ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Structural feature selection for wrapper methods

Gianluca Bontempi
ULB Machine Learning Group
Université Libre de Bruxelles
1050 Brussels - Belgium
email: gbonte@ulb.ac.be

Abstract. The wrapper approach to feature selection requires the assessment of several subset alternatives and the selection of the one which is expected to have the lowest generalization error. To tackle this problem, practitioners have often recourse to a search procedure in a very large space of subsets of features aiming to minimize a leave-one-out or more in general a cross-validation criterion. It has been previously discussed in literature, how this practice can lead to a strong bias selection in the case of high dimensionality problems. We propose here an alternative method, inspired by structural identification in model selection, which replaces a single global search by a number of searches into a sequence of nested spaces of features with an increasing number of variables. The paper presents some promising, although preliminary results on several real nonlinear regression problems.

1 Introduction

Consider a multivariate supervised learning problem where n is the size of the *input* vector $X = \{x_1, \ldots, x_n\}$. In the case of a very large n, it is common practice in machine learning to adopt feature selection algorithms [2] to improve the generalization accuracy. A well-known example is the *wrapper technique* [4] where the feature subset selection algorithm acts as a wrapper around the learning algorithm, considered as a black box that assesses (e.g. via cross-validation) feature subsets. If we denote by $S = 2^X$ the power set of X, the goal of a wrapper algorithm is to return a subset $s \in S$ of features with low prediction error. In this paper we will focus on the expected value of the squared error, also known as *mean integrated squared error* (MISE), as measure of the prediction error.

Since the MISE is not directly measurable but can only be estimated, the feature selection problem may be formulated in terms of a stochastic optimization problem [3] where the selection of the best subset s has to be based on a sample estimate $\widehat{\text{MISE}}$. Consider the stochastic minimization of the positive function $g(s) = E[\mathbf{G}(s)], s \in S$, that is the expected value function of a random function $\mathbf{G}(s) > 0$. Let $G(s)$ be a realization of $\mathbf{G}(s)$ and

$$\hat{s} = \arg\min_{s \in S} G(s) \tag{1}$$

In general terms, coupling the estimation of an expected value function $g(s)$ with the optimization of the function itself should be tackled very cautiously because

of the well-known relation [6]

$$E[\min_{s\in S} \mathbf{G}(s)] = E[\mathbf{G}(\hat{\mathbf{s}})] \leq \min_{s\in S} E[\mathbf{G}(s)] = \min_{s\in S} g(s) = g(s^*) = g^* \qquad (2)$$

where g^* is the minimum of $g(s)$ and $\hat{G} = G(\hat{s}) = \min_{s\in S} G(s)$ is the minimum of the resulting "approximation problem" (1) dependent on the realization G of the r.v. $\mathbf{G}$[1]. This relation states that the minimum of an expected value is optimistically estimated by the minimum of the corresponding sample function. Also, since $\forall \hat{s}, \min_{s\in S} g(s) \leq g(\hat{s})$, we have $\min_{s\in S} g(s) \leq E[g(\hat{\mathbf{s}})]$ and consequently from (2) that

$$E[\min_{s} \mathbf{G}(s)] = E[\mathbf{G}(\hat{\mathbf{s}})] \leq E[g(\hat{\mathbf{s}})]. \qquad (3)$$

This means that the minimum $G(\hat{s})$ of a sample function is also a biased estimate of the average value of the function $g(\cdot)$ in $\hat{\mathbf{s}}$.

Suppose now that, for a given subset of features $s \in S$, $g(s)$ denotes the mean integrated squared error MISE(s) and that $G(s) = \widehat{\text{MISE}}(s)$ is the (almost) unbiased estimate of MISE(s) returned by cross-validation or leave-one-out [1]. The wrapper approach to feature selection aims to return the minimum $\hat{s}$ of a cross-validation criterion $\widehat{\text{MISE}}(s)$.[2] According to the relations above, it follows that the quantity $\widehat{\text{MISE}}(\hat{s})$, returned by the wrapper selection, is a biased estimate both of the minimum MISE(s^*) and of the generalization performance $E[\text{MISE}(\hat{\mathbf{s}})]$. The first contribution of this paper is the adoption of an additional and external cross-validation loop to assess the expected value of MISE($\hat{\mathbf{s}}$) for a wrapper algorithm $\mathcal{W}$. In the following we will denote this estimate by $\widetilde{\text{MISE}}(\mathcal{W})$.

The risk of overfitting cross-validation data was already discussed in [7, 8] and references therein. The paper [7] proposed an alternative to cross-validation to avoid overfitting, called *percentile-cv*. Here we do not intend to propose an alternative criterion to cross-validation, but to stress that whatever the selection process may be, this is not able to return together with a minimum $\hat{s}$ also a reliable assessment of its average performance.

The second contribution of the paper is to propose an alternative to conventional wrappers with the aim to reduce the expected value of MISE($\hat{\mathbf{s}}$). The idea is that, for very large n, a conventional wrapper selection, intended as an intensive search in a huge space S, can return an almost unbiased $\hat{\mathbf{s}}$ but at the cost of a very large variance of $\hat{\mathbf{s}}$. This means that values of $\hat{s}$ far from the optimum s^* can occur with probability higher than zero, thus inducing a large mean value of MISE($\hat{\mathbf{s}}$). A solution based on early stopping was proposed in [5]. Here we propose an alternative approach based on the structuration of the search space S in a sequence of nested spaces $S_1 \subset \cdots \subset S_n = S$, where S_j is the class

[1] Throughout the paper, boldface denote random variables and normal font realizations of random variables.

[2] In this paper we will suppose that the wrapper algorithm is able to return the global minimum of $G(s)$, although in practice, this is known to be a NP-hard problem.

of all the subsets of X having a cardinality less or equal than j. This means that instead of performing a global search in the space of features, the algorithm aims at assessing n different search processes $\mathcal{W}_j$ in spaces $S_j, j = 1, \ldots, n$, of increasing cardinality. If we denote by $\hat{s}_j$ the subset returned by the search $\mathcal{W}_j$ in S_j, the final selection can be obtained either by selecting or combining the predictors based on $\hat{s}_j$ according to the the values $\widetilde{\text{MISE}}(\mathcal{W}_j)$ returned by an external cross-validation loop. The use of an external cross-validation loop was advocated in [10] for comparing alternative selection methods. Here we use it to improve the accuracy of a given selection technique.

The preliminary experimental results on several artificial and real regression tasks in the case of a forward selection search strategy appear to be very promising.

2 Feature selection as a stochastic optimization problem

Consider a supervised regression problem where the training set $D_N = \{\langle X_1, y_1\rangle, \langle X_2, y_2\rangle, \ldots, \langle X_N, y_N\rangle\}$ is made of N pairs $\langle X_i, y_i\rangle \in \mathcal{X} \times \mathcal{Y}$ i.i.d. distributed according to the joint distribution $P(\langle X, y\rangle) = P(y|X)P(X)$. Let us define a *learning machine* by the following components: (i) a parametric class of *hypothesis* functions $h(s, \alpha^s)$ with $\alpha^s \in \Lambda^s$, where $s \subseteq X$, (ii) a *quadratic cost* function $C(y, h) = (y - h)^2$, (iii) an *algorithm* of parametric identification that for a given subset $s \subseteq X$ and a given training set D_N returns a hypothesis function $h(\cdot, \alpha^s_{D_N})$ with $\alpha^s_{D_N} \in \Lambda^s$ such that $\sum_{\langle s,y\rangle \in D_N} C\left(y, h(s, \alpha^s_{D_N})\right) \leq \sum_{\langle s,y\rangle \in D_N} C\left(y, h(s, \alpha^s)\right)$ for all $\alpha^s \in \Lambda^s$. We may formulate the feature selection problem as a stochastic discrete optimization problem [3]

$$\min_{\mathbf{s}\in S} g(s) = \min_{\mathbf{s}\in S} \text{MISE}(\mathbf{s}) = \min_{\mathbf{s}\in S} \left\{ E_{\mathbf{D}_N} \left[\int_{\mathcal{X}} \int_{\mathcal{Y}} (y - h(s, \alpha^s_{\mathbf{D}_N}))^2 dP(y|s) dP(s) \right] \right\} = \\ = \min_{\mathbf{s}\in S} E\left[\mathbf{G}(s)\right] \quad (4)$$

where the mean integrated square error $\text{MISE}(s)$ of the subset $s \subseteq X$ is not observed directly but estimated by $\widehat{\text{MISE}}(s)$.

Let

$$s^* = \arg\min_{\mathbf{s}\in S} g(s), \quad \text{MISE}^* = \min_{\mathbf{s}\in S} g(s) = g(s^*) \quad (5)$$

be the optimal solution of the feature selection problem and the relative optimal generalization error, respectively.

3 The structured wrapper method

Consider a wrapper algorithm $\mathcal{W}$ (e.g. a forward search) which searches in the space S, assesses the subsets s of features according to the cross-validation measure $\widehat{\text{MISE}}(s)$ and that returns the subset

$$\hat{s} = \arg\min_{s\in S} \widehat{\text{MISE}}(s) = \mathcal{W}(D_N) \quad (6)$$

This algorithm can be considered as a mapping from the space of datasets of size N to the space S of subsets of X. Since $\mathbf{D}_N$ is a random variable, the variable $\hat{\mathbf{s}}$ is random too. From Equation (3), it follows that the internal cross-validation measure $\widehat{\text{MISE}}(\hat{s})$, i.e. the minimum attained by $\mathcal{W}$, is a biased estimate of $E[\text{MISE}(\hat{\mathbf{s}})]$. As an alternative to this inaccurate measure, we propose an external cross-validated estimate of $E[\text{MISE}(\hat{\mathbf{s}})]$. We define by $\hat{s}^{-k(i)} = \mathcal{W}(D_N^{-k(i)})$ the feature subset selected by the wrapper algorithm on the basis of $D_N^{-k(i)}$, i.e. the dataset D_N with the $k(i)$ part set aside. We adopt the quantity

$$\widetilde{\text{MISE}}(\mathcal{W}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - h(\hat{s}_i^{-k(i)}, \alpha_{D_N^{-k(i)}}^{\hat{s}^{-k(i)}})\right)^2 \tag{7}$$

to estimate the expected value of $E[\text{MISE}(\hat{\mathbf{s}})]$ for the wrapper algorithm $\mathcal{W}$. According to well-known properties of cross-validation, we can consider the estimator (7) as an almost unbiased estimator of $E[\text{MISE}(\hat{\mathbf{s}})]$ [1]. This estimator can be used to assess the quality of a given wrapper algorithm $\mathcal{W}$ or better to choose between different wrapper strategies. We argue that alternative wrapper strategies, differing for the size of the explored space, the assessment of the subsets and the exploration policy, may produce different distributions of $\hat{\mathbf{s}}$, and consequently different values of $E[\text{MISE}(\hat{\mathbf{s}})]$. The unbiased estimator $\widetilde{\text{MISE}}(\mathcal{W})$ provides an improved measure wrt the internal cross-validation to compare and choose among alternative wrapper strategies.

Here we propose to adopt the estimate $\widetilde{\text{MISE}}$ to perform a structured exploration in the power set $S = 2^X$ in the case of high dimensionality n. The rationale of the approach is that, in case of huge spaces S, the outcome $\hat{\mathbf{s}}$ of an intensive search policy is an unbiased estimator of s^*, but potentially with large variance. In order to better control the bias/variance trade-off we propose, accordingly to what is done in model selection tasks, to structure the space S into a nested sequence of spaces $S_1 \subset \cdots \subset S_n = S$ where $S_j = \{s : |s| \leq j\}$. The approach consists in running in parallel n wrapper strategies $\mathcal{W}_j$, $j = 1, \ldots, n$, each constrained to search in the space S_j. Each wrapper strategy returns a subset $\hat{s}_j \in S_j$, made of $|\hat{s}_j| \leq j$ features. The expected generalization error of each strategy is measured by $\widetilde{\text{MISE}}(\mathcal{W}_j)$.

The outcome of the structured wrapper algorithm can be obtained either by winner-takes-all policy

$$\tilde{s} = \hat{s}_{\tilde{j}}, \quad \text{where } \tilde{j} = \arg\min_{j=1,\ldots,n} \widetilde{\text{MISE}}(\mathcal{W}_j) \tag{8}$$

or by combining the models associated to the best B subsets, e.g. by using a weighted average of their predictions (*generalized ensemble method* [9]). Suppose the estimates $\widetilde{\text{MISE}}(\mathcal{W}_j)$ have been ordered creating a sequence of integers $\{j_b\}$ so that $\widetilde{\text{MISE}}(\mathcal{W}_{j_h}) \leq \widetilde{\text{MISE}}(\mathcal{W}_{j_k})$, $\forall h < k$. The resulting prediction is given by

$$h(X) = \frac{\sum_{b=1}^{B} \zeta_b \, h(\hat{s}_{j_b}, \alpha^{\hat{s}_{j_b}})}{\sum_{b=1}^{B} \zeta_b}, \qquad \hat{s}_{j_b} \subseteq X \tag{9}$$

Dataset	Housing	Cpu	Mpg	Servo	Ozone	Bodyfat	Bupa	Stock	Abalone
N	506	209	392	167	330	252	345	950	4177
n	13	6	7	8	8	13	6	9	10
Site	ML	ML	ML	ML	Breiman	CMU	ML	L. Torgo	ML

Kin_8fh	Kin_8nh	Kin_8fm	Kin_8nm	Bank-8fh	Bank-8nh	Bank-8nm	Bank-8fm
8192	8192	8192	8192	8192	8192	8192	8192
8	8	8	8	8	8	8	8
Delve	Delve	Delve	Delve	Delve	Delve	Delve	Delve

Table 1: A summary of the characteristics of the datasets considered.

where the weights are the inverse of the estimated mean squared errors: $\zeta_b = 1/\widetilde{\mathrm{MISE}}(\mathcal{W}_{j_b})$.

4 Experimental results

We test the effectiveness of the structured approach in the case of forward selection (FS), a well-known example of wrapper approach to feature selection. In order to perform the experiments we choose a Nearest Nearest Neighbor as learning algorithm.

The experimental session is based on 17 regression datasets (Table 1) downloaded from several dataset repository: the Machine Learning Repository[3] (ML), the Leo Breiman ftp site[4], the Luis Torgo Regression dataset repository[5], the DELVE[6] repository and the CMU[7] repository. In order to increase the dimensionality of the problem we add to each dataset a number $2n$ of irrelevant features obtained by randomizing the original features of the dataset.

For each dataset we generate five times a random training subset of $N = 50, 75, 100, 125, 150$ samples, respectively, and we use it to perform conventional forward selection and structured forward selection. The assessment of the feature selection procedures, in terms of squared prediction error, is made on the remaining test samples. The internal cross-validation $\widehat{\mathrm{MISE}}(s)$ and the external cross-validation $\widetilde{\mathrm{MISE}}(s)$ are obtained by leave-one-out, and ten-fold cross validation, respectively. Globally we carried out $17 * 5 = 85$ experiments. The first column of Table 2 reports the number of times that the structured forward selection (SFS) approach outperforms the conventional forward selection (FS) vs. the number of times that FS outperforms SFS. The second column of Table 2 reports the number of times that SFS significantly outperforms FS vs. the number of times that FS significantly outperforms SFS (paired t-test with p-value=0.05). The two rows concern the winner-takes-all (WTA) and the combined (CMB) version of SFS for $B = 10$, respectively.

[3] www.ics.uci.edu/∼mlearn/MLSummary.html
[4] ftp.stat.berkeley.edu/pub/users/breiman
[5] www.liacc.up.pt/∼ltorgo/Regression/DataSets.html
[6] www.cs.toronto.edu/∼delve/
[7] lib.stat.cmu.edu/

Method	SFS vs. FS	SFS vs. FS (paired t-test)
WTA	40-22	25-9
CMB	77-8	66-4

Table 2: Number of times (out of 85 experiments) that SFS (WTA and CMB) outperforms FS vs the number of times that FS outperforms SFS in terms of prediction accuracy.

These results, although preliminary, show the interest for exploring a structured approach to feature selection. The main open issue is, however, how to reduce the additional computational burden due to the outer cross-validation loop. We expect that possible solutions may derive from (i) the adoption of faster methods (e.g. racing and subsampling) for internal and external cross-validation assessment, (ii) the use of a stopping criterion for avoiding wrapper searches in large dimensional spaces based on the behaviour (e.g. local minimum) of the $\widetilde{\text{MISE}}$ quantity.

References

[1] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

[2] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[3] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal of Optimization*, 12:479–502, 2001.

[4] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[5] J. Loughrey and P. Cunningham. Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In *Proceedings of the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2004.

[6] W.K. Mak, D.P. Morton, and R.K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

[7] A. Y. Ng. Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.

[8] A. Y. Ng. On feature selection: Learning with exponentially many irrelevant features as training examples. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

[9] M. P. Perrone and L. N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 126–142. Chapman and Hall, 1993.

[10] J. Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382, 2003.