LARS Information Note 020871

# COMPARISON OF THE DIVERGENCE AND B-DISTANCE IN FEATURE SELECTION

BY

P. H. SWAIN
T. V. ROBERTSON
A. G. WACKER

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

LARS Information Note 020871

# Comparison of the Divergence and B-Distance in Feature Selection

by

P. H. Swain, T. V. Robertson, and A. G. Wacker

## Summary

The Bhattacharyya distance ("B-distance") is compared experimentally with the divergence as a criterion for feature selection in pattern recognition. The results obtained using B-distance generally approximate those obtained when the typewriter options available with LARS' divergence algorithm are used to best effect. This suggests that the B-distance offers a more automated approach to feature selection than has been available. A saturating function of the divergence is found to perform almost as well as B-distance, but is substantially more efficient in terms of the computations required.

## I. Introduction

The problem of feature selection in pattern recognition may be stated as follows:

Given N features (measurements on each pattern), select the k-feature subset (for a given k) which minimizes the probability of classification error.*

* A slightly different criterion may be used such as minimizing the "expected cost" (Bayesian criterion), but this does not significantly alter the problem since error probabilities are still involved.

It is often the case, however, that direct minimization of the error probability is not possible, either because an analytical expression for the error probability cannot be found, or because the analytical expression is too complicated for evaluation. An alternative approach which has been tried in such cases depends on the concept of a measure of "distance" between probability densities--the densities characterizing the pattern classes. If a distance measure could be found such that increasing distance between densities implies smaller classification error probability, then the best feature set could be found by simply maximizing this distance. Unfortunately, such a distance measure has not been found. However, the distance measures discussed in this report have the following property relating them to the probability of error, $P_e$ [1]:

For feature sets $\alpha$ and $\beta$, and distance measure d(.), if

$$d\ (\alpha) > d\ (\beta)$$

then there exists a set of prior class probabilities $\pi$ such that

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

## II. B-Distance

A statistical distance measure appearing in the work of Jeffreys [2] is defined for two densities $p_1(x)$ and $p_2(x)$ by

$$B \triangleq \int [\sqrt{p_1(x)} - \sqrt{p_2(x)}]^2\ dx \qquad (1)$$

For convenience, B will be referred to herein as the Bhattacharyya distance (B-distance), although this term has been used elsewhere in the literature to refer to the negative logarithm of the Bhattacharyya coefficient, i.e. $-\log_e \rho$, where

$$\rho = \int \sqrt{p_1(x)\ p_2(x)}\ dx$$

The B-distance defined by equation (1) is related to the Bhattacharya coefficient by

$$B = 2(1-\rho)$$

It can be interpreted as a distance between the two points $(\sqrt{p_1(x)})$ and $(\sqrt{p_2(x)})$ on the unit sphere in the space of x [1]. Kailath [1] gives $\rho$ for two multivariate Gaussian densities $\{p_i(x) = N(m_i, R_i), i = 1,2\}$ as

$$\rho = e^{-\sigma}$$

where

$$\sigma = \tfrac{1}{8}(m_1-m_2)^t R^{-1}(m_1-m_2) + \tfrac{1}{2}\log_e\left[\frac{\det R}{\det R_1 \cdot \det R_2^{1/2}}\right] \qquad (2)$$

and

$$2R = R_1 + R_2$$

Since $0 \leq \rho \leq 1$, B ranges from 0 to 2, with 2 being the largest separation attainable. Upper and lower bounds on error probability $P_e$ in terms of B have been found [1]. For equal prior probabilities they are

$$(2 - B)^2/_{16} \leq P_e \leq (2 - B)/_4$$

## III. Divergence

The divergence was first introduced by Jeffreys [2]. The divergence D for two densities $p_1(x)$ and $p_2(x)$ is defined as

$$D \triangleq \int [p_1(x) - p_2(x)] \ln \frac{p_1(x)}{p_2(x)}\, dx \qquad (3)$$

For $\{p_i(x) = N(m_i, R_i), i = 1,2\}$

$$D = \tfrac{1}{2}\mathrm{tr}[R_1-R_2][R_2^{-1} - R_1^{-1}] + \tfrac{1}{2}\mathrm{tr}[R_1^{-1} + R_2^{-1}][m_1-m_2][m_1-m_2]^t \qquad (4)$$

The range of D is 0 to $\infty$, with higher values implying greater separation. No upper bound on $P_e$ in terms of D is available, but a crude lower bound is [1]

$$P_e \geq \frac{1}{8}\exp(-D/2)$$

(See also [4].)

## IV. Feature Selection

The goal of feature selection is to minimize the overall probability of misclassification, which in the general multiclass case is not a simple function of the pairwise error probabilities. Even if a distance measure is an accurate representation of two-class error, it is not obvious how to extend the use of the distance measure to the multiclass case [3]. The method investigated in this study used the average pairwise distance between class densities as an indication of the overall error probability. In selecting the best k features from a group of N features, each of the possible combinations* of k features was used to construct a k-variate probability density function characterizing the m classes represented in the training data. The average of the $\binom{m}{2}$ pairwise distances between density functions was then used to rank the k-feature sets.

## V. Comparison of B-distance and Divergence

Under the assumption of Gaussian statistics, the computational complexities of the B-distance and divergence are quite different. The number of matrix inversions required by each distance measure is a reasonable criterion for comparison. To calculate all pairwise distances for a set of m classes, B-distance requires the

* The number of k-feature subsets of a set of N features is given by $\binom{N}{k} = \frac{N!}{k!(N-k)!}$

inversion of all pairwise sums of class covariance matrices plus evaluation of the determinant of each class covariance matrix (determinant evaluation is roughly equivalent to matrix inversion in terms of the amount of computation required), or a total of $(m^2 + m)/2$ inversions. Divergence requires only one inversion per class, a total of m inversions. Thus the relative computational cost of using B-distance rather than divergence grows rapidly as the number of classes increases.

The functional behavior of the two distance measures is also quite different. Consider, for example, that B-distance and divergence are related by the inequality

$$B \leq 2[1-\exp(-D/8)] \tag{5}$$

(again assuming Gaussian statistics). D is zero for identical statistics and increases without bound as the disparity between class statistics increases. B is also zero for identical statistics, but saturates (approaches the value 2.0 asymptotically) as the class statistics become more unlike. Significantly, the latter functional behavior is more like the behavior of the probability of correct classification than is the behavior of the divergence. This is of little importance when only two classes are considered, but it becomes very important in the general m-class case. In particular, the saturation effect of the B-distance tends to prevent widely separated class pairs from having overly large influence on the average pairwise distance, the criterion used to rank the feature subsets. Thus, the B-distance may provide a more reliable criterion for automatic feature selection. The experimental results presented below tend to support this conclusion.

In view of the high computation cost of using the B-distance rather than divergence, one is led to speculate as to whether some saturating function of D might have the desirable properties of B. The relationship (5) suggests such a function. The "transformed divergence" $D_T$, defined by

$$D_T \triangleq 2[1-\exp(-D/8)]$$

has been included in the experimental work discussed in the following section.

## VI. Experimental Results

Some experiments were performed to compare the effectiveness of feature selection algorithms based on the B-distance, divergence, and transformed divergence. The data consisted of two 12-channel flightlines.

Run 66000600: Purdue Flightline C1 (6/66)
Run 69004900: Purdue Flightline PF24 (8/69)

In Run 66000600 eleven classes were considered; 14 classes were considered in Run 69004900. The class probability densities were assumed to be multivariate Gaussian, characterized by the sample means and covariance matrices. Sets of 2, 3, and 4 features from both runs were ordered according

* This is accomplished in LARS' feature selection processor by means of the MAX option which allows the user to provide an artificial upper bound on the pairwise divergence.

to the three feature selection criteria. A selection of the best feature sets from these orderings were then used to classify the data originally used to calculate the class statistics.* The results are shown in the following tables in which the feature sets used for classification are ranked by actual classification performance and the rankings predicted by the feature selection criteria are listed for comparison.

In obtaining the results shown, none of the options in the LARS feature selection processor were invoked (such as the MAX option mentioned in an earlier footnote) so that the performance of the distance measures _per se_ could be evaluated. It was found that when the options _were_ used, all three measures yielded generally similar results.

Ideally, one would like a feature selection criterion to predict the same ranking as that arrived at by classification performance. Let i be the classification ranking and let p(i) be the ranking prediction; then a performance measure for comparing the feature selection criteria is given by:

$$M = \frac{1}{n}\sum_{i=1}^{n} \text{Max } \left[\frac{i}{p(i)}, \frac{p(i)}{i}\right]$$

where n is the number of k-feature subsets in the set of N features. The most effective feature selection criterion is

* Only "training samples" were used in the classification. For this experiment, we were looking for the "best" separability measure for distributions assumed _a priori_ well characterized by the statistics. Thus the use of test samples distinct from the data used to compute the statistics was not appropriate.

| Features | Overall % Correct | B rank | $D_T$ rank | D rank |
|---|---|---|---|---|
| 1,10 | 81.9 | 2 | 5 | * |
| 1,9 | 81.3 | 1 | 2 | 8 |
| 1,8 | 79.0 | 4 | 7 | 14 |
| 8,11 | 75.8 | 7 | 6 | 4 |
| 8,12 | 75.8 | 3 | 1 | 3 |
| 9,12 | 75.0 | 5 | 4 | 1 |
| 9,11 | 74.0 | * | 10 | 2 |
| 10,12 | 70.3 | 6 | 3 | 5 |
| 1,9,12 | 91.4 | 4 | 7 | 2 |
| 6,10,12 | 91.2 | 1 | 1 | * |
| 6,9,12 | 91.0 | 8 | 5 | 5 |
| 1,10,12 | 90.5 | 3 | 3 | * |
| 6,10,11 | 90.4 | 2 | 2 | * |
| 8,11,12 | 79.3 | * | * | 3 |
| 9,11,12 | 79.1 | * | * | 1 |
| 1,6,10,12 | 94.6 | 1 | 1 | * |
| 1,6,10,11 | 94.1 | 2 | 2 | * |
| 1,6,9,12 | 93.7 | 4 | 8 | 23 |
| 6,9,10,12 | 93.5 | 9 | 6 | 24 |
| 6,8,10,12 | 92.9 | 12 | 9 | * |
| 6,10,11,12 | 91.7 | * | 26 | 13 |
| 1,9,11,12 | 91.7 | * | * | 1 |
| 8,9,11,12 | 84.9 | * | * | 2 |

Run 66000600

Feature Ranking by B, $D_T$, and D

* Not ranked in top 30.

| Features | Overall % Correct | B rank | $D_T$ rank | D rank |
|---|---|---|---|---|
| 10,11 | 86.1 | 1 | 1 | 1 |
| 10,12 | 81.9 | 3 | 4 | 6 |
| 7,11 | 79.7 | 9 | 9 | 12 |
| 3,10 | 79.0 | 2 | 3 | 4 |
| 7,10 | 78.4 | 5 | 6 | 2 |
| 9,11 | 76.4 | 11 | 10 | 11 |
| 4,10,11 | 95.0 | 1 | 1 | 4 |
| 5,10,11 | 94.8 | 4 | 5 | 5 |
| 3,10,11 | 94.8 | 2 | 2 | 3 |
| 6,10,11 | 94.2 | 3 | 3 | 2 |
| 7,10,11 | 93.4 | 5 | 4 | 1 |
| 6,9,11 | 91.0 | 17 | 8 | * |
| 4,7,10,11 | 96.5 | 1 | 1 | 2 |
| 4,8,10,11 | 96.5 | 4 | 5 | 22 |
| 3,7,10,11 | 96.4 | 2 | 2 | 3 |
| 4,6,10,11 | 96.0 | 3 | 3 | 11 |
| 3,6,10,11 | 96.0 | 5 | 4 | 12 |
| 5,6,10,11 | 95.7 | 11 | 10 | 17 |
| 7,9,10,11 | 94.7 | * | * | 1 |

Run 69004900

Feature Ranking by B, $D_T$, and D

*Not ranked in top 30.

then the one for which M is minimum (M = 1 is perfect prediction). Applying this performance measure to the results shown in the tables* and using p(i) = 30 where * appears in the tables, the following results were obtained (averages have been taken over the 2, 3, and 4 feature cases):

| Run 66000600 | Run 69004900 |
|---|---|
| $M_B = 1.95$ | $M_B = 1.48$ |
| $M_{D_T} = 2.49$ | $M_{D_T} = 1.58$ |
| $M_D = 9.23$ | $M_D = 3.01$ |

The relative times required to rank all 4-feature combinations out of 12 features, assuming 11 classes, were as follows:

| | |
|---|---|
| Divergence | 3.29 |
| Transformed divergence | 3.35 |
| B-distance | 5.99 |

For this case, B-distance required 66 matrix inversions compared with 11 for divergence and 11 for transformed divergence.

## VII. Conclusions

The experimental results show that both the B-distance and transformed divergence are superior distance measures as compared with divergence when average pairwise distance is used as the criterion for feature selection. Although

* Only the best 5 feature sets, based on classification results, were used in each case.

the divergence aided by LARS' typewriter options can provide results similar to those obtained by the other distance measures, both the B-distance and transformed divergence offer a more automatic mode of operation. In view of the relative computational efficiency of the transformed divergence, this is probably the most practical distance measure to implement.

1. T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection, *"I.E.E.E. Trans. on Communication Technology*, Vol. COM-15, pp. 52-60, February 1967.

2. H. Jeffreys, *Theory of Probability*. Oxford University Press, 1948.

3. K.S. Fu and P.J. Min, "On Feature Selection in Multiclass Pattern Recognition," Tech. Rept. No. TR-EE68-17, School of Electrical Engineering, Purdue University, Lafayette, Indiana, July 1968.

4. T. Marill and D.M. Green, "On the Effectiveness of Receptors in Recognition Systems," *I.E.E.E. Trans. on Information Theory*, Vol. IT-9, pp. 11-17, January 1963.