D. Ververidis and C. Kotropoulos, "Accurate estimate of the cross-validated prediction error variance in Bayes classifiers," in Proc. Machine Learning for Signal Processing (MLSP), Thessaloniki, 2007.

ACCURATE ESTIMATE OF THE CROSS-VALIDATED PREDICTION ERROR VARIANCE IN BAYES CLASSIFIERS

Dimitrios Ververidis and Constantine Kotropoulos

Dept. of Informatics, Aristotle Univ. of Thessaloniki,
Box 451, Thessaloniki 54124, Greece.
E-mails:{jimver,costas}@aiia.csd.auth.gr

ABSTRACT

A relationship between the variance of the prediction error committed by the Bayes classifier and the mean prediction error was established by experiments in emotional speech classification within a cross-validation framework in a previous work. This paper theoretically justifies the validity of the aforementioned relationship. Furthermore, it proves that the new estimate of the variance of the prediction error, treated as a random variable itself, exhibits a much smaller variance than the usual estimate obtained by cross-validation even for a small number of repetitions. Accordingly, we claim that the proposed estimate is more accurate than the usual, straightforward, estimate of the variance of the prediction error obtained by applying cross-validation.

1. INTRODUCTION

Two popular methods for estimating the prediction error of a classifier are bootstrap and cross-validation. In these methods, the available dataset is divided repeatedly into a set used for desigining the classifier (i.e. the training set) and a set used for testing the classifier (i.e. the test set). By averaging the prediction error over all repetitions, hopefully a more accurate estimate of the prediction error is obtained than just using the prediction error in one repetition of the experiment. Both cross-validation [1] and bootstrap [2] stem from the jackknife method. Jackknife was introduced by M. Quenouille for finding unbiased estimates of statistics, such as the sample mean and the sample variance [3, 4]. Originally, jackknife meant to divide the dataset into two equal sets, to derive the target statistic over the two sets independently, and next to average the statistic estimates in order to obtain an unbiased statistic. Later, jackknife was about to split the dataset into many sets of equal cardinality. In another version, the statistic is estimated on the whole dataset except one sample, this procedure is repeated in a cyclic fashion, and the average estimate is found finally. The latter "leave-one-out" version dominates in practice. A review of jackknife variants can be found in [5].

The ordinary cross-validation is the extension of the jackknife method to derive an unbiased estimate of the prediction error in the "leave-one-out" sense [1]. The computational demands of the ordinary cross-validation are rising proportionally to the number of samples. A variant of cross-validation with a smaller number of repetitions is the s-fold cross-validation. During s-fold cross-validation the dataset is divided into s roughly equal subsets, the samples in the $s-1$ subsets are used for training the classifier, and the samples in the last subset are used for testing. The procedure is repeated for each one of the s subsets in a cyclic fashion and the prediction error is estimated by averaging the prediction errors measured in the test phase of the s repetitions. Burman proposed the repeated s-fold cross-validation for model selection, which is simply the s-fold cross-validation repeated many times [6]. The prediction error measured during cross-validation repetitions is a random variable that follows the Gaussian distribution. Therefore, according to the central limit theorem (CLT), the more the repetitions of the random variable are, the less varies the average prediction error. Throughout this paper, the repeated s-fold cross-validation is simply denoted as cross-validation for short.

The outline of the paper is as follows. Section 2 deals with the prediction error committed by the Bayes classifier. A theoretical analysis of the factors that affect the variance of the prediction error is made in Section 3. A comparison of the proposed method that predicts the variance of the cross-validated prediction error from a small number of repetitions against the usual estimate of the variance of the prediction error is presented in Section 4. Finally, Section 5, concludes the paper by indicating future research directions.

2. CLASSIFIER DESIGN

Let $\mathbf{u}^{\mathcal{W}} = \{\mathbf{u}_i^{\mathcal{W}}\}_{i=1}^{N}$ be a set of N samples, where $\mathcal{W} = \{w_k\}_{k=1}^{K}$ is the feature set comprising K features w_k. The samples can be considered as independent and identically distributed (i.i.d.) random variables (r.vs) distributed according to the multivariate distribution F of the feature set $\mathcal{W}$. Each sample $\mathbf{u}_i^{\mathcal{W}} = (\mathbf{y}_i^{\mathcal{W}}, l_i)$ is treated as a pattern consisting of a measurement vector $\mathbf{y}_i^{\mathcal{W}}$ and a label $l_i \in \{1, 2, \ldots, C\}$, where C is the total number of classes.

Let us predict the label of a sample by processing the feature vectors using a classifier. Cross-validation (CV) calculates the mean over $b = \{1, 2, \ldots, B\}$ prediction error estimates as follows. Let $s \in \{2, 3, \ldots, N/2\}$ be the number of folders the data should be divided into. To find the bth prediction error estimate, $N_{\mathcal{D}} = \frac{s-1}{s}N$ samples are randomly selected without resubstitution from $\mathbf{u}^{\mathcal{W}}$ to build the design set $\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}$ while the remaining $\mathbf{u}_{Tb}^{\mathcal{W}}$ of $\frac{N}{s}$ samples forms the test set.

The prediction error in CV repetition b is the error committed by the Bayes classifier. For sample $\mathbf{u}_i^{\mathcal{W}} = (\mathbf{y}_i^{\mathcal{W}}, l_i) \in \mathbf{u}_{Tb}^{\mathcal{W}}$, the

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

class label η predicted by the Bayes classifier is given by

$$\eta(\mathbf{y}_i^{\mathcal{W}}) = \arg\max_{c=1}^{C}\{p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c)P_b(\Omega_c)\}, \quad (1)$$

where $P_b(\Omega_c) = N_{cb}/N$ is the a priori class probability, N_{cb} is the number of samples that belong to class $\Omega_c,\ c = 1, 2, \ldots, C$ in the bth cross-validation repetition, and $p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c)$ is the class conditional probability density function (pdf) of the sample $\mathbf{u}_i^{\mathcal{W}}$ given Ω_c.

The class conditional pdf is modeled as a single Gaussian. Two parameters for each class Ω_c are required for a Gaussian pdf, namely the mean vector $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$. If $\mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}} = \{\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}} \cap \Omega_c\}$, and $N_{\mathcal{D}bc}$ is the number of samples in $\mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}}$, then the class sample mean vector and the class sample dispersion matrix can be used as estimates of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$, i.e.

$$\hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}} = \frac{1}{N_{\mathcal{D}bc}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}}} \mathbf{y}_i^{\mathcal{W}}, \quad (2)$$

$$\hat{\boldsymbol{\Sigma}}_{bc}^{\mathcal{W}} = \frac{1}{N_{\mathcal{D}bc}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}}} (\mathbf{y}_i^{\mathcal{W}} - \hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}})(\mathbf{y}_i^{\mathcal{W}} - \hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}})^T. \quad (3)$$

If $|\boldsymbol{\Gamma}|$ denotes the determinant of matrix $\boldsymbol{\Gamma}$ and $\mathcal{G}()$ denotes the Gaussian pdf, then the class conditional pdf is given by

$$p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c) = \mathcal{G}(\mathbf{y}_i^{\mathcal{W}}; \hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}}, \hat{\boldsymbol{\Sigma}}_{bc}^{\mathcal{W}}) = \frac{1}{(2\pi)^{\frac{K}{2}}|\boldsymbol{\Sigma}_{bc}^{\mathcal{W}}|^{\frac{1}{2}}}$$
$$\exp[-\frac{1}{2}(\mathbf{y}_i^{\mathcal{W}} - \hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}})^T(\hat{\boldsymbol{\Sigma}}_{bc}^{\mathcal{W}})^{-1}(\mathbf{y}_i^{\mathcal{W}} - \hat{\boldsymbol{\mu}}_{bc}^{\mathcal{W}})]. \quad (4)$$

Let $\mathcal{L}[l_i, \eta(\mathbf{y}_i^{\mathcal{W}})]$ denote the zero-one loss function between the label l_i and the predicted class label $\eta(\mathbf{y}_i^{\mathcal{W}})$ for $\mathbf{u}_i^{\mathcal{W}}$, i.e.

$$\mathcal{L}[l_i, \eta(\mathbf{y}_i^{\mathcal{W}})] = \begin{cases} 0 & \text{if } l_i = \eta(\mathbf{y}_i^{\mathcal{W}}) \\ 1 & \text{if } l_i \neq \eta(\mathbf{y}_i^{\mathcal{W}}) \end{cases}. \quad (5)$$

If $err(\hat{\mathcal{F}}(\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}), \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}})$ is the error predicted by the Bayes classifier $\hat{\mathcal{F}}$ that is designed using the set $\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}$ when it is applied to set $\mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}$ for classification, then the CV estimate of prediction error in a single repetition b is

$$CV_e^b(\mathbf{u}^{\mathcal{W}}) = err(\hat{\mathcal{F}}(\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}), \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}) = \frac{1}{N_{\mathcal{T}}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}} \mathcal{L}[l_i, \eta(\mathbf{y}_i^{\mathcal{W}})], \quad (6)$$

where $N_{\mathcal{T}} = card(\mathbf{u}_{\mathcal{T}b}^{\mathcal{W}})$ is the cardinality of the test set $\mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}$. The CV estimate of the prediction error over B repetitions is given by

$$MCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{1}{B}\sum_{b=1}^{B} CV_e^b(\mathbf{u}^{\mathcal{W}}), \quad (7)$$

and its variance is

$$VCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{1}{B}\sum_{b=1}^{B}[CV_e^b(\mathbf{u}^{\mathcal{W}}) - MCV_e^B(\mathbf{u}^{\mathcal{W}})]^2. \quad (8)$$

In [7], it is experimentally found by using linear regression that

$$VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) = \frac{s^2}{(s+\sqrt{2})N} MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big). \quad (9)$$

In the next section, theoretical evidence about (9) is provided and discussed.

3. THEORETICAL ANALYSIS

Lemma 1 *For a two-class pattern recognition problem, when each class pdf is modeled by a single Gaussian, the prediction error in one CV repetition $CV_e^b(\mathbf{u}^{\mathcal{W}})$ can be approximated as a function of the difference of the class means.*

Proof Let us assume that the set $\mathbf{u}^{\mathcal{W}} = \{\mathbf{u}_i^{\mathcal{W}}\}_{i=1}^{N}$ consists of infinite samples ($N = \infty$) that belong to two classes $\Omega_c,\ c = 1, 2$. Each class conditional pdf $p(y \mid \Omega_c)$ is a Gaussian pdf $\mathcal{G}(y; \mu_c, \sigma_c^2)$, where μ_c and σ_c^2 are the sample mean and the sample variance for each class $\Omega_c,\ c = 1, 2$, respectively. Without any loss of generality we assume $\mu_2 > \mu_1$. Let $P(\Omega_c) = N_c/N$ be the *a priori probability* of each class. In Figure 1, $P(\Omega_c)p(y \mid \Omega_c)$ is plotted for each class as a function of the measurement value y.

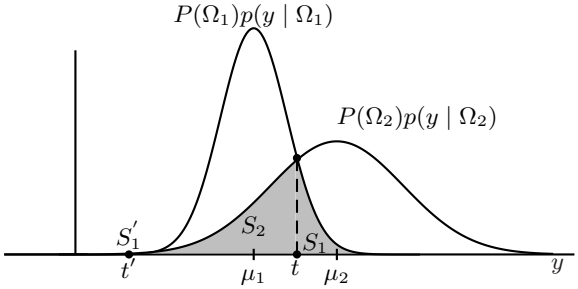

Fig. 1. Prediction error of a Bayes classifier based on a single measurement for a two-class problem.

Let t and t' denote the measurement values where $P(\Omega_c)p(y \mid \Omega_c)$ for both classes are equal. The points t, t' can be found by solving the equation $P(\Omega_1)p(y \mid \Omega_1) = P(\Omega_2)p(y \mid \Omega_2)$. The exact solutions are derived in the Appendix. The prediction error for Ω_1 is the sum of areas S_1 plus S_1', while the prediction error for class Ω_2 is the area S_2. The total prediction error is $P_e = (S_1 + S_1') + S_2$. Because $S_1' \ll S_1$, the term S_1' will be ignored. Then

$$P_e = S_1 + S_2 = P(\Omega_1)\int_t^{+\infty} p(y \mid \Omega_1)dy + P(\Omega_2)\cdot$$
$$\cdot \int_{-\infty}^{t} p(y \mid \Omega_2)dy = \frac{1}{2} - P(\Omega_1)\mathrm{sgn}(\frac{t-\mu_1}{\sigma_1})\mathrm{erf}(|\frac{t-\mu_1}{\sigma_1}|) +$$
$$P(\Omega_2)\mathrm{sgn}(\frac{t-\mu_2}{\sigma_2})\mathrm{erf}(|\frac{t-\mu_2}{\sigma_2}|) \quad (10)$$

where $\mathrm{sgn}(x)$ is the sign function, $\mathrm{erf}(x)$ is the error function defined as

$$\mathrm{erf}(x) = \int_0^x \frac{1}{\sqrt{2\pi}}\exp(-\frac{\xi^2}{2})d\xi, \quad (11)$$

The proof of (10) can be found in [8].

It is clearly seen that P_e is a function of variables $\mu_1, \mu_2, \sigma_1, \sigma_2$, and t. In the Appendix, it is shown that the ratio $(t - \mu_c)/\sigma_c,\ c = 1, 2$, is always a function of $\mu_2 - \mu_1$. Let $\varrho = \mu_2 - \mu_1$. Then, P_e can be rewritten as

$$P_e(t, \mu_1, \mu_2, \sigma_1, \sigma_2) = P_e(\varrho, \sigma_1, \sigma_2) \quad (12)$$

The maximum prediction error $\overline{P_e}$ is 0.5, when $\varrho = 0$ and $\sigma_1 = \sigma_2$. In Figure 2, $P_e(\varrho, \sigma_1, \sigma_2)$ is plotted for $\varrho \in (-\infty, +\infty)$, when $\sigma_1 = \sigma_2$ using (10).

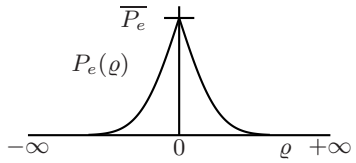

Fig. 2. Prediction error for a two-class problem, P_e, as a function of the difference of class means $\varrho = \mu_2 - \mu_1$.

Let us assume a finite number of samples. In each cross-validation repetition b, to estimate the parameters $\mu_{cb}, \sigma_{cb},\ c = 1, 2$, on the design sets, N_{1b} and N_{2b} samples are selected from the available dataset set. According to CLT [9], the sample means of measurements y are distributed as

$$\mu_{1b} \sim \mathcal{G}(\mu_1, \frac{\sigma_1^2}{N_{1b}}) = \mathcal{G}(\mu_1, \frac{\sigma_1^2}{P_b(\Omega_1)N}), \quad (13)$$

$$\mu_{2b} \sim \mathcal{G}(\mu_2, \frac{\sigma_2^2}{P_b(\Omega_2)N}). \quad (14)$$

If $\varrho_b = \mu_{2b} - \mu_{1b}$, then it can be shown that

$$\varrho_b \sim \mathcal{G}(\varrho, \sigma^2), \quad (15)$$

where

$$\varrho = \mu_2 - \mu_1, \quad (16)$$

$$\sigma^2 = \frac{\sigma_1^2}{P_b(\Omega_1)N} + \frac{\sigma_2^2}{P_b(\Omega_2)N}. \quad (17)$$

Let $VCV_e^B(\mathbf{u}^{\mathcal{W}})$ be an estimate of the variance of the r.v. $P_e(\varrho_b)$ and $MCV_e^B(\mathbf{u}^{\mathcal{W}})$ be an estimate of the mean of r.v. $P_e(\varrho_b)$, when the following assumptions are made to simplify the analysis:

- the design test is used for testing,
- $\sigma_{1b}^2, \sigma_{2b}^2, P_b(\Omega_1)$, and $P_b(\Omega_2)$ are invariant through cross-validation repetitions and equal to $\sigma_1^2, \sigma_2^2, P(\Omega_1)$, and $P(\Omega_2)$, respectively.

Accordingly, P_e can be expressed as a function of one r.v., i.e. ϱ_b, and (17) reduces to

$$\sigma^2 = \frac{\sigma_1^2}{P(\Omega_1)N} + \frac{\sigma_2^2}{P(\Omega_2)N} \quad (18)$$

which concludes the proof of Lemma 1. ∎

Henceforth, $P_e(\varrho)$ is treated as a function of the r.v. ϱ.

Theorem 1 *For a singleton feature set (i.e. one that contains only one feature), $VCV_e^B(\mathbf{u}^{\mathcal{W}})$ depends on $MCV_e^B(\mathbf{u}^{\mathcal{W}})$ for two classes (i.e. $C = 2$).*

Proof A qualitative proof will be made through an example that demonstrates the dependence of $P_e(\varrho)$ on ϱ. Let us derive the distribution of $P_e(\varrho_b^A)$, $P_e(\varrho_b^B)$, and $P_e(\varrho_b^C)$ when ϱ_b^A, ϱ_b^B, and ϱ_b^C are Gaussian r.vs. with means $\varrho^A = 0, \varrho^B$, and ϱ^C, respectively and equal standard deviations, as shown in Figure 3.

At the bottom of Figure 3, the pdfs of $\varrho_b^A, \varrho_b^B, \varrho_b^C$ are plotted downwards to maintain readability. In the right side, the pdfs of $P_e(\varrho_b^A), P_e(\varrho_b^B)$, and $P_e(\varrho_b^C)$ are calculated by the projection of $\varrho_b^A, \varrho_b^B, \varrho_b^C$ over the curve $P_e(\varrho)$, when $P_e(\varrho)$ is approximated by a straight line in a small area about ϱ^A, ϱ^B, and ϱ^C. From Figure 3, one can deduce that the variance of $P_e(\varrho_b^A)$ is smaller than the variance of $P_e(\varrho_b^B)$, and moreover, the variance of $P_e(\varrho_b^B)$ is greater than the variance of $P_e(\varrho_b^C)$. Let $P_e(\varrho_b^C) = \alpha_C \varrho_b + \beta$ where $\alpha_C = \tan(\phi_C)$. Then, the variance of $P_e(\varrho_b^C)$, according to the identity $Var(\alpha x + \beta) = \alpha^2 Var(x)$ and (15) is

$$Var\big(P_e(\varrho_b^C)\big) = \alpha_C^2 \sigma^2. \quad (19)$$

It can be seen in Figure 3 that as $\varrho_C \to \infty$, then $\alpha_C \to 0$, which combined with (19) yields

$$\lim_{\varrho^C \to \infty} Var\big(P_e(\varrho_b^C)\big) = 0. \quad (20)$$

So, it can be deduced from (19) and (20) that $Var\big(P_e(\varrho_b^C)\big)$ depends on ϱ^C. ∎

The just described Theorem 1 is extended to

Theorem 2 *$VCV_e^B(\mathbf{u}^{\mathcal{W}})$ is: (i) proportional to s, (ii) inversely proportional to N, and (iii) proportional to $MCV_e^B(\mathbf{u}^{\mathcal{W}})\big(1 - MCV_e^B(\mathbf{u}^{\mathcal{W}})\big)$. On the contrary, $VCV_e^B(\mathbf{u}^{\mathcal{W}})$ depends on neither the cardinality of feature set $\mathcal{W}$, nor the number of classes C, nor the prior probabilities $P(\Omega_c)$.*

Proof Let $\Upsilon_{b\mathcal{T}}^e$ be the number of samples of the test set that are misclassified in one CV repetition. From (6) we simply have

$$\Upsilon_{b\mathcal{T}}^e = \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}} \mathcal{L}[l_i, \eta(\mathbf{y}_i^{\mathcal{W}})]. \quad (21)$$

Let also $Prob\big\{\Upsilon_{b\mathcal{T}}^e = k\big\}$ denote the probability the r.v. $\Upsilon_{b\mathcal{T}}^e$ admits the integer value k at the bth CV repetition. If $P_{e\mathcal{D}}(\varrho_b)$ is the prediction error estimated from the design set at the bth repetition, then it can be inferred that $\Upsilon_{b\mathcal{T}}^e$ follows the binomial distribution

$$Prob\big\{\Upsilon_{b\mathcal{T}}^e = k\big\} = \binom{N_{\mathcal{T}}}{k} \big(P_{e\mathcal{D}}(\varrho_b)\big)^k \big(1 - P_{e\mathcal{D}}(\varrho_b)\big)^{N_{\mathcal{T}} - k}, \quad (22)$$

and therefore

$$Var(\Upsilon_{b\mathcal{T}}^e) = N_{\mathcal{T}} P_{e\mathcal{D}}(\varrho_b)\big(1 - P_{e\mathcal{D}}(\varrho_b)\big). \quad (23)$$

If we assume that

$$MCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{1}{B}\sum_{b=1}^{B} P_{e\mathcal{D}}(\varrho_b), \quad (24)$$

is a better estimate of prediction error than $P_{e\mathcal{D}}(\varrho_b)$, from (23) it can be inferred that

$$Var(\Upsilon_{\mathcal{T}}^e) = N_{\mathcal{T}} MCV_e^B(\mathbf{u}^{\mathcal{W}})\big(1 - MCV_e^B(\mathbf{u}^{\mathcal{W}})\big). \quad (25)$$

Given (6), (8), and (21),

$$VCV_e^B(\mathbf{u}^{\mathcal{W}}) = Var(\frac{\Upsilon_{\mathcal{T}}^e}{N_{\mathcal{T}}}) = \frac{1}{N_{\mathcal{T}}^2} Var(\Upsilon_{\mathcal{T}}^e). \quad (26)$$

Given that $N_{\mathcal{T}} = N/s$, (25), and (26), it is concluded that

$$VCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{s}{N} MCV_e^B(\mathbf{u}^{\mathcal{W}})\big(1 - MCV_e^B(\mathbf{u}^{\mathcal{W}})\big). \quad (27)$$

∎

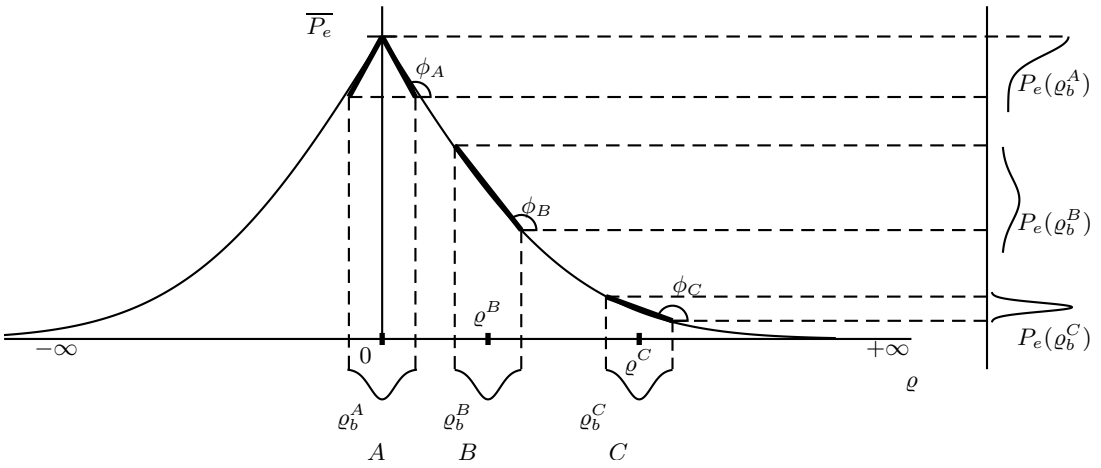

Fig. 3. Prediction error P_e in a two-class problem as a function of ϱ for three cases A, B, C.

The result (27) confirms that $VCV_e^B(\mathbf{u}^{\mathcal{W}})$ depends on $MCV_e^B(\mathbf{u}^{\mathcal{W}})$, as Theorem 1 asserts. From the comparison of (27) derived theoretically, and (9) obtained on experimental basis, it becomes evident that (27) should be multiplied by the factor $\frac{s}{s+\sqrt{2}}$ for $B = \infty$. However, when $s >> 1$ then $\frac{s}{s+\sqrt{2}} \rightarrow 1$. Accordingly, the difference between (9) and (27) becomes negligible. We believe that the factor $\frac{s}{s+\sqrt{2}}$ reflects a relationship between design and test sets.

An accurate estimate of $VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ can be obtained by just employing an estimate $MCV_e^{10}(\mathbf{u}^{\mathcal{W}})$ of $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ with $B = 10$ cross-validation repetitions, i.e.

$$\widehat{VCV}_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \simeq \frac{s^2}{(s+\sqrt{2})N} MCV_e^{10}(\mathbf{u}^{\mathcal{W}})\Big(1-MCV_e^{10}(\mathbf{u}^{\mathcal{W}})\Big). \tag{28}$$

The gains obtained by using (28) in order to estimate $VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ are theoretically derived in Section 4.

4. GAINS OBTAINED BY THE PROPOSED METHOD

Let $VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})$ be the variance of prediction error directly calculated using (8) for B repetitions, and $VCV_{e;prop}^B(\mathbf{u}^{\mathcal{W}})$ be the variance of prediction error estimated by (28), for B repetitions in general instead of 10. In order to show that $VCV_{e;prop}^B(\mathbf{u}^{\mathcal{W}})$ is more accurate than $VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})$, the following gain factor δ is defined

$$\delta \triangleq \frac{Var\big(VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})\big)}{Var\big(VCV_{e;prop}^B(\mathbf{u}^{\mathcal{W}})\big)}. \tag{29}$$

If $\delta > 1$, the proposed method finds an estimate of $VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ that varies much smaller than that of the variance estimate delivered by cross-validation, and therefore the proposed method is better than the straight forward cross-validation. The nominator and the denominator of (29) are derived separately.

Nominator: Since $CV_e^b(\mathbf{u}^{\mathcal{W}})$ is a binomial r.v., it can be approximated by a Gaussian r.v [10], if

$$NMCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big) > 25. \tag{30}$$

According to CLT its variance for B repetitions $VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})$ is a r.v. that follows the χ^2_{B-1} distribution, i.e.

$$\frac{B-1}{VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})} VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}}) \sim \chi^2_{B-1}. \tag{31}$$

Given that $Var(ax) = a^2 Var(x)$, and the fact that $Var(\chi^2_n) = 2n$, where a, n are constants, from (31) we obtain

$$Var(VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})) = 2\frac{\Big(VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2}{B-1}. \tag{32}$$

From (9) and (32), it is inferred that

$$Var\big(VCV_{e;dir}^B(\mathbf{u}^{\mathcal{W}})\big) = \frac{2s^4}{(B-1)(s+\sqrt{2})^2 N^2} \Big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2 \Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2. \tag{33}$$

Denominator: The variance of $CV_e^b(\mathbf{u}^{\mathcal{W}})$ from B repetitions with the proposed method is the function

$$VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}}) = \frac{s^2}{(s+\sqrt{2})N} MCV_e^B(\mathbf{u}^{\mathcal{W}}) \Big(1 - MCV_e^B(\mathbf{u}^{\mathcal{W}})\Big), \tag{34}$$

of the r.v. $MCV_e^B(\mathbf{u}^{\mathcal{W}})$, where according to CLT

$$MCV_e^B(\mathbf{u}^{\mathcal{W}}) \sim \mathcal{G}\Big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}), \frac{VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})}{B}\Big). \tag{35}$$

To derive approximately $Var\big(VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})\big)$, the fundamental theorem governing the transformation of one r.v. [9] is applied. In Figure 4 the function

$$R\Big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big) = \frac{s}{(s+\sqrt{2})N} MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big), \tag{36}$$

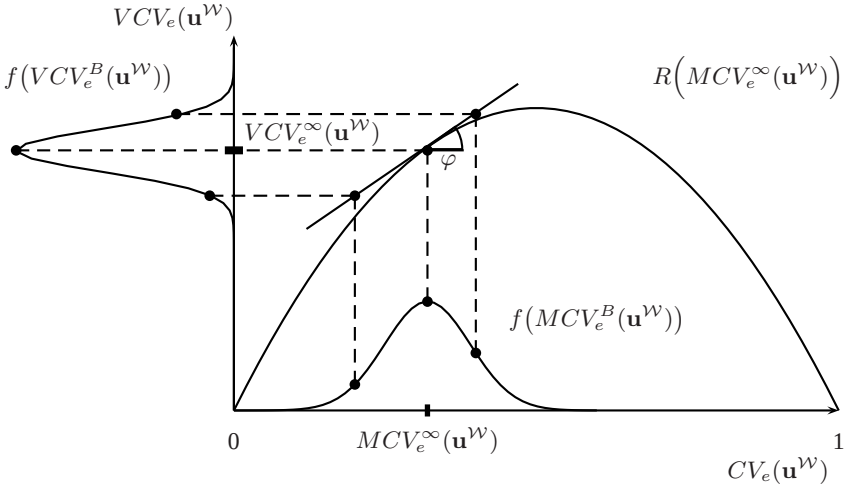

Fig. 4. Approximation of $f\big(VCV_e^B(\mathbf{u}^{\mathcal{W}})\big)$ by using the derivative of the curve R.

is plotted.

The pdf of $MCV_e^B(\mathbf{u}^{\mathcal{W}})$ is plotted on axis x and the pdf of $VCV_e^B(\mathbf{u}^{\mathcal{W}})$ is plotted on axis y. It can be seen that the pdf of $VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})$ is the projection of $MCV_e^B(\mathbf{u}^{\mathcal{W}})$ on the curve $R\big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\big)$. The curve $R\big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\big)$ can be approximated with a straight line $y = \tan(\varphi)x + \beta$ over the area above the pdf of $MCV_e^B(\mathbf{u}^{\mathcal{W}})$, where

$$\tan(\varphi) = \frac{dR\Big(MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)}{dMCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})} = \frac{s^2}{(s+\sqrt{2})N} \Big(1 - 2MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big). \quad (37)$$

Given that $Var(\tan(\varphi)x + \beta) = \big(\tan(\varphi)\big)^2 Var(x)$, from (34) and (37) we obtain

$$Var\Big(VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})\Big) = \frac{s^4}{(s+\sqrt{2})^2 N^2} \Big(1 - 2MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2 Var\Big(MCV_e^B(\mathbf{u}^{\mathcal{W}})\Big). \quad (38)$$

Then, from (9), (35) and (38), we find

$$Var\Big(VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})\Big) = \frac{s^6}{(s+\sqrt{2})^3 N^3 B} MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)\Big(1 - 2MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2. \quad (39)$$

By using (29), (33), and (39), δ is found to be

$$\delta = 2N\frac{B(s+\sqrt{2})}{(B-1)s^2} \frac{MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big(1 - MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)}{\Big(1 - 2MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})\Big)^2}. \quad (40)$$

From (40), it is inferred that the gain factor δ is:

- proportional to the total number of samples N,
- not affected dramatically from the number of cross-validation repetitions B,
- almost inversely proportional to folder number s,
- maximized when $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \to 0.5$, whereas it is minimized when $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \to 0$ or $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \to 1$.

By ignoring B in (40), the gain δ is greater than 1 when

$$0.5 - 0.5\sqrt{\frac{N(s+\sqrt{2})}{N(s+\sqrt{2}) + 2s^2}} < MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) < 0.5 + 0.5\sqrt{\frac{N(s+\sqrt{2})}{N(s+\sqrt{2}) + 2s^2}}. \quad (41)$$

For example, when the number of folders, s, equals 2 and the total number of samples is 1000, it can be inferred from (41) that the gain is greater than 1 if $0.001 < MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) < 0.999$. The gain is smaller than 1 when $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ approaches 0, i.e. the classes are well separated or when $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}}) \to 1$, which means random classification for a great number of classes. Gain values higher than 900 are obtained as $MCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$ tends to 0.5, for any number of classes C. In such cases, $Var\big(VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})\big) \to 0$, and therefore $VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}}) \to VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$, i.e. $VCV_{e;prop.}^B(\mathbf{u}^{\mathcal{W}})$ can be a very accurate estimator of $VCV_e^{\infty}(\mathbf{u}^{\mathcal{W}})$, even for $B = 10$ repetitions.

5. CONCLUSIONS

In this paper, we studied the cross-validation method, when it is applied to obtain an unbiased estimator of the prediction error. On the grounds of experimental findings [7] and the presented theoretical analysis, we derived Eq. (9) that relates the variance of the prediction error with the mean value of the prediction error by employing an infinite number of cross-validation repetitions. The theoretical analysis began with the variance of the prediction error committed by the Bayes classifier using univariate Gaussian class pdfs (Theorem 1) and extended for any dimensionality and any number of classes (Theorem 2). The main result came out of this theoretical analysis, i.e. Eq. (27), indicates that by a multiplicative factor $\frac{s}{s+\sqrt{2}}$ the experimentally derived relationship (9) conforms with the theoretically derived one (27).

Although the proposed equation (9) is valid for an infinite number of cross-validation repetitions, it is proved that the va-

riance of the prediction error, treated as an r.v. itself, exhibits a variance that could be 900 smaller than that delivered by cross-validation even for a finite number of repetitions, say 10. By exploiting the main result of the paper in (28) we succeeded to speed up floating forward feature selection algorithm [11] within the framework of emotional speech classification [7]. The relationships between the sample estimates of the variance and the mean of the prediction error can be extended to other estimates such as the bootstrap estimates.

Appendix

Given that $p(y|\Omega_1)$ and $p(y|\Omega_1)$ are Gaussians, then we prove that the ratios $\frac{t-\mu_i}{\sigma_i}$, $i = 1, 2$ for all the solutions of

$$\frac{P(\Omega_1)}{\sigma_1}\exp\{-\frac{1}{2}(\frac{t-\mu_1}{\sigma_1})^2\} = \frac{P(\Omega_2)}{\sigma_2}\exp\{-\frac{1}{2}(\frac{t-\mu_2}{\sigma_2})^2\}, \tag{42}$$

are always functions of $\mu_2 - \mu_1$.

Proof (42) leads to

$$t^2(\sigma_2^2-\sigma_1^2)+t(2\mu_2\sigma_1^2-2\mu_1\sigma_2^2)+\mu_1^2\sigma_2^2-\mu_2^2\sigma_1^2-2\sigma_1^2\sigma_2^2\Lambda = 0$$

where $\Lambda = \ln\frac{\sigma_2 P(\Omega_1)}{\sigma_1 P(\Omega_2)}$.

- If $\sigma_1 \neq \sigma_2$ and $P(\Omega_1) \neq P(\Omega_2)$, there are two solutions

$$\begin{aligned} t_{1,2} &= \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}, \quad \text{where} && (44)\\ \alpha &= \sigma_2^2 - \sigma_1^2, && (45)\\ \beta &= 2\mu_2\sigma_1^2 - 2\mu_1\sigma_2^2, && (46)\\ \gamma &= (\mu_1\sigma_2)^2 - (\mu_2\sigma_1)^2 - 2(\sigma_1\sigma_2)^2\Lambda. && (47) \end{aligned}$$

  If

$$\beta^2 - 4\alpha\gamma = (2\sigma_1\sigma_2)^2\left\{(\mu_2-\mu_1)^2 + 2(\sigma_2^2-\sigma_1^2)\Lambda\right\} > 0,$$

  then (44) leads to

$$t_{1,2} = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{\sigma_2^2-\sigma_1^2} \pm \sigma_1\sigma_2\sqrt{\frac{(\mu_2-\mu_1)^2}{(\sigma_2^2-\sigma_1^2)^2} + \frac{2\Lambda}{\sigma_2^2-\sigma_1^2}}.$$

  Then

$$\frac{t_{1,2}-\mu_1}{\sigma_1} = -(\mu_2-\mu_1)\frac{\sigma_1}{\sigma_2^2-\sigma_1^2} \pm \sigma_2\sqrt{\frac{(\mu_2-\mu_1)^2}{(\sigma_2^2-\sigma_1^2)^2} + \frac{2\Lambda}{\sigma_2^2-\sigma_1^2}}, \tag{48}$$

  and

$$\frac{t_{1,2}-\mu_2}{\sigma_2} = (\mu_2-\mu_1)\frac{\sigma_2}{\sigma_2^2-\sigma_1^2} \pm \sigma_1\sqrt{\frac{(\mu_2-\mu_1)^2}{(\sigma_2^2-\sigma_1^2)^2} + \frac{2\Lambda}{\sigma_2^2-\sigma_1^2}}, \tag{49}$$

  which are indeed functions of $\mu_2 - \mu_1$.

- If $\sigma_1 = \sigma_2 = \sigma$, there is a single solution

$$t = \frac{-\gamma}{\beta} = \frac{\mu_2+\mu_1}{2} + \frac{\sigma^2}{\mu_2-\mu_1}\ln(\frac{P(\Omega_1)}{P(\Omega_2)}), \tag{50}$$

  and therefore,

$$\begin{aligned} \frac{t-\mu_1}{\sigma} &= \frac{\mu_2-\mu_1}{2\sigma} + \frac{\sigma}{\mu_2-\mu_1}\ln(\frac{P(\Omega_1)}{P(\Omega_2)}), && (51)\\ \frac{t-\mu_2}{\sigma} &= -\frac{\mu_2-\mu_1}{2\sigma} + \frac{\sigma}{\mu_2-\mu_1}\ln(\frac{P(\Omega_1)}{P(\Omega_2)}), && (52) \end{aligned}$$

  which are functions of $\mu_2 - \mu_1$.

- If $\sigma_1 = \sigma_2 = \sigma$ and $P(\Omega_1) = P(\Omega_2)$ then,

$$t = \frac{-\gamma}{\beta} = \frac{-\sigma^2(\mu_1^2-\mu_2^2)}{2\sigma^2(\mu_2-\mu_1)} = \frac{1}{2}(\mu_2+\mu_1) \tag{53}$$

  and the ratios are

$$\begin{aligned} \frac{t-\mu_1}{\sigma} &= \frac{\mu_2-\mu_1}{2\sigma}, && (54)\\ \frac{t-\mu_2}{\sigma} &= -\frac{\mu_2-\mu_1}{2\sigma}. && (55) \end{aligned}$$

From (48), (49), (51), (52), (54), and (55), it can be inferred that the ratios $\frac{t-\mu_i}{\sigma_i}$, $i = 1, 2$, are always functions of $\mu_2 - \mu_1$. ∎

6. REFERENCES

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

[2] B. Efron, "Bootstrap methods: another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.

[3] M. H. Quenouille, "Approximate tests of correlation in time-series," *J. R. Statist. Soc. (Series B)*, vol. 11, pp. 68–84, 1949.

[4] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, 1956.

[5] R. G. Miller, "The jackknife - a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.

[6] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503, 1989.

[7] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, N.Y.: Academic Press, second edition, 1990.

[9] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, N.Y.: McGraw-Hill, 2002.

[10] M. Evans, N. Hastings, and J. B. Peacock, *Statistical distributions*, N.Y. Wiley, 2000.

[11] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pat. Rec. Letters*, vol. 15, pp. 1119–1125, 1994.