Y}^i - M_{1Y}^i)^T\left[\frac{\Sigma_{1Y}^i+\Sigma_{2Y}^i}{2}\right]^{-1}(M_{2Y}^i - M_{1Y}^i) + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma_{1Y}^i+\Sigma_{2Y}^i}{2}\right|}{\sqrt{|\Sigma_{1Y}^i||\Sigma_{2Y}^i|}}\right]\right)$$

In the present paper this technique will be used as a preprocessing method of multispectral data. From the initial number of channels we can produce a linear combination to obtain a preliminary set of features. The advantage of this method is that one can determine a combination of features whose number is greater than the number of samples preserving information. After preprocessing the data in this way, we can use any method of feature extraction known to obtain the final set of features that will be used to classify the data.

IV. Experiments

The multispectral data used in these experiments is a segment of an AVIRIS frame taken of NW Indiana's Indian Pines test site. Only 88 features were used from the original 220 spectral channels. This data was obtain in June 1992. By that time most of the crops in the agricultural portion of the test site had not reached their maximum height. In this circumstance, species classification is a challenging problem, because the data gathered came not only from the vegetation but also from variations in soils, moisture, and residues.

Experiment I

Thirteen classes were defined after using a clustering algorithm and ground truth information. The number of training samples was greater than or equal to 60 samples/class. One of the classes reached the minimum (60 training samples). As a consequence Decision Boundary feature extraction could not be used for this preprocessing step. Four algorithms were used to reduce the number of dimensions. Three of them used to project the data from an 88 feature space to an 11 feature space: (1) averaging contiguous groups of eight features, (2) choosing 1 in every eight consecutive features, and (3) combining every eight contiguous features using Projection Pursuit with minimum Bhattacharyya as an index. The fourth algorithm used was Discriminant Analysis where the first 11 features were used.

The next step was the use of a feature selection algorithm to choose the combination of features for classification. The algorithm selects the best subset of features based on an average pairwise Bhattacharyya distance measure. After choosing the best combination of features from 1 to 10, the data was classified using a standard Maximum Likelihood classifier. A threshold was applied to the classifier results whereby 2% of the least likely points were thresholded. Figure 2 illustrates the results. The maximum accuracy was obtained by Projection Pursuit (70.9) with the use of three features. For that particular method, after the maximum was reached, the accuracy was maintained at near to the maximum level. The other methods, after reaching their maximum, began to decrease in accuracy significantly as the dimensionality increased. The results show that Projection Pursuit was able to function effectively with more features than other methods. This is particular relevant when it is recognized that typically as the number of classes increases the number of features needed to reach the maximum classification accuracy increases as well. With the use of Bhattacharyya index Projection Pursuit consolidates the data in different classes more than other algorithms. Consequently, more information is preserved. Among all the other algorithms, channel averaging preserved less information than any other method.

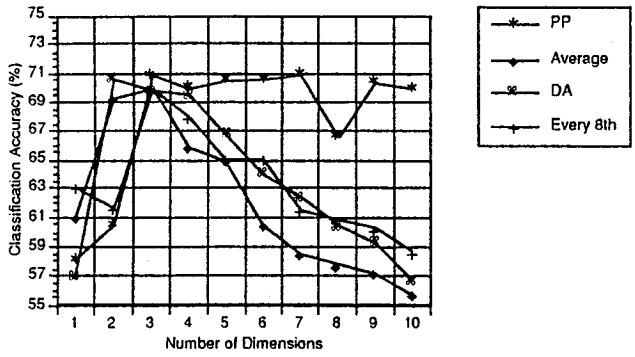

Figure 2. Classification accuracy vs. dimensionality.

Experiment II

In this experiment 5 classes where defined and only 60 samples/class were used to train the classifier. This experiment was designed to test how robust each pre-processing method is in relation to the Hughes phenomena. Because of the limit on the number of training samples, neither Decision Boundary nor Discriminant Analysis could be used. As a result we were constrained to use a pre-processing algorithm. Observe that Projection Pursuit has the advantage of being able to use a number of labeled samples (60) less than the number of features (88). After using the same procedure used in Experiment I, we plot the results in Figure 3. The figure shows that Projection Pursuit is more successful in terms of mitigating the Hughes phenomena. At almost all the dimensionality Projection Pursuit provided the greater accuracy.

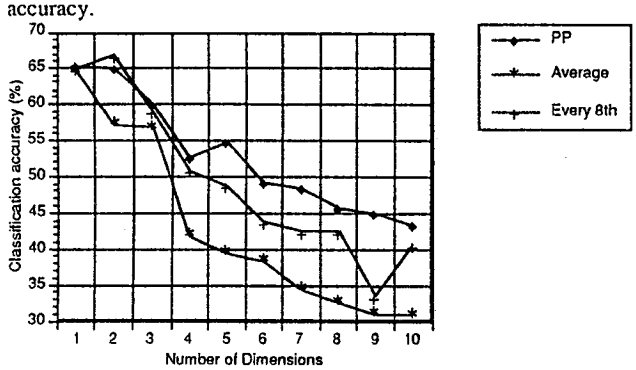

Figure 3. Classification accuracy vs. dimensionality.

V. Concluding Remarks

The increased number of features of modern data sources increase the amount of information which should be extractable from multispectral data. At the same time, since there is usually a limit on the number of training samples, degrading factors such as the Hughes phenomena and other characteristics of high dimensional data also increase as the number of dimensions increases. The challenge is to reduce the number of dimensions avoiding the obstacles inherent in the above mentioned phenomenon, while preserving maximum information and using a priori data. The use of Projection Pursuit as a pre-processing algorithm carries out this task better than using feature extraction or feature selection algorithms alone or using the two other pre-processing algorithms mentioned in figures 2 and 3, without a priori information.

A significant advantage of Projection Pursuit is that it enables the analyst to choose a particular index that relates to the specific application at hand. In our particular case an "interesting" projection is one that increases the distance among classes. For that reason we used Bhattacharyya distance as the index with the consequence that the data is better separated into the different classes.

VI. References

[1] Blahut, R.E. "Principles and Practice of Information Theory." Massachusetts: Addison-Wesley Publishing Company, 1987, pp 246.

[2] Huber, P.J. "Projection Pursuit." The Annals of Statistics Vol. 13 No 2 (1985): pp 435-475.

[3] Hughes, G.F. "On the Mean Accuracy of Statistical Pattern Recognizers." IEEE Trans. Info. Theory Vol. IT-14 No 1 (1968): pp 55-63.

[4] Jones, M.C., Sibson, R. "What is Projection Pursuit.", J. R. Statistics Soc. A Part 1 (1987): pp 1- 36.

[5] Lee, C., Landgrebe, D. A. "Feature Extraction and Classification Algorithms for High Dimensional Data." Technical Report, TR-EE 93-1 School of Electrical Engineering Purdue University, 1993.

[6] Scott, D.W. "Multivariate Density Estimation." New York: John Wiley & Sons, 1992.

[7] Swain P.H., Davis, S.M., eds. "Remote Sensing: The Quantitative Approach." New-York: McGraw-Hill, 1978, pp 340.