URC97063

Supervised Classification Techniques for Hyperspectral Data*

Luis O. Jimenez
Tropical Center for Earth and Space Studies
Electrical & Computer Engineering Department
University of Puerto Rico, Mayaguez, Puerto Rico
jimenez@exodo.upr.clu.edu

Abstract

The recent development of more sophisticated remote sensing systems enables the measurement of radiation in many mm-e spectral intervals than previous possible. An example of this technology is the AVIRIS system, which collects image data in 220 bands. The increased dimensionality of such hyperspectral data provides a challenge to the current techniques for analyzing such data. Human experience in three dimensional space tends to mislead one's intuition of geometrical and statistical properties in high dimensional space, properties which must guide our choices in the data analysis process. In this paper high dimensional space properties are mentioned with their implication for high dimensional data analysis in order to illuminate the next steps that need to be taken for the next generation of hyperspectral data classifiers.

I. Introduction

The complexity 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 an exponential growth of the computational effort with the number of dimensions. In statistics, it manifests itself as a problem with parameter or density estimation due to the paucity of data. The negative effect of this paucity results from some geometrical, statistical and asymptotical properties of high dimensional feature space. These characteristics exhibit surprising behavior of data in higher dimensions.

There are many assumptions that we make about characteristics of lower dimensional spaces based on our experience in three dimensional Euclidean space. There is a conceptual barrier that makes it difficult to have proper intuition of the properties of high dimensional space and its consequences in high dimensional data behavior. Most of the assumptions that are important for statistical purposes we tend to relate to our three dimensional space intuition, for example, as to where the concentration of volume is of such figures as cubes, spheres, and ellipsoids or where the data concentration is in known density function families such as normal and uniform. Other important perceptions that are relevant for statistical analysis are, for example, how the diagonals relate to the coordinates, the number of labeled samples required for supervised classification, the assumption of normality in data, and the importance of mean and covariance difference in the process of discrimination among different statistical classes. In the next section some characteristics of high dimensional space will be mentioned, and their impact in supervised classification data analysis will be discussed. Most of these properties do not fit our experience in three dimensional Euclidean space as mentioned before.

II. Geometrical, Statistical and Asymptotical Properties

In this section we illustrate some unusual or unexpected hyperspace characteristics including a discussion of its implications for supervised classification. These illustrations are intended to show that higher dimensional space is quite different from the three dimensional space with which we are familiar.

As dimensionality increases:

* Work reported herein was funded in part by NASA Grant NAGW-3924.

A. The volume of a hypercube concentrates in the corners and the volume of a hypersphere concentrates in an outside shell [Scott 1992].

These characteristics have two important consequences for high dimensional data that appear immediately. The first one is that high dimensional space is mostly empty, which implies that multivariate data in R^d is usually in a lower dimensional structure. As a consequence high dimensional data can be projected to a lower dimensional subspace without losing significant information in terms of separability among the different statistical classes. The second consequence of the foregoing, is that normally distributed data will have a tendency to concentrate in the tails; similarly, uniformly distributed data will be more likely to be collected in the corners, making density estimation more difficult. Local neighborhoods are almost surely empty, requiring the bandwidth of estimation to be large and producing the effect of losing detailed density estimation. Support for this tendency can be found in the statistical behavior of normally and uniformly distributed multivariate data at high dimensionality. It is expected that as the dimensionality increases the data will concentrate in an outside shell. As the number of dimensions increases that shell will increase its distance from the origin as well. Under these circumstances it would be difficult to implement any density estimation procedure and to obtain accurate results. Generally nonparametric approaches will have even greater problems with high dimensional data.

B. The required number of labeled samples for supervised classification increases as a function of dimensionality.

Fukunaga [Fukunaga 1989] proves that the required number of training samples is linearly related to the dimensionality for a linear classifier and to the square of the dimensionality for a quadratic classifier. That fact is very relevant, especially since experiments have demonstrated that there are circumstances where second order statistics are more relevant than first order statistics in discriminating among classes in high dimensional data [Lee and Landgrebe, July 1993], In terms of nonparametric classifiers the situation is even more severe. It has been estimated that as the number of dimensions increases, the sample size needs to increase exponentially in order to have an effective estimate of multivariate densities [Scott 1992, pp 208-212] [Hwang, Lay, Lippman 1994],

It is to be expected that high dimensional data contains more information. At the same time the above characteristics tell us that it is difficult with the current techniques, which are usually based on computations at fill dimensionality, to extract such information unless the available labeled data is substantial. A concrete example of this is the so-called Hughes phenomena. Hughes proved that with a limited number of training samples there is a penalty in classification accuracy as the number of features increases beyond some point [Hughes 1968].

C. For most high dimensional data sets', low linear projections have the tendency to be normal, or a combination of normal distributions, as the dimension increases.

That is a significant characteristic of high dimensional data that is quite relevant to its analysis. It has been proved [Diaconis and Freedman 1984] [Hall and Li 1993] that as the dimensionality tends to infinity, lower dimensional linear projections will approach a normality model with probability approaching one (see Figure 6). Normality in this case implies a normal or a combination of normal distributions.

In all the cases above we can see the advantage of developing an algorithm that will estimate the projection directions that separate the explicitly defined classes, doing the computations in a lower dimensional space. The vectors that it computes will separate the classes, and at the same time, the explicitly defined classes will behave asymptotically more like a normal distribution. The assumption of normality will be better grounded in the projected subspace than in fill dimensionality.

D. The role of the second order statistics become as important as the first order statistics.

Lee and Landgrebe [Lee and Landgrebe July 1993] performed an experiment where they classified some high dimensional data in order to see the relative role that first and second order statistics played.

In that particular experiment as the number of dimension grew the role played by the second order statistics increased in discriminating among classes. Under these circumstances, the shape of the distribution given by the second order statistics becomes as important as the location provided by the first order statistics.

III. High dimensional characteristics implications for supervised classification

Based on the characteristics of high dimensional data that the volume of hypercubes have a tendency to concentrates in the corners, and in a hyperellipsoid in an outside shell, it is apparent that high dimensional space is mostly empty, and multivariate data is usually in a lower dimensional structure. As a consequence it is possible to reduce the dimensionality without losing significant information and separability. Due to the difficulties of density estimation in nonparametric approaches, a parametric version of data analysis algorithms maybe expected to provide better performance where only limited numbers of labeled samples are available to provide the needed a priori information.

The increased number of labeled samples required for supervised classification as the dimensionality increases presents a problem to current feature extraction algorithms where computation is done at fill dimensionality, e.g. Principal Components, Discriminant Analysis and Decision Boundary Feature Extraction [Lee & Landgrebe, April 1993]. A new method is required that, instead of doing the computation at full dimensionality, computes in a lower dimensional subspace. Performing the computation in a lower dimensional subspace that is a result of a linear projection from the original high dimensional space will make the assumption of normality better grounded in reality, giving a better parameter estimation, and better classification accuracy.

A preprocessing method of high dimensional data based on such characteristics has been developed based on a technique called Projection Pursuit. The preprocessing method is called Parametric Projection Pursuit [Jimenez and Landgrebe IGARSS 95] [Jimenez and Landgrebe SMC 95].

Parametric Projection Pursuit reduces the dimensionality of the data maintaining as much information as possible by optimizing a Projection Index that is a measure of separability. The projection index that is used is the minimum Bhattacharyya distance among the classes, taking in consideration first and second order characteristics. The calculation is performed in the lower dimensional subspace where the data is to be projected. Such preprocessing is used before a feature extraction algorithm and classification process, as shown in Figure 1.

In Figure 1 the different feature spaces have been named with Greek letters in order to avoid confusion. Φ is the original high dimensional space. Γ is the subspace resulting from a class-conditional linear projection from Φ using a preprocessing algorithm, e.g. Parametric Projection Pursuit. Y is the result of a feature extraction method. Y could be projected directly from Φ or, if preprocessing is used, it is projected from 17. Finally Ω is a one dimensional space that is a result of classification of data from Y space. Note that the three procedures, preprocessing, feature extraction and classification use labeled samples as a priori information.

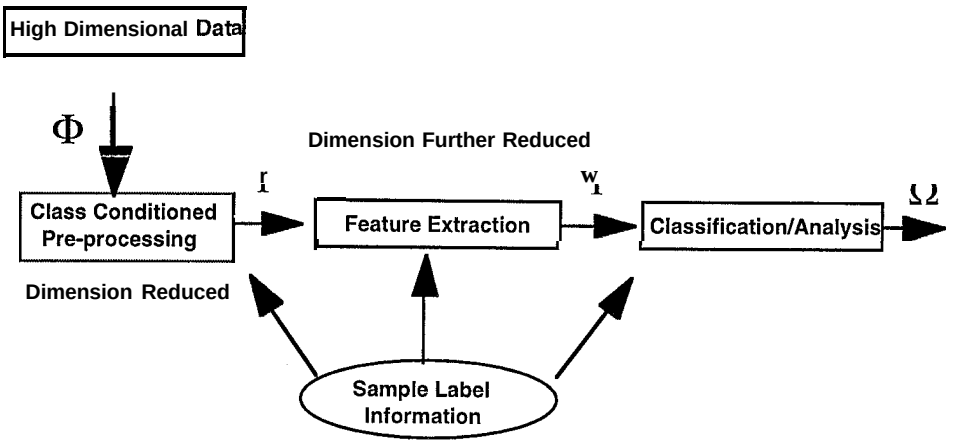

Figure 1. Classification of high dimensional data including preprocessing

IV. Experiment

In order to see the relevance of high dimensional geometrical and statistical properties for high dimensional data analysis purposes two experiments were designed. In both experiments a comparison is provided between high dimensional feature extraction and the method that uses a Parametric Projection Pursuit based preprocessing to reduce the dimensionality before a feature extraction method is used. The multispectral data used in these experiments are a segment of AVIRIS data taken of NW Indiana's Indian Pine test site. From the original 220 spectral channels 200 were used, discarding the atmospheric absorption bands.

The classification task for several classes in this and the next experiment are particularly difficult ones. The data were collected early in the growing season when the canopy of both corn and soybeans covered only about 5% of the area, There were three levels of tillage, no till in which there would be a great deal of residue on the soil surface from last year's crop, minimum till leaving a moderate amount of residue, and clean till for which there would be little or no residue. Add to this the normal amount of spectral variability due to the varying soil types present in the fields. Thus the 95% background would be highly variable, as compared to the relatively small difference in spectral response between corn and soybeans.

In this experiment four classes were defined: corn, corn-notill, soybean-rein, soybean-notill. The total number of training samples is 179 (less than the number of bands used) and the total number of test samples is 3501. Observe that this is an extreme case that is used to show the potentials of Parametric Projection Pursuit. Two types of dimensional reduction algorithms were used. The first is Discriminant Analysis (DA 200-3) that reduces the dimensionality from 200 to 3. It directly projects the data from Φ space to Y subspace. In the second method Parametric Projection Pursuit was used to reduce the dimensionality from 200 to 22. It projected the data from the Φ space to the Γ subspace. After that preprocessing method was used, Discriminant Analysis was used (PPDA 200-3) in order to linearly project the data from the Γ subspace to the Ψ subspace. As mentioned before, this has the advantage of doing the computation with the same number of training samples but at lower dimensionality. In both cases the best three features were used for classification purposes.

Four types of classifiers were used. The first one is ML classifier, the second is ML with 2% threshold. The third classifier is a spectral-spatial classifier named ECHO [Kettig & Landgrebe 1976] [Landgrebe 1980] and the fourth is ECHO with a 2% threshold. In the second and the fourth, a threshold was applied to the standard classifiers whereby in case of true normal

distributions of the data, 2% of the least likely points will be thresholded. These 2% thresholds provide one indication of how well the data fit the normal model.

The results are shown in Figure 2. Parametric Projection Pursuit followed by Discriminant Analysis at lower dimensionality performed substantially better than using Discriminant Analysis at fill dimensionality. The application of a threshold to Discriminant Analysis at full dimensionality reduced its classification accuracy more severely than when a threshold was applied in the case where Projection Pursuit was first applied, followed by Discriminant Analysis at lower dimensionality. This is due to Parametric Projection Pursuit preprocessing being better fitted to the assumption of normality.

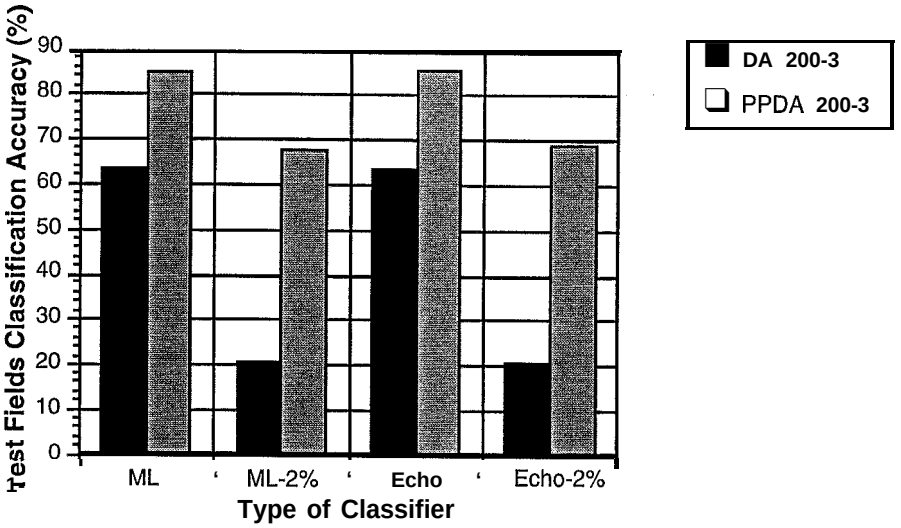

Figure 2. Test fields classification accuracy for two feature extraction methods and four classifiers.

Observe how significantly the performance of classifiers with 2% thresholds improves when using Parametric Projection Pursuit. The reason is that making the computation at low dimensional space, Γ, the assumption of normality has greater validity. In the case of having less samples and classes Discriminant Analysis will be significantly affected by the high dimensional geometrical and statistical characteristics. The next experiment will show this difficulty.

VI. Conclusion

In this section we will consider some implications of what has been discussed for supervised classification. In terms of parameter estimation, a large number of samples are required to make a given estimation in multispectral data to adequate precision. In a nonparametric approach, the number of samples required to satisfactorily estimate the density is even greater. Both kinds of estimations confront the problem of high dimensional space characteristics. As a consequence, it is desirable to project the data to a lower dimensional space where high-dimensional geometric characteristics and the Hughes phenomena are reduced. Commonly used techniques such as Principal Components, Discriminant Analysis, and Decision Boundary Feature Extraction have the disadvantage of requiring computations at full dimensionality in which the required number of labeled samples is very large. The procedures use estimated statistics that are not necessarily accurate. Another problem is the assumption of normality. Nothing guarantees that at full dimensionality, that model fits well.

It has been shown that high dimensional spaces are mostly empty, indicating that the data structures involved exist primarily in a subspace. The problem is which subspace it is to be

found in is situation-specific. Thus the goal is to reduce the dimensionality of the data to the right subspace without losing separability information. The approach is to make the computations in a lower dimensional space, i.e. in Γ instead of Φ, where the projected data produce a maximally separable structure and which, in turn, avoids the problem of dimensionality in the face of the limited number of training samples. Further, a linear projection to a lower dimensional subspace will make the assumption of normality in the Γ subspace more suitable than in the original Φ. In such a lower dimensional subspace any method used for feature extraction could be used before a final classification of data, even those that have the assumption of normality.

References

Chulhee Lee and David A. Landgrebe, "Feature Extraction Based On Decision Boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol 15, No. 4, April 1993, p p 388-400.

Chulhee Lee and David A. Landgrebe, "Analyzing High Dimensional Multispectral Data," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 31, No, 4, pp 792-800, July, 1993.

Diaconis, P. , Freedman, D. "Asymptotic of Graphical Projection Pursuit." The Annals of Statistics Vol 12, No 3 (1984): pp 793-815.

Fukunaga, K. "Introduction to Statistical Pattern Recognition," San Diego, California, Academic Press, Inc., 1990.

Hall, P., Li, K. "On Almost Linearity Of Low Dimensional Projections From High Dimensional Data." The Annals of Statistics, Vol. 21, No. 2 (1993): pp 867-889.

Hughes, G. F., "On the mean accuracy of statistical pattern recognizes," IEEE Transactions on Information Theory, Vol. IT-14, No. 1, January 1968.

Hwang, J., Lay, S., Lippman, A., "Nonparametric Multivariate Density Estimation: A Comparative Study.", IEEE Transactions on Signal Processing, Vol. 42, No. 10, 1994, p p 2795-2810.

Jimenez, L., Landgrebe, D., "Projection Pursuit For High Dimensional Feature Reduction: Parallel And Sequential Approaches," presented at the International Geoscience and Remote Sensing Symposium (IGARSS'95), Florence Italy, July 10-14, 1995.

Jimenez, L., Landgrebe, D., "Projection Pursuit in High Dimensional Data Reduction: Initial Conditions, Feature Selection and the Assumption of Normality", To be presented at IEEE International Conference on Systems, Man and Cybernetics (SMC 95), Vancouver Canada, October 22-25, 1995.

R. L. Kettig and D. A. Landgrebe, "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," *IEEE Transactions on Geoscience Electronics,* Volume GE-14, No. 1, pp. 19-26, January 1976.

D.A. Landgrebe, "The Development of a Spectral-Spatial Classifier for Earth Observational Data," *Pattern Recognition,* Vol. 12, No. 3, pp. 165-175,1980.

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