LARS Technical Memorandum T-13 040473

AN INVESTIGATION OF CLASSIFICATION RESULTS BASED ON USE OF DIRECTION COMPONENTS

By

Philip H. Swain

An Investigation of Classification Results Based on Use of Direction Components

It has been suggested (for example, by work done at Penn State) that multispectral data might profitably be represented and analyzed in polar coordinate form. An alternative but related approach is to use vector components giving (1) distance of a point from the origin, and (2) the direction components of the points. (Refer to Kaplan: Advanced Calculus, pp 9-12.) Specifically, the components are given by:

$$y_1 = \left[\sum_{j=1}^{n} x_j^2\right]^{1/2}$$

$$y_i = x_i / y_1, \quad i = 2, 3, \ldots, n$$

where x_j is the $j^{\underline{th}}$ (out of n) component of the original data and y_j is the $j^{\underline{th}}$ (out of n) component of the "new" data. For convenience, we will refer to the y-form as the "direction components" representation of the data.

For example, for 3-dimensional data, we would have

$$y_1 = \sqrt{x_1^2 + x_2^2 + x_3^2}; \quad y_2 = \frac{x_2}{\sqrt{x_1^2 + x_2^2 + x_3^2}}$$

$$y_3 = \frac{x_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}}$$

Note that since the original data vector is n-dimensional, it is of no value to add a component $y_4 = x_1/y_1$ since we would then have a linearly dependent set of components. Furthermore, replacing the given y_1 by the quotient x_1/y_1 would cause loss of information - the distance of the point from the origin.

The rationale for this approach is as follows. There is some reason to believe that to a considerable extent our data classes are really characterized by their directions relative to the coordinate axes much more than by their distance from the origin. This is because distance from the origin is a measure of response intensity whereas the direction characterizes the spectral character of the response. (See, for example, the IEEE Proceedings article by Holmes and MacDonald.)

It is proposed to investigate this question by means of the following steps:

1. Write a computer program to transform data from rectangular coordinates to direction components. The program will create a new data tape with the direction components appropriately scaled.

2. Use LARSYS to classify the new data and compare the results with classifications obtained from conventional data. Clustering, feature selection, etc., should be used in the conventional manner on both sets.

Suggestion: In creating the new tape, define the new data coordinates as follows:

$$y_i = \frac{x_i}{\left[\sum_{j=1}^{n} x_j^2\right]^2} \quad i = 1, 2, \ldots, n$$

$$y_{n+1} = \left|\sum_{j=1}^{n} x_j^2\right|^{1/2}$$

so that it will be possible to study the effects of both using and ignoring the component representing "intensity." Care must be exercised, however, that at most n of the available n+1 coordinates are used at once for any stage of the analysis.

Related questions: Is the multivariate Gaussian assumption reasonable for the transformed data?

How do linear classifiers perform on this data?