LARS Technical Note 042673

TWO EFFECTIVE FEATURE SELECTION CRITERIA FOR MULTISPECTRAL REMOTE SENSING*

P. H. Swain and R. C. King
Laboratory for Applications of Remote Sensing
Purdue University
W. Lafayette, Indiana 47907

Presented at The International Joint Conference on Pattern Recognition,
Washington, D.C.,
November 1973

Summary

In an earlier study, Swain et al.[1] reported on two statistical separability measures which for multiclass feature selection were shown experimentally to be more reliable than divergence. However, the empirical results of that study together with the best theoretical results in the literature left open some practical questions regarding the quantitative characterization of these separability measures. This paper is concerned with an empirical study aimed at answering such questions. It has been possible to further substantiate that the Jeffreys-Matusita Distance and a saturating transform of divergence are effective feature selection criteria for remote sensing applications. In fact, an explanation as to why this should be the case has now been made apparent.

Introduction

The feature selection problem in pattern recognition may be stated as follows: Given a set of N features (e.g., measurements on an object to be classified), find the beat subset consisting of k features to be used for classification. Usually the objective is to optimize a trade-off between classification accuracy (which is generally reduced when fewer than the N available features are used) and computational speed and cost (fewer features require fewer computations and hence less time).

Ideally this problem would be solved by computing the probability of classification error associated with each k-feature subset and then selecting the subset yielding minimal error. However, it is generally not practical to perform the required computations; even under the simplifying assumption of Gaussian statistics, the numerical integration required to compute the errors is impractical to carry out. Alternative methods have therefore been sought for feature selection.

An approach which has been widely investigated[2] depends on the concept of a measure of "statistical distance" between the probability densities characterizing the pattern classes. Intuitively one would like to have a distance measure with the property that if the distance between two class densities were greater for feature set α than for feature set β, then the error probability obtained for set α would be less than for set β.

Unfortunately, none of the distance measures which have been proposed can be shown to have This property exactly. However several, including the distance measures discussed 2 herein, have the following weaker property

For feature sets α and β and distance measure $d(*)$, if $d(\alpha) > d(\beta)$ then there exists a set of prior probabilities π for the pattern classes such that

$$P_e(\alpha,\pi) < P_e(\beta,\pi) \qquad (1)$$

where $P_e(\alpha,\pi)$ is the probability of error associated with feature set α under the assumption of prior probability set π; similarly

*This research was supported by NASA Grant NGL 15-005-112.

for $P_e(\beta,\pi)$. Distance measures having this property have been found quite useful for feature selection.

Candidate Distance measures

Divergence is a distance measure long ago proposed for this purpose.[3,4] The divergence D for two densities $p_l(x)$ and $p_2(x)$ is defined as

$$D = \int_X [p_1(x) - p_2(x)] \log_e \frac{p_1(x)}{p_2(x)} dx \qquad (2)$$

where the integral is taken over the entire feature space. If the $p_i(x)$, i = 1, 2, are multivariate Gaussian densities with U_i and covariance matrices $\sum_i$ then

$$D = \frac{1}{2} \mathrm{tr}\left[\Sigma_1 - \Sigma_2\right]\left[\Sigma_2^{-1} - \Sigma_1^{-1}\right] + \frac{1}{2} \mathrm{tr}\left[\Sigma_1^{-1} + \Sigma_2^{-1}\right]\left[U_1 - U_2\right]\left[U_1 - U_2\right]^T \qquad (3)$$

where trA denotes trace of matrix A, A^{-1} is the inverse of A, and A^T is the transpose of A.

Although divergence only provides a measure of the distance between two class densities, its use is extended to the multiclass case by taking the average over all class pairs.4 If D_{ij} is the divergence between classes i and j, then the multiclass feature selection criterion is

$$D_{AVE} = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_{ij} \qquad (4)$$

Even for the 2-class case, the relationship between divergence and classification accuracy is highly nonlinear (in fact, divergence increases without bound as class separability increases, whereas probability of correct classification must "saturate" at 100 percent), and it is found that widely separable classes make too much of a contribution to D_{AVE} as compared with less separable classes. As a result, in problems involving a wide range of class separabilities, D_{AVE} is not a reliable criterion for feature selection.

Swain et al.[1] have shown experimentally that a separability measure referred to as the Jeffreys-Matusita Distance (JM-distance)[+] provides a much more reliable criterion, presumably because as a function of class separability it behaves much more like probability of correct classification. For two densities $p_l(x)$ and $p_2(x)$, the JM-distance J is given by

$$J = \int_X \left[\sqrt{p_1(x)} - \sqrt{p_2(x)}\right]^2 dx \qquad (5)$$

which can also be written in the form J=2(1-p) where

$$p = \int_X \sqrt{p_1(x) p_2(x)} dx \qquad (6)$$

If $p_i(x)$, i = 1, 2, are multivariate Gaussian densities as above then $\rho=e^{-\alpha}$ and $J=2(1-e^{-\alpha})$, where

$$\alpha = \frac{1}{8}\left(U_1 - U_2\right)^T \Sigma^{-1} \left(U_1 - U_2\right) + \frac{1}{2} \log_e \left[\frac{\det \Sigma}{\det \Sigma_1 \cdot \det \Sigma_2}\right] \qquad (7)$$

and

$$\Sigma = \frac{1}{2}\left[\Sigma_1 + \Sigma_2\right]$$

In Eq. 7, det A means the determinant of matrix A. Since $0<e^{-\alpha}<1$, J ranges from 0 to 2 with 2 corresponding to the largest separation. It was observed by Swain et al. that this "saturating" behavior of J is responsible for its utility as a feature selection criterion in multiclass problems. If Jij is the JM-distance between classes i and j, then the multiclass feature selection criterion is taken as

$$J_{AVE} = \frac{1}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} J_{ij} \qquad (8)$$

[+] In Swain et al.1 this is referred to as B-distance because of its relationship to the Bhattacharyya coefficient. However other works have attributed this distance measure to Jeffreys and Matusita.

Since in this case Jij increases with separability in a saturating fashion (rather than in an unbounded fashion as does divergence), widely separable classes do not make an undue contribution to the average separability criterion. Thus the average separability criterion better reflects the overall classification accuracy which will be attained.

In some respects, divergence is easier to compute than JM-distance, however, so that the following observation is of interest: JM-distance and divergence are related by the inequality

$$J \leq 2\left[1 - \exp\left(-\frac{D}{8}\right)\right] \qquad (9)$$

(again assuming Gaussian statistics). This relation provides the motivation for defining a "saturating transform" of divergence, DT, as

$$D_T = 2\left[1 - \exp\left(-\frac{D}{8}\right)\right]. \qquad (10)$$

The transformed divergence varies from D_T=0 (when D=0) to D_T=2 (when D–>∞). Thus DT is another candidate of potential utility for averaging in the multiclass case.

Experimental Investigations

Experiments with remote sensing data collected by airborne multispectral scanner systems have demonstrated that the saturating measures of separability are preferable to ordinary divergence for multiclass feature selection.1 In fact, the JM-distance yields much more reliable results, with the transformed divergence running a close second.

However, these experimental results taken together with the best available results from the literature (both theoretical and experimental) leave open some nagging questions; to wit:

1. Is it possible to characterize the relations between the separability measures discussed above and classification error in a manner which more fully explains the superior performance of the saturating separability measures?

2. Given that at best only inequality relationships between the separability measures and classification error are available (i.e. upper and/or lower bounds on error for a given separability), what is the nature of the distribution of performance for a given separability?

To derive answers to these questions in the remote sensing context, an experiment was performed on a digital computer'. Based on typical second order statistics derived from real remote sensing data, 2790 sets of Gaussianly distributed artificial data were generated; each set contained 1000 observations for each of two pattern classes in a feature space of dimensionality ranging from 1 to 6 (465 sets were generated for each dimension 1,2,...,6). For each set the divergence, transformed divergence, and JM-distance were computed, and the actual classification error for the 2000 observations was taken as the associated probability of error. The results are summarized in Figures 1 through 5, with P_c, the probability of correct classification (1.0 - probability of error), plotted against the respective distance measures. Figures 1, 2 and 3 show the superimposed results for all 2790 data sets. Also shown are least-squares polynomial approximations (of degree 3 for J and D_T and degree 10 for D), and the theoretically derived bounds[2] on performance as functions of separability (a lower bound is available only for J). Clearly, the relationship between probability of correct classification and the measure of separability is nonlinear in each case. But for the range of classification accuracy likely to be encountered in real problems -- say, 75 percent to 100 percent -- divergence increases much more than linearly with separability, which is precisely why well separated classes have too much influence on average divergence in the multiclass case. The other two separability measures are, in this sense, much more "well-behaved. "

For historical reasons plus the fact that transformed divergence is slightly easier to calculate than JM-distance (one less matrix

inversion is required for each feature combination evaluated), our work has been mostly directed toward the former. As mentioned earlier, however, our experiments have shown the JM-distance to have a small edge over transformed divergence with respect to accurately predicting the best features for multiclass recognition. Figures 1 and 2 provide additional reasons for preferring JM-distance. One reason is the lower bound on classification accuracy as a function of JM-distance; no such lower bound is available for transformed divergence. Another reason is the much tighter clustering of the experimental results about the regression curve for the case of JM-distance so that performance is "much more predictable" as a function of JM-distance.

Figures 4 and 5 show how the experimental results vary as a function of the number of features for the case of transformed divergence. The principal effect is the increasing difficulty of obtaining observations at the lower end of the scale as more features are added. Compounded with this, there may be a tendency for the results to cluster somewhat more tightly about the regression curve. Practically speaking, however, the same regression curve can be used for 1,2,3,4,5, or 6 features to approximate the functional relationship between classification accuracy and transformed divergence (a similar conclusion applies for the JM-distance).

Concluding Remarks

One cannot help but notice in Figures 1 and 2 the apparent looseness of the theoretical bounds relative to the experimental observations. This almost certainly is at least in part the result of characteristics of the remote sensing data which were the basis for the artificial data generation. We leave it to anyone needing a more general result to repeat the experiment with completely random selection of means and covariance matrices.

As another extension of this experiment, the multiclass case may be investigated explicitly in order to verify the inferences we have drawn relative to the multiclass case. This would involve generating a large number of randomly specified (though presumably, but not necessarily, Gaussianly distributed) multiclass data sets. Graphs similar to Figures 1 through 5 could then be produced to determine trends, empirical bounds, etc. Does anyone have an idle computer at his disposal?

In conclusion, our results have reinforced and provided an explanation for previous observations that JM- distance and a saturating transform of divergence are highly useful for feature selection in the multiclass case. Although our experiments have been limited to a very specific instance from remote sensing, our general conclusions are certainly applicable to a much wider range of problems, since they depend only on the functional behavior of the separability measures discussed.

References

1. P. H. Swain, T. V. Robertson, and A. G. Wacker, "Comparison of the Divergence and B-Distance in Feature Selection," Information Note 020871, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana 47907, February 1971.

2. T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection." IEEE Trans. on Comm. Technology, vol . 15, no. 1, Feb. 1967, pp. 52-60.

3. T. Marill and D. M. Green, "On the Effectiveness of Receptors in Recognition Systems," IEEE Trans. Information Theory, vol. IT-9, pp. 11-17, January 1963,

4. K. S. Fu and P. J. Min, "On Feature Selection in Multiclass Pattern Recognition," Tech. Rept. TR-EE68-17, School of Electrical Engineering, Purdue University, West Lafayette$ Indiana 47907, July 1968.

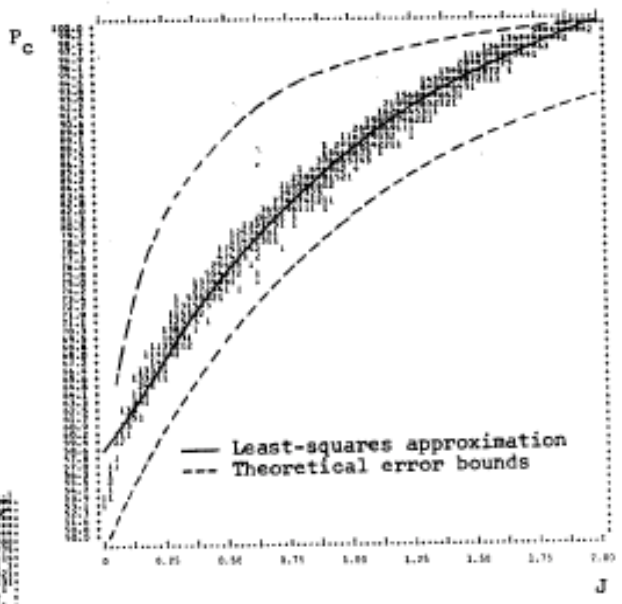

Figure 1.
Probability of Correct Classification (P_c) vs. JM-Distance (J)

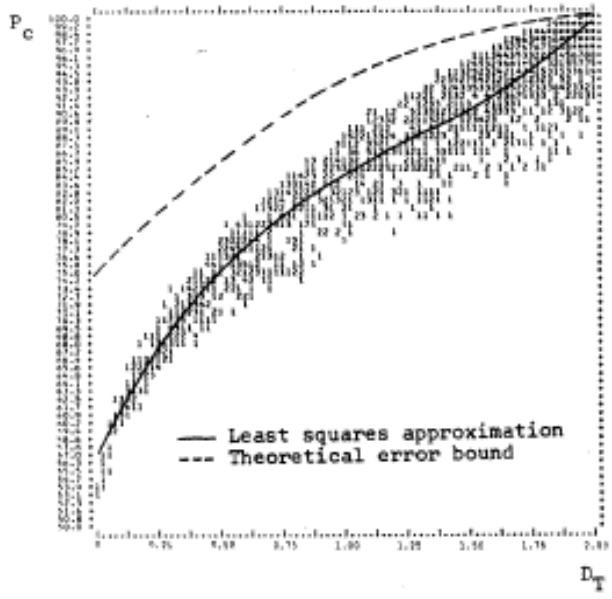

Figure 2.
Probability of Correct Classification (P_c) vs. Transformed Divergence (D_T)

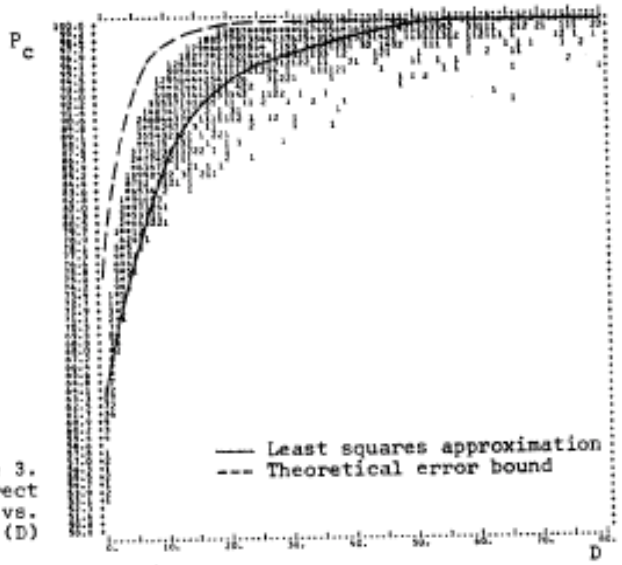

Figure 3.
Probability of Correct Classification (P_c) vs. Divergence (D)

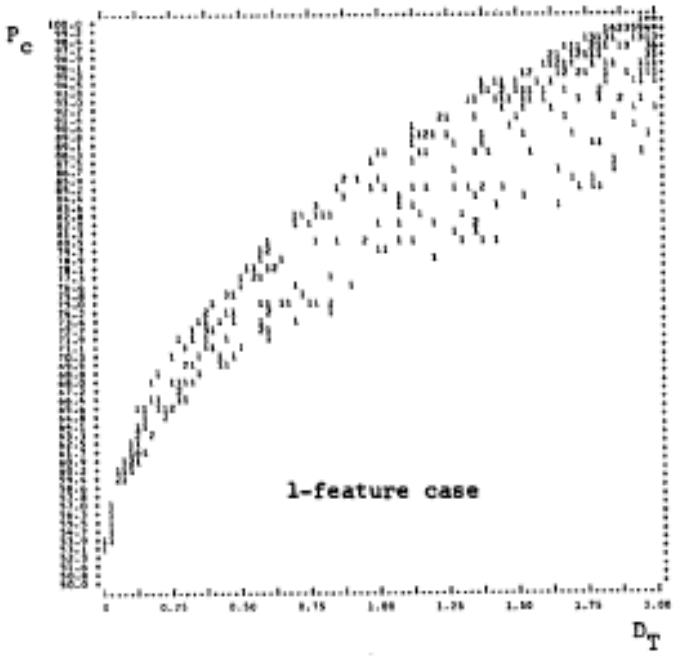

Figure 4.
Probability of Correct Classification (P_c) vs. Transformed Divergence (D_T) for One Feature

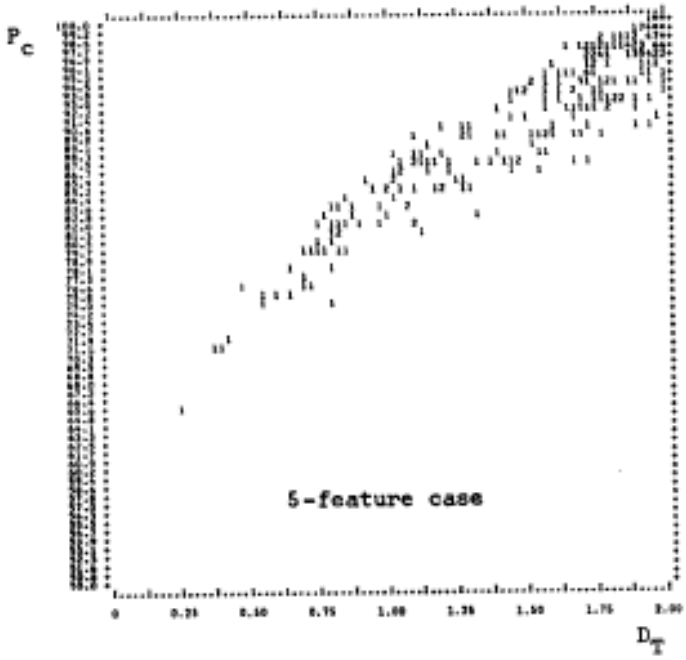

Figure 5.
Probability of Correct Classification (P_c) vs. Transformed Divergence (D_T) for Five Features