Kybernetika

Lorenzo Bruzzone; Sebastiano B. Serpico

A simple upper bound to the Bayes error probability for feature selection

*Kybernetika*, Vol. 34 (1998), No. 4, [387]--392

Persistent URL: `http://dml.cz/dmlcz/135220`

**Terms of use:**

© Institute of Information Theory and Automation AS CR, 1998

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* `http://project.dml.cz`

KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 387–392

# A SIMPLE UPPER BOUND TO THE BAYES ERROR PROBABILITY FOR FEATURE SELECTION

LORENZO BRUZZONE AND SEBASTIANO B. SERPICO[1]

In this paper, feature selection in multiclass cases for classification of remote-sensing images is addressed. A criterion based on a simple upper bound to the error probability of the Bayes classifier for the minimum error is proposed. This criterion has the advantage of selecting features having a link with the error probability with a low computational load. Experiments have been carried out in order to compare the performances provided by the proposed criterion with the ones of some of the widely used feature-selection criteria presented in the remote-sensing literature. These experiments confirm the effectiveness of the proposed criterion, which performs slightly better than all the others considered in the paper.

## 1. INTRODUCTION

One of the most critical phases of a pattern recognition processing chain is the identification of a reliable set of features that distinguish the information classes to be recognized by a classifier. In particular, given a redundant set of features, the problem is to select reduced subsets of these features to obtain the best separation of the information classes in the feature space. This process is known as "feature selection" and allows features not useful to the classification process to be neglected [3]. By reducing the number of features given as input to a classifier, feature selection makes it possible to decrease the computational time required by the classification process. Moreover, in practical situations involving a limited number of samples, the reduction in the number of features may also increase classification accuracy (Hughes phenomenon) [3].

In the literature, several feature-selection techniques have been proposed [3, 5]. Usually, they involve both a criterion function and a search algorithm. The former is aimed at evaluating the effectiveness of feature sebsets; the latter identifies a subset of features that well satisfy the adopted criterion function. Generally, criterion functions are defined for problems in which two classes are to be recognized. Several strategies can be used to apply them to multiclass cases [1, 5].

[1] This research was conducted within the framework of the research project "Sviluppo di metodi integrati di classificazione agroecologica tramite dati di telerilevamento," supported by the Italian Space Agency.

In this paper, we address the problem of feature selection in multiclass cases for remote-sensing image classification. In Section 2, we provide a brief overview of some feature-selection techniques commonly used in remote-sensing applications. Then, in Section 3, we present a feature-selection criterion that is based on a simple upper bound to the error probability of the Bayes classifier for the minimum error, formulated under some simplifying hypotheses. Similar techniques have previously been proposed in the literature, but mainly for two-class situations and not for remote-sensing applications. Experimental results on an agricultural remote-sensing data set are reported in Section 4, and conclusions are drawn in Section 5.

## 2. PREVIOUS WORK

Many feature-selection criteria have been proposed in the literature [3, 4, 5, 6]. In this section, we briefly recall some of the widely used ones in remote-sensing applications, i. e., the Bhattacharyya distance [3], the Jeffreys–Matusita distance [5], and an index based on scatter matrices [3, 4].

The Bhattacharyya distance ($B_{ij}$) and the Jeffreys–Matusita (J–M) distance ($J_{ij}$) [3, 5] are based on separability indexes between pairs of classes:

$$B_{ij} = -\ln\left\{ \int_x \sqrt{p(x|\omega_i)\, p(x|\omega_j)}\, \mathrm{d}x \right\} \tag{2.1}$$

$$J_{ij} = \left\{ \int_x \left[\sqrt{p(x|\omega_i)} - \sqrt{p(x|\omega_j)}\right]^2 \mathrm{d}x \right\}^{\frac{1}{2}} = \left[2(1 - e^{-B_{ij}})\right]^{\frac{1}{2}} \tag{2.2}$$

where $p(x|\omega_i)$ and $p(x|\omega_j)$ are conditional probability density functions for the feature vector $x$, given the classes $\omega_i$ and $\omega_j$, respectively. Both indexes allow one to evaluate the separability between classes by computing statistical distance measures in the feature space. One can apply feature selection by selecting subsets of features that maximize (2.1) or (2.2). Although, in two-class cases, the above criteria select the same features, their behaviours are significantly different. The Bhattacharyya distance increases even when classes are well-separated; on the contrary, the J–M distance exhibits a saturation effect, that is, it does not significantly increase over distance values corresponding to well-separated classes.

The above criteria can be adopted when, in a given pattern-recognition problem, two classes are considered. Many authors proposed both theoretically based and empirical generalizations of these criteria to multiclass cases. The most common strategy to apply the above distance measures to multiclass cases is to use the weighted average distances computed for all pairs of classes [5]:

$$B_{\text{ave}} = \sum_{i=1}^{C} \sum_{j=1}^{C} P(\omega_i)\, P(\omega_j)\, B_{ij} \tag{2.3}$$

$$J_{\text{ave}} = \sum_{i=1}^{C} \sum_{j=1}^{C} P(\omega_i)\, P(\omega_j)\, J_{ij} \tag{2.4}$$

where $C$ is the number of classes considered, and $P(\omega_i)$ and $P(\omega_j)$ are the apriori probabilities of the classes $\omega_i$ and $\omega_j$, respectively. From (2.3) and (2.4), it is easy to observe that, unlike two-class cases, multiclass cases allow the Bhattacharyya and the J–M criteria to select different sets of features. Thanks to the relationship existing between the J–M distance and the error probability behaviour, the criterion based on the J–M distance is usually more effective [5].

In the literature, other generalizations of the criterion based on the J–M distance to multiclass cases have been presented. Bruzzone et al [1] proposed to apply the J–M distance to multiclass cases according to the Bhattacharyya bound to the Bayes error:

$$J_{\text{bh}} = \sum_{i=1}^{C} \sum_{j>1}^{C} \sqrt{P(\omega_i)\, P(\omega_j) J_{ij}^2}. \tag{2.5}$$

One can use the J–M distance for feature selection in multiclass cases also by selecting the set of features that maximize the separability index $J_{\min}$ given by [5]:

$$J_{\min} = \min_{i,j} \left\{ J_{ij} \right\} \quad i = 1, \ldots, C;\ j = 1, \ldots, C;\ i \neq j. \tag{2.6}$$

Other feature-selection criteria are implicitly oriented toward multiclass cases (e.g., the criteria based on scatter matrices [3, 4]). The criteria based on scatter matrices evaluate the effectiveness of features by computing within-class ($S_w$) and between-class ($S_b$) scatter matrices [3]. From these matrices, several separability indexes can be derived, like for example, the following one [4]:

$$F = \frac{\mid S_w + S_b \mid}{\mid S_w \mid}. \tag{2.7}$$

This index evaluates the effectiveness of features by considering their capability to provide a large inter-class separation and a small intra-class spread by analyzing together samples of all classes.

## 3. AN UPPER BOUND TO THE BAYES ERROR PROBABILITY FOR FEATURE SELECTION

Let us consider two classes, $\omega_i$ and $\omega_j$, that are to be separated by a classifier. It is well-known that the error probability of the Bayes classifier for the minimum error is given by [3, 6]:

$$P_e(\omega_i, \omega_j) = P(\omega_i) \int_{D_j} p(x|\omega_i)\, \mathrm{d}x + P(\omega_j) \int_{D_i} p(x|\omega_j)\, \mathrm{d}x \tag{3.1}$$

where $D_i$ and $D_j$ are the "decision regions" for the classes $\omega_i$ and $\omega_j$, respectively [3, 6]. Under the hypotheses of Gaussian distributions and of two classes with equal covariance matrices (i.e., $\Sigma_i = \Sigma_j = \Sigma_{ij}$), equation (3.1) can be rewritten considering the direction that connects the mean vectors of the two classes [6]:

$$P_e(\omega_i, \omega_j) = P(\omega_i)\, Q(\sqrt{\alpha_{ij}}) + P(\omega_j)\, Q(\sqrt{\alpha_{ji}}) \tag{3.2}$$

where $Q(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \int_x^{\infty} e^{-\frac{\xi^2}{2}}\, d\xi$, and the values of the distances $\alpha_{ij}$ and $\alpha_{ji}$ depend on the optimal decision threshold computed by the maximum a-posteriori probability (MAP) rule [6]. An upper bound to (3.2) is provided by:

$$e_{ij} = \Big[P(\omega_i) + P(\omega_j)\Big] Q\left(\frac{\sqrt{d_{ij}}}{2}\right) \geq P_e(\omega_i, \omega_j) \tag{3.3}$$

where $d_{ij}$ is the Mahalanobis distance between the two classes $\omega_i$ and $\omega_j$ and is given by:

$$d_{ij} = (M_i - M_j)^t \Sigma_{ij}^{-1} (M_i - M_j) \tag{3.4}$$

and $M_i$ and $M_j$ are the mean vectors for the classes $\omega_i$ and $\omega_j$, respectively. The approximation made in (3.3) corresponds to fixing the threshold at the middle point of the segment connecting the two mean vectors, instead of using the optimal point obtained by the MAP rule.

When more than two classes are present, an upper bound to the Bayes error $P_e$ is provided by a combination of the pairwise bounds, computed for all pairs of classes as [1, 3]:

$$E = \sum_{i=1}^{C} \sum_{j>1}^{C} e_{ij} \geq P_e. \tag{3.5}$$

Features can be selected according to the minimization of the upper bound $E$.

Equation (3.3) has been derived under the hypothesis of classes with equal covariance matrices. But, in practical cases, covariance matrices may be different. Therefore, for each pair of classes, we empirically compute $\Sigma_{ij}$ as the mean value of the two covariance matrices $\Sigma_i$ and $\Sigma_j$, i.e.,

$$\Sigma_{ij} = \frac{(\Sigma_i + \Sigma_j)}{2}. \tag{3.6}$$

## 4. EXPERIMENTAL RESULTS

In order to compare the performances of the feature-selection techniques presented in Section 2 with those of the proposed approach, we considered a set of remote-sensing images related to an agricultural area near the village of Feltwell (U.K.). The images (each of 250 by 350 pixels) were obtained by using a multiband optical sensor installed on an airplane (i.e., an Airborne Thematic Mapper sensor). In our experiments, we selected from the available eleven spectral bands the six ones corresponding to the Thematic Mapper bands of the Landsat satellite (with the exception of the thermal channel). In addition, eleven nonlinear combinations of spectral bands (the so-called "vegetation indexes" [5]) were considered. As the feature selection was carried out on a pixel basis, each pixel was characterized by a seventeen-element feature vector containing the reflectances measured in the six optical bands and the eleven vegetation indexes. For our experiments, we selected 1431 samples belonging to five agricultural classes (i.e., sugar beets, stubble, bare soil, potatoes, and carrots). For all the criteria considered, we selected the best subsets of features by adopting the optimal Branch and Bound search algorithm [3]. The effectiveness of each subset of features was then assessed by evaluating the accuracies provided by the Bayes classifier. All the experiments were performed under the assumption of classes with Gaussian distributions.

The features selected by the different criteria provided the overall classification accuracies given in Figure 4.1. In particular, the accuracies were plotted with relation to the number of selected features. From the behaviours shown in the diagram, it is easy to deduce that, for the present data set, to consider more than 8 features is not useful. In fact, 8 features allow the classification accuracy to reach saturation for all the criteria (i.e, addition of further features does not significantly improve classification accuracy). In order to better compare the performances provided by the different criteria, in Table 4.1 the average classification accuracies computed for all the algorithms considered are given (the average accuracies were computed considering only the significant subsets of features, that is, the ones containing a number of features between 1 and 8). The best performances were provided by the proposed criterion, which yielded the highest average classification accuracy among the considered techniques. Concerning the classical criteria, the highest average accuracy was obtained by $J_{\mathrm{bh}}$, which performed slightly worse than the proposed criterion. $J_{\mathrm{ave}}$ and $J_{\mathrm{min}}$ gave accuracies close to that of $J_{\mathrm{bh}}$. By contrast, the performances of $F$ and $B_{\mathrm{ave}}$ were definitely worse.

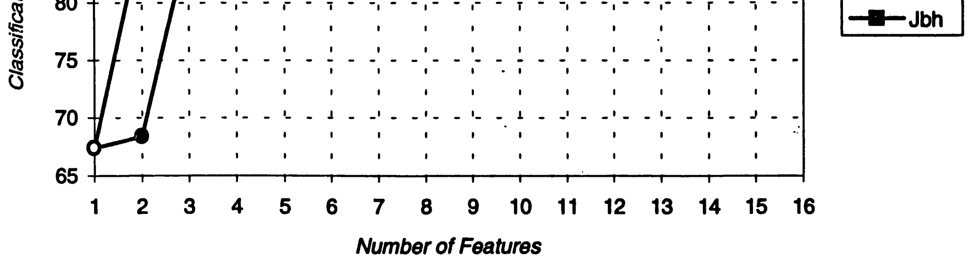

Fig. 4.1. Comparison of the classification accuracies provided by the considered techniques.

## 5. CONCLUDING REMARKS

In this paper, a criterion for feature selection based on a simple upper bound to the Bayes error probability has been presented and compared with some of the widely

**Table 4.1.** Average classification accuracies provided by the considered techniques.

| Criterion | Average classification accuracy (%) |
|---|---|
| $B_{\text{ave}}$ | 83.44 |
| $J_{\text{ave}}$ | 91.88 |
| $J_{\text{bh}}$ | 92.06 |
| $J_{\text{min}}$ | 91.92 |
| $F$ | 88.56 |
| $E$ | 93.05 |

used feature-selection criteria proposed in the remote-sensing literature. Experiments performed on an agricultural remote-sensing data set have shown that the described technique selected features that provided the best average classification accuracy. The effectiveness of the proposed criterion was also confirmed by other experiments we carried out on different remote-sensing data sets [2]. This suggests that the proposed criterion can be regarded as a valid alternative to other classical criteria presented in the remote-sensing literature.

From the computational viewpoint, the proposed technique is less expensive than all the others considered in this paper, except for the index $F$, which is the fastest (however, for the considered data set, the index $F$ was less effective).

Among the classical criteria, $J_{\text{bh}}$ turned out to be the most effective one, providing accuracies slightly worse than those obtained by the proposed approach.

(Received December 18, 1997.)

## REFERENCES

[1] L. Bruzzone, F. Roli and S. B. Serpico: An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection. IEEE Trans. Geoscience Remote Sensing *33* (1995), 6, 1318–1321.

[2] L. Bruzzone and S. B. Serpico: Feature selection in multiclass cases: a proposal and an experimental investigation. In: Proceedings of the 1st Workshop on Statistical Techniques in Pattern Recognition, Prague 1997, pp. 19–24.

[3] K. Fukunaga: Introduction to Statistical Pattern Recognition. Second edition. Academic, New York 1990.

[4] S. S. Liu and M. E. Jernigan: Texture analysis and discrimination in additive noise. Computer Vision Image Processing *49* (1990), 52–67.

[5] P. H. Swain and S. M. Davis: Remote Sensing: The Quantitative Approach. McGraw–Hill, New York 1994.

[6] J. T. Tou and R. C. Gonzalez: Pattern Recognition Principles. Addison–Wesley, London 1974.

*Lorenzo Bruzzone and Sebastiano B. Serpico, Department of Biophysical and Electronic Engineering – University of Genova, Via Opera Pia, 11A, 16145 Genova. Italy.*
*e-mail: lore@dibe.unige.it*