Model Selection in Classification: the Swapping Method

Jean-Jacques Daudin and Tristan Mary-Huard

INA-PG (dépt OMIP) / INRA (dépt MIA)
16 rue Claude Bernard, Paris Cedex 05, France
(e-mail: daudin@inapg.fr, maryhuar@inapg.fr)

Abstract. In this article, the bias of the empirical error rate in supervised classification is studied. The exact formula and a robust estimator of the bias are given. From these results, we propose a new penalized criterion to perform model selection in classification. Applications to simulated and real data are presented.
Keywords: Classification, Model Selection, Covariance Penalty.

1 Introduction

The aim of supervised classification is to predict the unknown label Y of an observation (here $Y = 0$ or 1), according to some collected information X. A classifier $\phi_n^* : x \mapsto \phi_n^*(x) = \hat{y}$ is constructed on the basis of a collection of i.i.d. examples $(X_i, Y_i),\ i = 1, ..., n$ for which both the label and the information are known. An important problem is to estimate the conditional error rate (CER)

$$L_x(\phi_n^*) = \frac{1}{n}\sum_{i=1}^{n} P(\phi_n^*(x_i) \neq Y)$$

of the constructed classifier, where the x_i were observed on the training set. A natural estimator of $L_x(\Phi_n^*)$ is the empirical error rate (EER)

$$L_n(\phi_n^*) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi_n^*(X_i)\neq Y_i\}} \quad ,$$

but this estimator is known to be optimistically biased, and we would like to gain insight into the bias of the EER estimator.
In this paper, we study the behavior of the random variable $B(\Phi_n^*) = L_x(\Phi_n^*) - L_n(\Phi_n^*)$, where Φ_n^* is constructed on the basis of an independent copy of the Y_i's, and with the same x_i's as in the initial dataset. We give an exact formula for the bias

$$E_Y\left(B(\Phi_n^*)\right) = E_Y\left(L_x(\Phi_n^*) - L_n(\Phi_n^*)\right) \quad , \tag{1}$$

along with an estimator S_n of $E_Y\left(B(\Phi_n^*)\right)$.
An important motivation for estimating (1) is to perform complexity regularization in pattern recognition. When the CER is close to the true error

rate $P(\phi_n^*(X) \neq Y)$ (TER), it should be relevant to minimize the criterion

$$C(\Phi_n^*) = L_n(\Phi_n^*) + S_n \tag{2}$$

to find a classifier with good generalization performance. We call the minimization of criterion (2) the swapping method (designated by (S)). We analyse the empirical behavior of (S) on a theoretical example, and we compare (S) with cross-validation (CV). We then present the adaption of (S) to the popular k-nearest neighbors algorithm (kNN), where (S) is used to select k. Applications to experimental data are presented to assess the performance of (S).

2 Bias estimation in classification

Let (X_i, Y_i), $i = 1, ..., n$ be n i.i.d. random vectors with distribution P. We note $p_x = P(Y = 1|X = x)$. We define ϕ_n^* as a fixed classification function obtained from a given sample. The "*" indicates that the function was found by optimization of some criterion. We also define Φ_n^* as the corresponding random classification rule obtained for any sample with the same x_is and random Y_is. In practice we would like to obtain some mathematical properties about ϕ_n^* which is the classification function we will use for prediction. However, these properties are difficult to obtain, and we must use Φ_n^* as an intermediate trick.
The following theorem gives the exact form for the bias of the EER in the general classification case:

Theorem 1 *For any classification rule Φ_n^* we have:*

$$E_Y(B(\Phi_n^*)) = \frac{2}{n}\sum_{i=1}^{n} p_{x_i}(1 - p_{x_i})E_Y[\Phi_n^*(x_i|Y_i = 1) - \Phi_n^*(x_i|Y_i = 0)] \ , \tag{3}$$

where $\Phi_n^*(.\ |Y_i = 1)$ is the decision rule computed from the learning dataset with Y_i set to 1.

The proof is not given here. It is worthwhile to interpret this result. The label of each observation is swapped alternatively and the consequence on the decision rule is observed. If the swap does not change the decision for the observation under concern, its contribution to the bias estimate is null. Conversely, if the decision is changed, the contribution is equal to $2p_x(1-p_x)$ with a sign - or +, usually +. Thus if a decision rule is "too versatile" the bias of the EER is high.
From Theorem 1 we can derive an unbiased estimator for the bias of any classification method:

Corollary 1 *With the notations of Theorem 1, an unbiased estimator of* $E_Y(B(\Phi_n^*))$ *is*

$$S_n = \frac{2}{n}\sum_{i=1}^{n} p_{x_i}(1-p_{x_i})[\phi_n^*(x_i|Y_i=1) - \phi_n^*(x_i|Y_i=0)] \ .$$

Of course this estimator is theoretical, since p_x is unknown. Many classification methods provide estimations of the posterior probabilities $\hat{p}_x$ that could be used in place of p_x in Lemma 1. But this method leads to an inconsistent estimation of the bias. We propose a robust version of the plug-in estimator:

$$\hat{p}_{x,B} = \frac{n_x\hat{p}_x + n_0 \times (1/2)}{n_x + n_0} \ , \tag{4}$$

where $\hat{p}_x$ is the plug-in estimator, n_x is the number of points used to compute $\hat{p}_x$ and n_0 is a fixed integer. The "B" index stands for "Bayesian". If $n_0 = 0$, then $\hat{p}_B = \hat{p}_x$ and we find the plug-in estimator. Inversely, if $n_0 = \infty$, then $\hat{p}_B = 1/2$ which corresponds to the worst case in classification
The behavior of the swapping estimate may be closely related to the value of n_0. For high levels of noise in the data and rich classes of classification functions, n_0 should be large. Conversely for low level of noise and poor classes of functions, n_0 should be small. In the following, n_0 is fixed to 10, that seems to be an omnibus compromise (see section 3).

3 Model selection by swapping

3.1 Model selection

Classification aims at finding a classifier ϕ_n^* in a class of functions $\mathcal{C}$ on the basis of data $((X_1,Y_1),...,(X_n,Y_n))$. Of course, we want the TER of ϕ_n^* to be close to the Bayes error rate, i.e. the error rate L^* of the Bayes classifier

$$\Phi^*(x) = \begin{cases} 1 \text{ if } \mathbf{P}\{Y=1|X=x\} > 1/2 \\ 0 \text{ otherwise .} \end{cases}$$

In practice, ϕ_n^* is selected by empirical risk minimization on $\mathcal{C}$. Since we do not know how to choose $\mathcal{C}$, we consider many classes $\mathcal{C}_k$ with different complexities. In the classical complexity regularization framework, the EER minimizer $\phi_{n,k}^*$ is computed for each class. Then among all the candidate classifiers we choose the one that minimizes a given penalized criterion, which usually is an upper bound of the TER.
We propose to use the swapping method (S) to perform model selection. The selection among all the candidate classifiers is performed by minimizing:

$$\begin{aligned} C(\phi_{n,k}^*) &= L_n(\phi_{n,k}^*) + S_n \\ &= L_n(\phi_{n,k}^*) + \frac{2}{n}\sum_{i=1}^{n} \hat{p}_{x_i}(1-\hat{p}_{x_i})[\phi_n^*(x_i|Y_i=1) - \phi_n^*(x_i|Y_i=0)] \end{aligned} \tag{5}$$

While this strategy is also based on the minimization of a penalized criterion, the difference with the preceding strategy is the meaning of the criterion. In (5), the criterion is an estimator of the conditional error risk, while in the regularization framework the criterion is an upper bound for the true error rate. The (S) strategy can be justified with the following break-down:

$$\begin{aligned} L(\phi_n^*) &= L_n(\phi_n^*) + [L_x(\phi_n^*) - L_n(\phi_n^*)] + [L(\phi_n^*) - L_x(\phi_n^*)] \\ &= L_n(\phi_n^*) + B(\phi_n^*) + A(\phi_n^*) \ , \end{aligned}$$

where $A(\phi_n^*) = L(\phi_n^*) - L_x(\phi_n^*)$. In this paper we make the assumption that $A(\phi_n^*)$ does not strongly depend on the complexity of ϕ_n^*, and therefore can be neglected for model selection.

3.2 The Kearn's example

[Kearns *et al.*, 1997] proposed the following model for comparison of model selection methods. The interval $[0, 1]$ is divided into d equal subintervals, alternatively labelled 0 and 1. Let $((X_1, Y_1), ...(X_n, Y_n))$ be an i.i.d. sample, where X_i and Y_i are the position and label of observation i, respectively. The X_i's are drawn from the uniform distribution on $[0, 1]$. Y_i equals the label of the interval to which X_i belongs with probability $1 - \eta$, and the alternative label with probability η. η denotes the noise level of the problem.
We performed simulations according to this model with $d = 10$, $\eta = 0.1$, 0.2, 0.3 and 0.4 and $n = 20, 100, 500$. Simulations performed with $d = 100$ lead to similar findings (not shown).
Figure 1 (left) shows the EER, CER and TER averaged on 100 trials, for $\eta = 0.2$ and $n = 100$, displayed along the number of intervals k. One can see that the curves of the conditional and true error rates are nearly parallel for $k \geq d$. This behavior is observed for any value of η, d and n (data not shown). Therefore the basic condition on $A(\phi_n^*)$ assumed in this paper is satisfied for the Kearns example.
Figure 1 (right) shows the behavior of the estimate of the conditional bias given by the swapping method (S). In this example with $\eta = 0.2$ the bias is overestimated. This overestimation is higher for $\eta = 0.1$, vanishes for $\eta = 0.3$ and becomes an underestimation for $\eta = 0.4$ (not shown).
Figure 2 (left) gives the behavior of the empirical and (S) error rates (y-axis) according to the number of intervals k (x-axis), for 3 trials with $\eta = 0.2$ and $n = 100$. One can see that the empirical error rate decreases to zero. Conversely (S) estimate of the error rate decreases till $k \simeq 10$ and then grows for $k > 10$. Figure 2 (right) shows the mean values of 100 trials of the two error rate estimates with the same parameters as above.

3.3 Comparison between cross validation and swapping

We compared the swapping method selection with $n_0 = 10$ (S) to its natural competitors, the $(n-1, 1)$ cross validation (CV) and the best possible classifi-

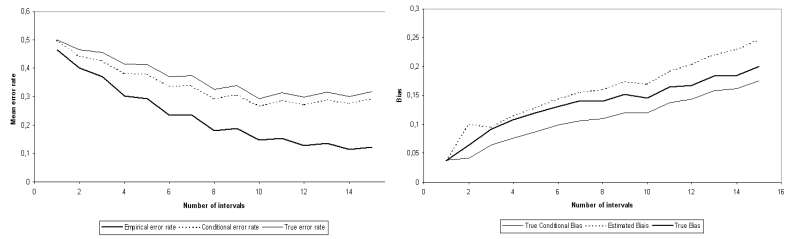

Fig. 1. Left: EER (bold line), CER (dotted line) and TER (solid line) along the number of intervals k. Average on 100 trials, with $\eta = 0.2$ and $n = 100$. **Right:** Estimated bias (dotted line), conditional bias (solid line) and true bias (bold line) along the number of intervals k.

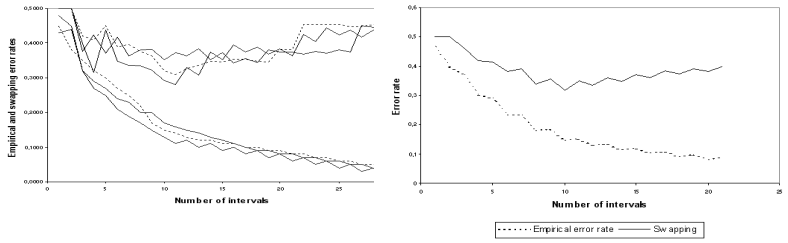

Fig. 2. Left: Empirical and (S) error rates along the number of intervals k for 3 trials. **Right:** Empirical and (S) error rates along the number of intervals k, averaged on 100 trials.

cation function "oracle" (O). (O) is the classification function that minimizes the true error rate for each sample. Figure 3 shows the results for $\eta = 0.2$. Considering Figure 3 and the results obtained for other values of η (not shown here), we draw the following conclusions:

• (S) outperforms (CV) for $\eta \leq 0.3$. The relative gain $(100(L_{CV} - L_S)/(L_{CV} - L_O)$ of (S) on (CV) for $\eta \leq 0.3$ lies between 20% and 80% (not shown here). When $\eta = 0.4$ the gain exists but is tiny.

• The (S) 95% quantile of $L_S - L_O$ is always lower than the (CV) 95% quantile of $L_{CV} - L_O$.

• The empirical error rate penalized by the (S) method gives a better estimate of the true error rate of the selected classification function. This estimate is optimistic for $\eta \geq 0.2$ and pessimistic for $\eta \leq 0.1$. (CV) systematically gives an optimistic view of the true error rate of the selected classification function.

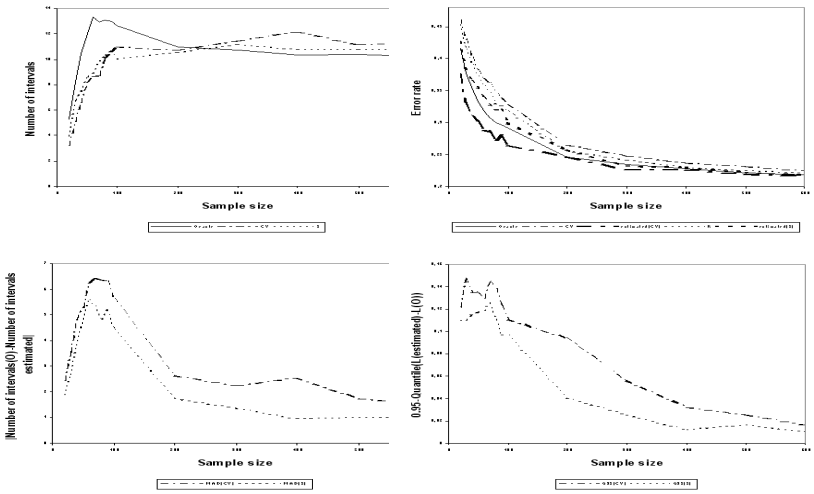

Fig. 3. Results of the (S) model selection for $\eta = 0.2$. **Top left:** Mean number of intervals k_O, k_{CV} and k_S obtained by (O), (CV) and (S), respectively. **Top right:** Mean of the true error rate of the classifiers obtained by (O) selection (solid line), (CV) selection (dashed line) and (S) selection (dotted line), respectively. Dashed lines correspond to (CV) TER estimated by (CV), and (S) TER estimated by (S). **Bottom left:** Mean of $|k_O - k_{CV}|$ and $|k_O - k_S|$. **Bottom right:** 95% quantile of $L_{CV} - L_O$ and $L_S - L_O$.

4 Application to k-nearest-neighbors

We present a simple computational trick to efficiently apply the (S) method to kNN. We then compare the performance of (CV) and (S) on a benchmarking microarray dataset.

4.1 Computation of (S) for kNN

To avoid any concern about the parity of k, in the following we consider only odd values for k, as proposed in [Fort and Lambert-Lacroix, 2004]. For a given k, we need to compute for each observation x_i the quantity $p_{x_i}(1 - p_{x_i})[\phi_n^*(x_i, 1) - \phi_n^*(x_i, 0)]$. The posterior probability p_{x_i} can be estimated according to the Bayes method presented in section 2. In this case, the Bayes estimator of p_{x_i} for the kNN is:

$$\hat{p}_{x_i,B} = \frac{k \times (m/k) + n_0 \times 1/2}{k + n_0} = \frac{m + n_0/2}{k + n_0} \quad , \tag{6}$$

where m is the number of 1 among the k neighbors of point x_i. Clearly, this posterior probability can be obtained from the kNN classifier without

additional computational time.
The difference $\phi_n^*(x_i,1)-\phi_n^*(x_i,0)$ can also be easily obtained from the kNN classifier considering the following argument: when the label of point x_i is swapped, its classification is not changed except in the case where x_i belongs to the majority and the majority is "short", i.e. $m=(k-1)/2$ or $m=(k+1)/2$ (remember that k is odd). Hence, the difference $\phi_n^*(x_i,1)-\phi_n^*(x_i,0)$ will be 1 if $m=(k-1)/2$ or $m=(k+1)/2$, and 0 otherwise. So this difference is easily obtained from the kNN algorithm.
This shows that (S) is a competing method from a computational point of view. In practice, for samples of size $n\sim 100$ and a number of variables as big as 2000, the minimization of the penalized empirical risk to select k is performed within a few seconds.

4.2 Microarray data

We consider the Colon microarray dataset, described in [Alon *et al.*, 1999]. It contains 62 tissue samples for which 2,000 genes were observed. Among the 62 observations, 40 of them are tumor tissues and 22 are normal. For comparison with other published studies, the data normalization, the preliminary gene selection, and the re-randomization study to assess the performance of (S) and (CV) were performed according to the procedures described in [Fort and Lambert-Lacroix, 2004]. It should be noticed that the high level of noise in the data along with the high number of variables considered (with possibly many of them irrelevant) should be in favor of (CV). We display the average performance of the classification rules obtained with (S) and (CV) selection methods.

Table 1 shows that (S) outperforms (CV) for three gene selections: $g=$

	Oracle		Swapping		Cross-Valid.	
Nb. Genes	N	R	N	R	N	R
2000	6.0	19.0	9.11	**28.6**	6.7	28.8
1000	7.6	13.8	12.8	21.4	11.2	**21.1**
500	6.4	13.1	12.1	**18.1**	15.7	18.7
100	4.8	12.0	12.0	**15.6**	20.7	16.0

Table 1. Results for the Colon dataset, over 500 resamplings. First column indicates the number of selected genes. For each selection method (Oracle, Swapping and Cross-Valid.) the mean number of neighbors (N) and the mean test error (R) are computed.

$100, 500, 2000$. As for simulations, both methods are far from the oracle results, even for the simpler case where the number of genes is 100 (which corresponds to the low level of noise case). We conclude that the (S) method for kNN is competing on simulated and real data.

5 Discussion

The methods proposed in this paper to estimate the conditional error rate are connected to some recent papers. A review of the field of prediction error estimation in a quite general context has been made by [Efron, 2004] who divides the methods into two classes: covariance penalties, assuming a parametric model, and nonparametric methods such as cross validation and bootstrap. The swapping method is clearly a covariance penalty method, but it may be applied to non parametric statistical methods. Its only requirement is that a conditional probability $P(Y = 1/X = x)$ may be estimated for each observed value x. This is true because the field is reduced to the error rate in classification, where the p.d.f. of the response variable Y reduces to only one parameter.

The swapping expression in Theorem 1 was present in an earlier paper of [Efron, 1986], but the idea of estimating $E_Y\left(B(\Phi_n^*)\right)$ by its sample estimate (Corollary 1), and the application to model selection in classification are new. Moreover we propose a robust estimate of p_x, which attempts to correct the over-learning bias. In this study n_0 was fixed to 10, but simulations performed with values ranging from 5 to 20 give similar results. However the choice of n_0 is an open problem.

References

[Alon *et al.*, 1999]U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. 96(12):6745–6750, 1999.

[Efron, 1986]B. Efron. How biased is the apparent error rate of a prediction rule? pages 461–470, 1986.

[Efron, 2004]B. Efron. The estimation of prediction error: covariance penalties and cross-validation. 99, 2004.

[Fort and Lambert-Lacroix, 2004]G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 2004.

[Kearns *et al.*, 1997]M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1):7–50, 1997.