SEQUENTIAL FORWARD FEATURE SELECTION WITH LOW COMPUTATIONAL COST

Dimitrios Ververidis and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, costas}@zeus.csd.auth.gr

ABSTRACT

This paper presents a novel method to control the number of crossvalidation repetitions in sequential forward feature selection algorithms. The criterion for selecting a feature is the probability of correct classification achieved by the Bayes classifier when the class feature probability density function is modeled by a single multivariate Gaussian density. Let the probability of correct classification achieved by the Bayes classifier be a random variable. We demonstrate by experiments that the probability density function of the latter random variable can be modeled by a Gaussian density. Based on this observation, a method for reducing the computational burden in sequential forward selection algorithms is proposed. The method predicts the number of crossvalidation repetitions by employing a t-test to guarantee that a statistically significant improvement in the probability of correct classification is obtained by increasing the number of selected features. The proposed method is twice feaster than the sequential forward selection algorithm that uses a fixed number of crossvalidation repetitions and it maintains the performance of the sequential floating forward selection algorithm.

1. INTRODUCTION

To estimate the probability of correct classification achieved by the Bayes classifier one usually divides the available data-set into a set used for designing the classifier (i.e. the training set) and a set used for testing the classifier (i.e. the test set). There are methods frequently employed to estimate the probability of correct classification, namely the *resubstitution* method [1], the *leave-one-out* method (or Jacknife test) [2], and the *crossvalidation* method [3, 4].

The resubstitution method uses the whole data-set for training and testing the classifier resulting in a biased estimate of the probability of correct classification. The leave-one-out method alleviates the lack of independence between the training set and the test set in the resubstitution method, because the training is performed using all the samples of the data-set but one. The latter is used for testing. The procedure is repeated for all the samples of the data-set resulting in a long execution time. The crossvalidation method reduces the high computational burden of the leave-one-out method, as it chooses randomly a number of samples greater than one from the data-set to construct the test set and uses the remaining samples for training the classifier. This procedure is repeated for several times defined by the user and the estimated probability of correct classification is the average correct classification rate for all repetitions. The number of crossvalidation repetitions is usually set between 10 and 100. In this paper, the number of crossvalidation repetitions in sequential forward selection algorithms is controlled through a *t-test* that guarantees statistically significant improvements in the probability of correct classification for the Bayes classifier, when new features are added. The method results in low computational demands while maintains the same high accuracy achieved by the sequential floating forward selection algorithm for a fixed number of crossvalidation repetitions.

The outline of the paper is as follows. In Section 2, the data extracted from the Danish Emotional Speech (DES) database [5] are briefly described. The prosody features extracted from the speech utterances are presented in Section 3. Section 4 treats the probability of correct classification achieved by the Bayes classifier as a random variable and studies its distribution with respect to the number of crossvalidation repetitions and the selection of the feature set. Based on the observations of Section 4, a mechanism that controls the number of crossvalidation repetitions is developed in the next section. The mechanism is incorporated into the sequential floating forward selection (SFFS) algorithm and the sequential forward selection (SFS) algorithm to speed up their execution. A comparison of the time savings between the proposed variant of SFS algorithm and the SFS for a fixed number of crossvalidation repetitions is reported in Section 6. We also demonstrate that the proposed SFFS variant maintains the performance achieved by the SFFS algorithm for a fixed number of crossvalidation repetitions. Finally,conclusions are drawn in Section 7.

2. DATA

The audio data used in the experiments consist of 1300 utterances, 800 more than those used in [6], that are manually extracted from DES. Each utterance is a speech segment between two silence pauses. The 800 utterances, that are now included, are detached from paragraphs, whereas the old 500 utterances corresponded to isolated words and sentences. The utterances are expressed by four professional actors, two males and two females, in 5 emotional styles such as anger, happiness, sadness, surprise, and neutral.

3. FEATURE EXTRACTION AND PREPROCESSING

Pitch estimates are obtained from the peaks of the short-term autocorrelation function of the speech amplitude. The short-term analysis is performed using windows of duration 15 msec. We assume that pitch frequencies are limited to the range 60-320 Hz. For esti-

This work has been supported by the FP6 European Union Network of Excellence MUSCLE "Multimedia Understanding through Semantics, Computation and LEarning" (FP6-507752).

mating the 4 formants, we use a method based on linear prediction analysis. The method finds the angle of the poles in the $\mathcal{Z}$ plane for an all-pole model and considers the poles that are further from zero as indicators of the formant frequencies. To estimate the energy, a simple short-term energy function has been used. After the evaluation of the primary raw features, secondary statistical features were extracted from the primary ones. The statistical features employed in our study are grouped in several classes. The speech features computed and their corresponding indices can be found in [6].

Let $\mathcal{X}$ denote the feature set. Each feature $X_k \in \mathcal{X}$, $k = 1, \ldots, 87$ has its own dynamic range. Features with variance of order 10^6 such as the fourth formant, have greater influence in the classifier design than features with a variance of order 10^2 such as the mean value of pitch. Thus, a linear transformation is applied to each one of the 87 features. Let $a_k = \min_i\{X_{ki}\}$ and $b_k = \max_i\{X_{ki}\}$ for $i = 1, \ldots, N_S$, where N_S equals to the total number of utterances. A linear transformation from $[a_k, b_k]$ to $[0, 1]$ is applied for each X_k.

The exponentially distributed features may lead to an increased computational time and underflow warnings, as they become too dense near the lower bound which in our case is 0_+. Accordingly, after the linear transformation, we apply a logarithmic transformation to the exponentially distributed features.

4. VARIATION OF THE MEAN PROBABILITY OF CORRECT CLASSIFICATION DURING CROSSVALIDATION PROCEDURE

The probability of correct classification achieved by the Bayes classifier that uses the feature set $\mathcal{Z} \subseteq \mathcal{X}$ can be estimated by crossvalidation for $nrep$ repetitions using

$$J_{nrep}(\mathcal{Z}) = 1 - E[\{\varepsilon(\mathcal{Z}, \mathcal{T}_r; \mathcal{D}_r)\}_{r=1}^{nrep}] \quad (1)$$

where $\varepsilon(\mathcal{Z}, \mathcal{T}_r; \mathcal{D}_r)$ is the probability of error for the Bayes classifier designed using $\mathcal{D}_r$ during training when it is applied to $\mathcal{T}_r$. In (1) the expectation is applied over the sequence of error probabilities measured over $\mathcal{T}_r, r = 1, 2, \ldots, nrep$ and the dependence of J_{nrep} on the feature set $\mathcal{Z}$ is explicitly stated. 90% of the available utterances are used to build $\mathcal{D}_r$ and the remaining 10% create $\mathcal{T}_r$. The training and the test sets are complementary.

Let the features X_k be treated as elements of a d-dimensional random vector $\mathbf{x}$ (e.g. a pattern). Let Ω_m denote the mth class, $P(\Omega_m)$ be the a priori probability of class Ω_m, and $p_r(\mathbf{x}|\Omega_m)$ be the class conditional probability density function (pdf). At each crossvalidation repetition r, we assume that the labels of the training set are known, whereas the labels of the test set are unknown. If the number of classes is c, the samples in the training set $\mathcal{D}_r$ can be divided into c subsets $\mathcal{D}_{r;m}$ with cardinalities $N_{\mathcal{D}_{r;m}}$, $m = 1, 2, \ldots, c$, such that

$$N_{\mathcal{D}_r} = \sum_{m=1}^{c} N_{\mathcal{D}_{r;m}}. \quad (2)$$

Since the labels of the training set are known the pdf $p_r(\mathbf{x}|\Omega_m)$ of each class Ω_m, $m = 1, 2, \ldots, c$, can be estimated.

The Bayes classifier assigns the pattern $\mathbf{x} \in \mathcal{T}_r$ to Ω_m if

$$P(\Omega_m)\, p_r(\mathbf{x}|\Omega_m) > P(\Omega_j)\, p_r(\mathbf{x}|\Omega_j) \quad (3)$$

$j = 1, \ldots, m-1, m+1, \ldots, c$. Let $\mathcal{L}_m$ be the region where $\mathbf{x}$ is classified to Ω_m and $\mathcal{L} = \cup_{m=1}^{c} \mathcal{L}_m$. We also define the complement of $\mathcal{L}_m$ as $\mathcal{L}_m^c = \mathcal{L} - \mathcal{L}_m$. The probability of error for the Bayes classifier is given by

$$\varepsilon = \sum_{m=1}^{c} P(\Omega_m) \int_{\mathcal{L}_m^c} p_r(\mathbf{x}|\Omega_m)\, d\mathbf{x}. \quad (4)$$

Let the pdf $p_r(\mathbf{x}|\Omega_m)$ be modeled by a single multivariate Gaussian density

$$p_r(\mathbf{x}|\Omega_m) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{r;m})^T\, \boldsymbol{\Sigma}_{r;m}^{-1}\, (\mathbf{x} - \boldsymbol{\mu}_{r;m})\right]}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_{r;m}|^{1/2}} \quad (5)$$

where $\boldsymbol{\mu}_{r;m}$ is the mean vector and $\boldsymbol{\Sigma}_{r;m}$ is the covariance matrix of the feature vectors.

Let us treat the probability of correct classification achieved by the Bayes classifier $J_{nrep}(\mathcal{Z})$ as a random variable. Its pdf $f(J_{nrep}(\mathcal{Z}))$ for a particular feature set $\mathcal{Z}$ and several choices of $nrep$ is plotted in Figure 1 when the class pdfs $p_r(\mathbf{x}|\Omega_m)$ are modeled as Gaussian distributions (5). It is seen that $J_{nrep}(\mathcal{Z})$ follows a Gaussian distribution.

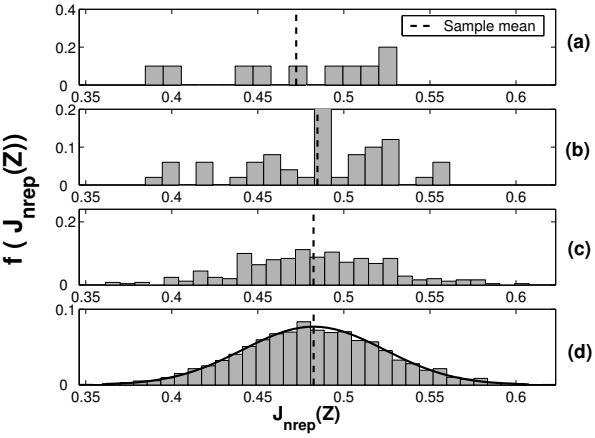

Fig. 1. Probability density function of $J_{nrep}(\mathcal{Z})$ for several choices of $nrep$: (a) $nrep$=10; (b) $nrep$=50; (c) $nrep$=300; (d) $nrep$=1000.

The pdf of $J_{nrep}(\mathcal{Z}_i)$ for several feature sets $\mathcal{Z}_i$ is plotted in Figure 2 when $nrep$=1000. The pdfs marked by A, B, C correspond to three emotional speech feature sets. For several other pdfs that correspond to real feature sets, the peak at the mode of each pdf is marked with $*$. Moreover, pdfs for artificially created feature sets whose class pdfs are modeled by (5) for five classes have been created. For each pdf, the peak at its mode is marked with $\circ$. It can be seen that the variance $\sigma^2(J_{1000}(\mathcal{Z}_i)) = \mathrm{var}\{J_{1000}(\mathcal{Z}_i)\}$ depends on the level of $J_{1000}(\mathcal{Z}_i)\}$ as is quantified by its mean value (i.e. the mode of the pdf)

$$\mu(J_{1000}(\mathcal{Z}_i)) = 1 - E[\{\varepsilon(\mathcal{Z}_i, \mathcal{T}_r; \mathcal{D}_r)\}_{r=1}^{1000}] = J_{1000}(\mathcal{Z}_i). \quad (6)$$

Let $g(J_{1000}(\mathcal{Z}_i))$ be a polynomial of degree 3 fitted to the peaks of $f(J_{1000}(\mathcal{Z}_i))$ in mean squared error (MSE) sense. The variance of $J_{1000}(\mathcal{Z})$ can be estimated by

$$\sigma^2(J_{1000}(\mathcal{Z}_i)) = \frac{1}{2\pi g^2(J_{1000}(\mathcal{Z}_i))} \quad (7)$$

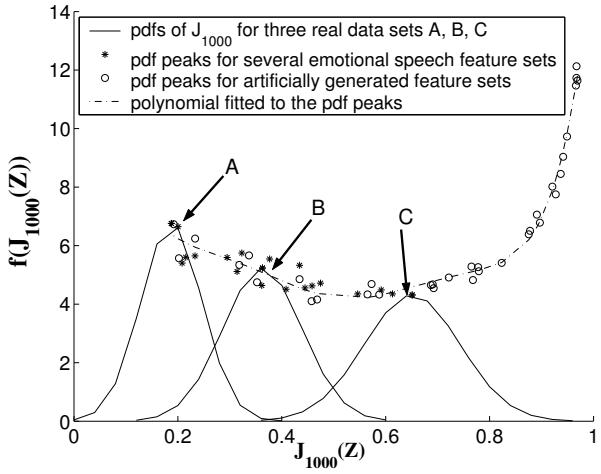

Fig. 2. Probability density function of $J_{1000}(\mathcal{Z}_i)$ for several feature set selections $\mathcal{Z}_i$.

Let us assume that $f(J)$ for infinite crossvalidation repetitions follows a Gaussian distribution with mean μ_∞ and variance σ^2_∞. If the number of crossvalidation repetitions is set to $nrep$ then $f(J_{nrep}(\mathcal{Z}))$ follows a Gaussian distribution with the same mean μ_∞ and variance

$$\sigma^2(J_{nrep}(\mathcal{Z})) = \frac{\sigma^2_\infty}{nrep}. \tag{8}$$

Let us estimate σ^2_∞ by $\sigma^2(J_{1000}(\mathcal{Z}))$. Then $f(J_{nrep}(\mathcal{Z}))$ is a Gaussian pdf with mean $\mu(J_{1000}(\mathcal{Z}))$ and variance

$$\sigma^2(J_{nrep}(\mathcal{Z})) = \frac{1}{2\pi\, nrep\, g^2(J_{1000}(\mathcal{Z}))}. \tag{9}$$

If our aim is to keep the variance $\sigma^2(J_{nrep}(\mathcal{Z}))$ constant, i.e.

$$\sigma^2(J_{nrep}(\mathcal{Z})) = \gamma \tag{10}$$

the number of crossvalidation repetitions should be set equal to

$$nrep = \frac{1}{2\pi\, \gamma\, g^2(J_{1000}(\mathcal{Z}))}. \tag{11}$$

For example, by inspecting Figure 2, it can be seen that the number of crossvalidation repetitions estimated by (11) for the feature set B, that gives a mean probability of correct classification equal to 0.38 when employed in the Bayes classifier, is smaller than that for the feature set C for which the Bayes classifier attains a mean probability of correct classification equal to 0.64.

In the following, we are interested in testing the hypothesis

$$J_{nrep_1}(\mathcal{Z}_1) > J_{nrep_2}(\mathcal{Z}_2). \tag{12}$$

This is accomplished by using the test statistic

$$q = \frac{J_{nrep1}(\mathcal{Z}_1) - J_{nrep2}(\mathcal{Z}_2)}{\sqrt{\frac{\sigma^2_{J_{nrep1}(\mathcal{Z}_1)}}{nrep1} + \frac{\sigma^2_{J_{nrep2}(\mathcal{Z}_2)}}{nrep2}}} = \frac{J_{nrep1}(\mathcal{Z}_1) - J_{nrep2}(\mathcal{Z}_2)}{\sqrt{2\cdot\gamma}} \tag{13}$$

which is distributed approximately as Student's t-statistic with a number of degrees of freedom equal to

$$\kappa = \frac{\left[\frac{1}{nrep1} + \frac{1}{nrep2}\right]^2}{\frac{1}{nrep1^2(nrep1-1)} + \frac{1}{nrep2^2(nrep2-1)}}. \tag{14}$$

The hypothesis (12) is accepted when $q > t_{1-a}(\kappa)$ where a equals 0.05. The test-statistic depends only on the distance between the corresponding probabilities of correct classification and it is greater when γ becomes smaller. The user selects γ with respect to the computation speed, as it can be inferred from (11). When γ becomes smaller the number of crossvalidation repetitions $nrep$ becomes greater and accordingly the computational time increases.

5. APPLICATION TO SEQUENTIAL FEATURE SELECTION ALGORITHMS

In this section, we will augment the sequential forward feature selection algorithms, either the standard SFS or its floating variant SFFS, by a mechanism that controls the number of crossvalidation repetitions based on the analysis of Section 4. The SFS consists of a forward step which is as follows: starting from an initially empty set of features $\mathcal{Z}_0$, at each forward (inclusion) step at the level l we seek the feature $X^+ \in (\mathcal{X} - \mathcal{Z}_{l-1})$ such that for $\mathcal{Z}_l = \mathcal{Z}_{l-1} \cup \{X^+\}$ the probability of correct classification achieved by the Bayes classifier $J(\mathcal{Z}_l)$ is maximized. In addition to the aforementioned inclusion step the SFFS algorithm [7] applies a number of backward steps (exclusions) as long as the resulting subsets are better than the previously derived ones at this level. Consequently, there are no backward steps at all when the performance cannot be improved. The exclusion step is as follows. We exclude at level l those $Z^- \in \mathcal{Z}_l$ as long as the correct classification of the Bayes classifier for the feature set $\mathcal{Z}_l^- = \mathcal{Z}_l - \{Z^-\}$, $J(\mathcal{Z}_l^-)$, is greater than $J(\mathcal{Z}_l)$.

The goal at level l is to find which non-selected feature $X \in (\mathcal{X} - (\mathcal{Z}_{l-1}))$ yields the greatest improvement in the probability of correct classification for the Bayes classifier among the non-selected features. That is, the feature that yields

$$J_{max} = \max_{X \in (\mathcal{X} - \mathcal{Z}_{l-1})} J_{nrep}(\mathcal{Z}_{l-1} + \{X\}). \tag{15}$$

If $nrep$ is a large number, then $J_{nrep}(\mathcal{Z}_{l-1} + \{X\})$ is an accurate estimate of the maximum probability of correct classification one might expect from the Bayes classifier. But the computation is time consuming. If $nrep$ is small, $J_{nrep}(\mathcal{Z}_{l-1} + \{X\})$ is computed faster, but it is not accurate.

In the proposed method, the number of crossvalidation repetitions $nrep$ in (15) is predicted by (11) from the first 10 crossvalidation repetitions for the feature set $(\mathcal{Z}_{l-1} + \{X\})$ and it is denoted as $\widehat{nrep}$. $\mu(J_{1000}(\mathcal{Z}_{l-1} + \{X\}))$ can be replaced with $\mu(J_{10}(\mathcal{Z}_{l-1} + \{X\}))$, because the sample mean does not change dramatically by varying the number of crossvalidation repetitions as can be seen in Figure 1.

Let us separate the features $X \in (\mathcal{X} - \mathcal{Z}_{l-1})$ in potentially expressive features and potentially bad features. The former features yield $J_{nrep1}(\mathcal{Z}_{l-1} + \{X\}) \geq J_{\widehat{nrep}}(\mathcal{Z}_{l-1})$, while the latter ones consistently yield $J_{nrep2}(\mathcal{Z}_{l-1} + \{X\}) < J_{\widehat{nrep}}(\mathcal{Z}_{l-1})$ where $10 < nrep1, nrep2 < \widehat{nrep}$. We propose to formulate a t-test in order to test the hypothesis $H_0^A : J_{nrep2}(\mathcal{Z}_{l-1} + \{X\}) < J_{\widehat{nrep}}(\mathcal{Z}_{l-1})$ at 95% significance level for a small number of repetitions (e.g. $nrep2$=10). If the hypothesis is accepted, we discard

the feature X. Otherwise, we perform more repetitions checking each time the validity of the hypothesis H_0^A. If $nrep2$ has reached $\widehat{nrep}$ then we perform another t-test to check whether the hypothesis $H_0^B : J_{\widehat{nrep}}(\mathcal{Z}_{l-1} + \{X\}) \geq J_{\widehat{nrep}'}(\mathcal{Z}_{l-1})$ can be accepted. If H_0^B is accepted then the feature X is added to $\mathcal{Z}_{l-1}$. Figure 3 explains in detail the proposed mechanism to be incorporated in SFS or SFFS algorithms.

Initialize
Set $J^{max} = J_{nrepA}(Z_{l-1} + \{X_1\})$ where $X_1 \in (\mathcal{X} - \mathcal{Z}_{l-1})$, $X^{opt} = X_1$. $nrepA$ is calculated from (11): $nrepA = [2\pi \cdot g^2(J_{10}(Z_{l-1} + \{X_1\})) \cdot \gamma]^{-1}$
Loop 1: For all $X \in (\mathcal{X} - \mathcal{Z}_{l-1} - \{X_1\})$,
  $nrepA = [2\pi \cdot g^2(J_{10}(\mathcal{Z}_{l-1} + \{X\})) \cdot \gamma]^{-1}$
  $nrep = 10$
  Loop 2: while $nrep \leq nrepA$
    Test $H_0^A : J_{nrep}(\mathcal{Z}_{l-1} + X) < J^{max}$ at 95% significance level.
      if H_0 is accepted, no further crossvalidation repetitions are required so we proceed to the next non-selected feature.
      go to Loop 1
      end if
      if H_0 is rejected then the feature X might improve J. More repetitions are need to validate its usefulness.
      $nrep = nrep + 1$ **end if**
  end for Loop 2
  So, feature X with $J_{nrepA}(\mathcal{Z}_{l-1} + \{X\})$ was never rejected for $nrepA$ crossvalidation repetitions, but is it actually better than the best feature found up to now X^{opt} that achieves J^{max}?
  Test $H_0^B : J_{nrepA}(\mathcal{Z}_{l-1} + X) > J^{max}_{nrepB}$ with a two sided t-test at 95% significance level using the test statistic (13).
    if H_0^B is accepted then,
    $J^{max} = J_{nrepA}(\mathcal{Z}_{l-1} + \{X\})$
    $X^{opt} := \{X\}$ **end if**
end for Loop 1
$X^+ := X^{opt}$ // This feature must be included

Fig. 3. Algorithm to determine the best feature among the non-selected ones with low computational cost while maintaining high accuracy.

6. RESULTS

To demonstrate the usefulness of the proposed method we compare the SFS and SFFS for fixed number of 70 crossvalidation repetitions against our variants of SFS and SFFS with variable number of crossvalidation repetitions up to 70. To evaluate the speed of the proposed method, we compare the execution time needed for our variant of SFS against the execution time needed for the SFS with a fixed number of crossvalidation repetitions. SFS is more suitable for speed evaluation than SFFS, because the former employs a fixed number of forward steps. The results are shown in Table 1. The proposed method is twice faster than the ordinary SFS. In order to determine the accuracy of the proposed method, we compare the probability of correct classification achieved by the Bayes classifier when the ordinary SFFS algorithm with a fixed number of crossvalidation repetitions is employed with that when the proposed variant of SFFS is used. SFFS is preferred than SFS for accuracy comparison, because the former is not sensitive to nesting problems. As can be seen from Table 2 there is no performance deterioration.

Table 1. Time lapsed in secs for SFS.

Data set	SFS	SFS with t-test
500 utterances	2547	1231
1300 utterances	2710	1159

Table 2. Probability of correct classification for SFFS.

Data set	SFFS	SFFS with t-test
500 utterances	0.563	0.558
1300 utterances	0.485	0.487

7. CONCLUSIONS

In this paper, we studied the distribution properties of the probability of correct classification achieved by the Bayes classifier and demonstrated by experiments that it follows a Gaussian distribution. We have proposed a t-test that can be incorporated within the SFFS or SFS algorithms to control the number of crossvalidation repetitions. The proposed SFS variant is twice faster than the SFS algorithm for a fixed number of crossvalidation repetitions. Moreover, the proposed SFFS variant maintains the same accuracy with the SFFS for a fixed number of crossvalidation repetitions. In all the cases, the class pdfs are modeled by a single multivariate Gaussian density. The proposed method can be applied to other sequential forward or backward feature selection algorithms as well.

8. REFERENCES

[1] D. Foley, "Consideration of sample and feature size," *Trans. Information Theory*, vol. 18, no. 5, pp. 618–626, 1972.

[2] P. A. Lachenbruch and R. M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.

[3] M. Stone, "Cross-validatory choice and assesment of statistical predictions," *J. Roy. Statist. Soc. B*, vol. 36, no. 2, pp. 111–147, 1974.

[4] B. Efron and R. E. Tibshirani, *An Introduction to the Bootstrap.*, N.Y.: Chapman & HALL/CRC, 1993.

[5] I. S. Engberg and A. V. Hansen, *Documentation of the Danish Emotional Speech Database,*, Internal report, Center for Person Kommunikation, 1996.

[6] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. 2004 European Signal Processing Conf.*, Sep. 2004, pp. 341–344.

[7] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. Int. Conf. Pattern Recognition*, 1994, vol. 1, pp. 279–283.