On the Evaluation of Independent Binary Features

ROBERT P. W. DUIN, CHRIS E. VAN HAERSMA BUMA, AND
LUITZEN ROOSMA

***Abstract*—For the case of independent binary features, conditions under which the addition of a new feature does not decrease the Bayes error are derived. These conditions lead to illustrations of families of distributions for which the best two independent measurements are not the two best.**

I. INTRODUCTION

We consider the problem of classifying the K-dimensional binary vector $\mathbf{x} = (x^1, x^2, \cdots, x^K)$ into one of the two classes A and B. The features are assumed to be statistically independent for both classes so that the probability distribution of $\mathbf{x}$, given class i, can be written

$$F_i(\mathbf{x}) = \prod_{j=1}^{K} \{p_i^j x^j + (1 - p_i^j)(1 - x^j)\}, \tag{1}$$

where $i = A,B$ and $p_i^j = \text{Prob}\,(x^j = 1 | \mathbf{x} \in \text{class } i)$. The Bayes error ϵ made by using (1) for classification is

$$\epsilon = \sum_{\mathbf{x}} \min \{cF_A(\mathbf{x}), (1 - c)F_B(\mathbf{x})\} \tag{2}$$

in which c is the *a priori* probability for class A. The main purpose of this note is to investigate the effect upon ϵ of the addition of a $K + 1$st feature. Conditions under which ϵ does not decrease will be given, that is, cases in which the addition of a new feature does not result in an improvement of the probability of classification.

The Bayes error (2) can be expressed in the contributions $\epsilon_{\mathbf{x}}$ of all points $\mathbf{x}$ by

$$\epsilon = \sum_{\mathbf{x}} \epsilon_{\mathbf{x}} F(\mathbf{x}), \tag{3}$$

where $F(\mathbf{x}) = cF_A(\mathbf{x}) + (1 - c)F_B(\mathbf{x})$ is the probability of $\mathbf{x}$. We will write the error ϵ', if the dimensionality is raised from K to $K + 1$, as a sum over all points $\mathbf{x}$ of the K-dimensional space, let us say

$$\epsilon' = \sum_{\mathbf{x}} \epsilon'_{\mathbf{x}} F(\mathbf{x}), \tag{4}$$

where $\epsilon'_{\mathbf{x}}$ can be interpreted as the probability of error for a given K-dimensional point $\mathbf{x}$ if the additional $K + 1$st feature is used. The probabilities ϵ and ϵ' will be compared by comparing $\epsilon_{\mathbf{x}}$ and $\epsilon'_{\mathbf{x}}$ for all points $\mathbf{x}$. Let

$$\alpha(\mathbf{x}) = (1 - c)F_B(\mathbf{x})/\{cF_A(\mathbf{x})\} \tag{5}$$

be the probability ratio of class B to class A for a given point $\mathbf{x}$. It can be shown (see [5]) that $\epsilon'_{\mathbf{x}} = \epsilon_{\mathbf{x}}$ when both p_A^{K+1}/p_B^{K+1} and $(1 - p_A^{K+1})/(1 - p_B^{K+1})$ are either simultaneously larger than $\alpha(\mathbf{x})$ or smaller than $\alpha(\mathbf{x})$. If this is valid for all $\mathbf{x}$, then $\epsilon' = \epsilon$, and the addition of the new feature give no improvement. When features with probabilities p_A^{K+1} and p_B^{K+1} are plotted in a (p_A, p_B) plane, there is an area (shaded in Fig. 1) where these conditions apply simultaneously. For the proof, see [4]. A feature in the shaded area of Fig. 1 therefore gives no improvement when it is added to the feature set. The important thing to note is that such a feature is not necessarily a feature such that $p_A^{K+1} = p_B^{K+1}$.

Manuscript received July 1, 1975; revised April 26, 1977.

R. P. W. Duin and L. Roosma are with the Department of Applied Physics, Delft University of Technology, Delft, The Netherlands.

C. E. van Haersma Buma is with Philips Audio Division, Eindhoven, The Netherlands.

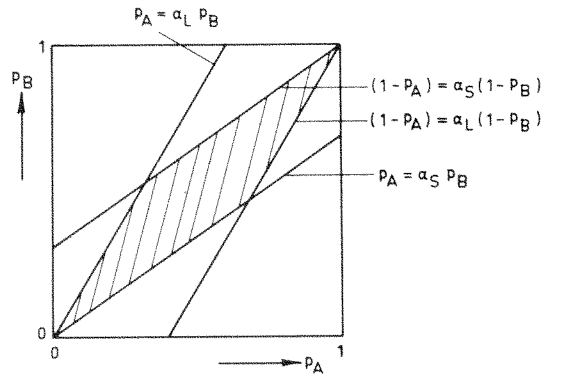


Fig. 1. For features in the shaded area, $\epsilon' = \epsilon$ is valid because $\epsilon_{\mathbf{x}}' = \epsilon_{\mathbf{x}}$, for all $\mathbf{x}$. α_l: largest α smaller than one. α_s: smallest α larger than one.

II. Results for Identically Distributed Features

The above result can be simplified when the features are identically distributed. Let

$$p_A^j = p_A, \qquad p_B^j = p_B, \qquad \text{for all } j,$$

so that

$$F_i(\mathbf{x}) = (p_i)^N(1-p_i)^{K-N}, \qquad i = A,B \tag{6}$$

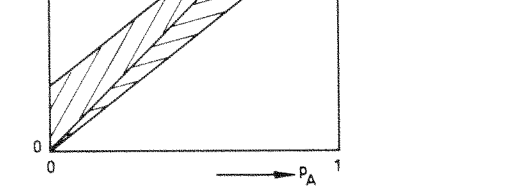
where N is the number of ones and $K-N$ is the number of zeros in $\mathbf{x}$. It can be easily shown (see [5]) that no improvement is reached by the addition of a new feature if one of the following conditions holds.

1) The point (p_A,p_B) is in the shaded area of Fig. 2, an area that grows smaller with increasing K. Features within this area give an error of $\epsilon = \min\{c, 1-c\}$.

2) There exists a point $\mathbf{x}$ with the following number of ones:

$$N = \left\{\ln\left(\frac{c}{1-c}\right) - \ln\left(\frac{p_A}{p_B}\right) + K\ln\left(\frac{1-p_A}{1-p_B}\right)\right\}\left\{\ln\left(\frac{p_A}{p_B}\right) + \ln\left(\frac{1-p_A}{1-p_B}\right)\right\}^{-1}. \tag{7}$$

This condition is fulfilled for those combinations of c, p_A, p_B, and K for which (7) is an integer. If $c = 0.5$ and $p_A = 1 - p_B$, (7) simplifies to $N = \frac{1}{2}(K-1)$, which is an integer for K odd. The error will then decrease only for an even number of features. This result is the same as that for the error probability of a binary symmetric channel as a function of the number of repetitions of the message.



$1-p_B = \left(\frac{c}{1-c}\right)^{\frac{k+1}{2}}(1-p_A)$
$p_B = \left(\frac{c}{1-c}\right)^{\frac{k+1}{2}} p_A$
p_B
p_A
0
1
0
1

Fig. 2. Features which give no improvement if K of them are available and a K + 1st one is added, $c < 0.5$.

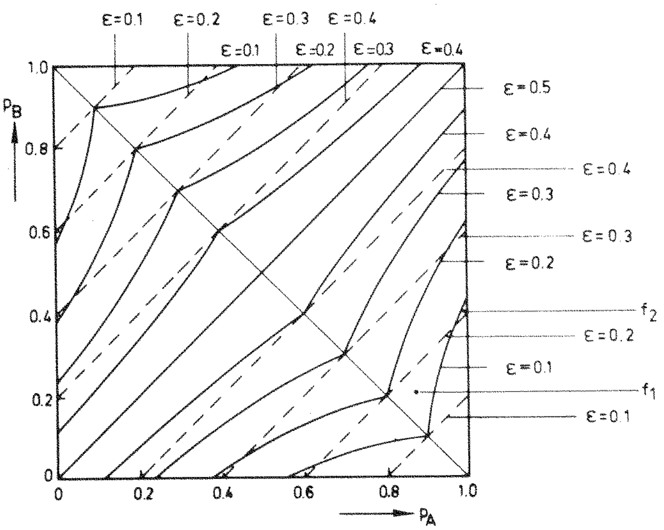


Fig. 3. Lines of constant Bayes error ϵ in the (p_A, p_B) plane for a single feature (broken lines) and for two independent identically distributed features with parameter values p_A and p_B, $c = 0.5$.

For the case of identically distributed features, it is possible to compute lines of constant Bayes error in the (p_A,p_B) plane [4]. In Fig. 3, these lines are given for $K = 1$ and $K = 2$. They illustrate the result mentioned above. From this figure, it is clear that it is possible to have two features $f_1 = (p_A^1, p_B^1)$ and $f_2 = (p_A^2, p_B^2)$ such that

$$\epsilon(f_1) < \epsilon(f_2),$$

and for $K = 2$,

$$\epsilon(f_1, f_1') > \epsilon(f_2, f_2'),$$

in which f_1' and f_2' are identical with f_1 and f_2, respectively. This shows that the two best features (f_1 and f_1') are not the best two features. Cover [3] has also given an example of this occurrence, also based upon pairs of identically distributed features. Other examples given by Elashoff [1] and Toussaint [2] cannot be interpreted in Fig. 3 because they are based upon pairs of nonidentically distributed features.

Acknowledgment

The authors wish to thank Prof. D. B. Cooper of Brown University for stimulating and fruitful discussions on the subject of this paper during his stay at Delft University, and the referees for their useful suggestions.

References

[1] J. D. Elashoff, R. M. Elashoff, and G. E. Goldman, "On the choice of variables in classification problems with dichotomous variables," *Biometrika*, vol. 54, pp. 668–670, 1967.

[2] G. T. Toussaint, "Note on optical selection of independent binary-valued features for pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-17, p. 618, Sept. 1971.

[3] T. M. Cover, "The best two independent measurements are not the two best," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-4, pp. 116–117, Jan. 1974.

[4] C. E. van Haersma Buma, "Evaluation and use of independent binary features," Thesis, Dep. Applied Physics, Delft Univ. Technology, The Netherlands, June 1973.

[5] R. P. W. Duin, C. E. van Haersma Buma, and L. Roosma, "Evaluation of independent binary features," Internal report, Pattern Recognition Group, Dep. Applied Physics, Delft Univ. Technology, The Netherlands, June 1975.