ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

A multiple testing procedure for input variable selection in neural networks

Michele La Rocca and Cira Perna

Department of Economics and Statistics - University of Salerno
Via Ponte Don Melillo, 84084, Fisciano (SA) - Italy

Abstract. In this paper a novel procedure to select the input nodes in neural network modeling is presented and discussed. The approach is developed in a multiple testing framework and so it is able to take under control the well known data snooping problem which arises when the same sample is used more than once for estimation and model selection.

1 Introduction

When using neural networks the selection of an adequate model is always a hard task, due to the "atheoretical" nature of the tool and its intrinsic misspecification. The problem is not new and several approaches have been proposed in the literature both in a frequentist and Bayesian framework [9, 7] (see also [2] and the papers in the same special issue).

In this paper a strategy for input selection in neural network modeling is proposed. The novel approach is in the same spirit of those based on relevance measures [1] but, to avoid the data snooping problem, familywise error rate is controlled by using a multiple testing scheme [10]. When compared to existing testing solutions, the approach does not require *a priori* identification of a proper set of variables to test, which can often lead to sequential testing schemes and, as a consequence, to loose control over the true size of the test. The sampling distribution of the test statistic involved is approximated by subsampling, a resampling scheme which is able to deliver consistent results under very weak assumptions [5]. The paper is organized as follows. In section 2 neural network modeling is briefly reviewed and the variable selection procedure is presented and discussed. In section 3 the subsampling scheme is described. In section 4 some results on a small Monte Carlo study are reported to show the performance of the proposed approach.

2 Variable selection in neural network modeling

Let $\left\{\mathbf{Z}_i = \left(Y_i, \mathbf{X}_i^T\right)^T\right\}$ be *iid* random vectors of dimension $(d+1)$. The variable Y_i represents a target and it is usually of interest its relationship with the (explanatory) variables $\mathbf{X}_i$. If $\mathbb{E}\left(Y_i\right) < \infty$, then $\mathbb{E}\left(Y_i \,|\mathbf{X}_i\right) = g\left(\mathbf{X}_i\right)$ and we can write

$$Y_i = g\left(\mathbf{X}_i\right) + \varepsilon_i \tag{1}$$

where $\varepsilon_i \equiv Y_i - g\left(\mathbf{X}_i\right)$ and g is a function satisfying general regularity conditions. Clearly, by construction the error term ε_i is such that $\mathbb{E}\left(\varepsilon_i \,|\mathbf{X}_i\right) = 0$.

The function g embodies the systematic part of the stochastic relation between Y_i and $\mathbf{X}_i$. It can be approximated by using the output of a single hidden layer feedforward artificial neural network of the form:

$$f(\mathbf{x}, \mathbf{w}) = w_{00} + \sum_{j=1}^{r} w_{0j} \psi\left(\tilde{\mathbf{x}}^T \mathbf{w}_{1j}\right) \tag{2}$$

where $\mathbf{w} \equiv \left(w_{00}, w_{01}, \ldots w_{0r}, \mathbf{w}_{11}^T, \ldots, \mathbf{w}_{1r}^T\right)^T$ is a $r(d+2)+1$ vector of network weights, $\mathbf{w} \in \mathbf{W}$ with $\mathbf{W}$ compact subset of $\mathbb{R}^{r(d+2)+1}$, and $\tilde{\mathbf{x}} \equiv \left(1, \mathbf{x}^T\right)^T$ is the input vector augmented by a bias component 1. The network (2) has d input neurons, r neurons in the hidden layer and identity function for the output layer. The (fixed) hidden unit activation function ψ is a sigmoidal function.

To select a proper set of input variables, we focus on a selection rule which involves: (i) definition of variable's relevance to the model; (ii) estimation of the sampling distribution of the relevance measure; (iii) testing the hypothesis that the variable is irrelevant.

Following White and Racine [12], the hypotheses that the independent variable X_j has no effect on Y, in model (1) can be formulated as:

$$\frac{\partial g(\mathbf{x})}{\partial x_j} = 0, \forall x. \tag{3}$$

Of course the function g is unknown but we equivalently investigate the hypotheses

$$f_j(\mathbf{x}; \mathbf{w}_0) = \frac{\partial f(\mathbf{x}; \mathbf{w}_0)}{\partial x_j} = 0, \forall x. \tag{4}$$

since f is known and $\mathbf{w}_0$ can be closely approximated. So, if a given variable X_j has no effect on Y we have $\mathbb{E}\left[f_j^2(\mathbf{x}, \mathbf{w}_0)\right] = 0$, where the square function is used to avoid cancelation effects.

In this perspective, the hypothesis that a given set of variables has no effect on Y can be formulated in a multiple testing framework as

$$H_j : \theta_j = 0 \quad vs \quad H_j' : \theta_j > 0, \quad j = 1, 2, \ldots, d. \tag{5}$$

where $\theta_j = \mathbb{E}\left[f_j^2(\mathbf{x}, \mathbf{w}_0)\right]$. So, the problem here is how to decide which hypotheses to reject, accounting for the multitude of tests. In such a context, several approaches have been proposed to control the familywise error rate (FWE), defined as the probability of rejecting at least one of the true null hypotheses. The most familiar multiple testing methods for controlling the FWE are the Bonferroni method and the stepwise procedure proposed by Holm [3]. In any case, both procedures are conservative since they do not take into account the dependence structure of the individual p-values. These drawbacks can be successfully avoided by using a recent proposal by Romano and Wolf [10], suitable for joint comparison of multiple misspecified models.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Each null H_j can be tested by using the statistic,

$$\widehat{T}_{n,j} = n^{-1} \sum_{i=1}^{n} f_j^2 \left(\mathbf{X}_i, \widehat{\mathbf{w}}_n \right) \tag{6}$$

where the parameter vector $\widehat{\mathbf{w}}_n$ is a consistent estimator of the unknown parameter vector $\mathbf{w}_0$. Clearly, large values of the test statistics indicate evidence against H_j.

Now, relabel the hypothesis from H_{r_1} to H_{r_d} in redescending order with respect to the value of the test statistics $\widehat{T}_{n,j}$, that is $\widehat{T}_{n,r_1} \geq \widehat{T}_{n,r_2} \geq \ldots \geq \widehat{T}_{n,r_d}$.

The stepdown procedure begins by testing the joint null hypothesis that all hypotheses H_j are true. This hypothesis is rejected if $\widehat{T}_{n,r1}$ is large, otherwise all hypotheses are accepted. In other words, in the first step the procedure constructs a rectangular joint confidence region for the vector $(\theta_{r_1}, \ldots, \theta_{r_d})^T$, with nominal joint coverage probability $1 - \alpha$. The confidence region is of the form $\left[\widehat{T}_{n,r_1} - c_1, \infty\right) \times \cdots \times \left[\widehat{T}_{n,r_d} - c_1, \infty\right)$ where the common value c_1 is chosen to ensure the proper joint (asymptotic) coverage probability. If a particular individual confidence interval $\left[\widehat{T}_{n,r_j} - c_1, \infty\right)$ does not contain zero, the corresponding null hypothesis H_{r_s} is rejected. Once a hypothesis is rejected, it is removed and the remaining hypotheses are tested by rejecting for large values of the maximum of the remaining test statistics. If the first R_1 relabeled hypotheses are rejected in the first step, then $d - R_1$ hypotheses remain, corresponding to the labels $r_{R_1+1}, \ldots, r_d$. In the second step, a rectangular joint confidence region for the vector $(\theta_{R_1+1}, \ldots, \theta_{r_d})^T$ is constructed with, again, nominal joint coverage probability $1 - \alpha$. The new confidence region is of the form $\left[\widehat{T}_{n,r_{R_1+1}} - c_2, \infty\right) \times \cdots \times \left[\widehat{T}_{n,r_d} - c_2, \infty\right)$, where the common constant c_2 is chosen to ensure the proper joint (asymptotic) coverage probability. Again, if a particular individual confidence interval $\left[\widehat{T}_{n,r_j} - c_2, \infty\right)$ does not contain zero, the corresponding null hypothesis H_{r_j} is rejected. The stepwise process is repeated until no further hypotheses are rejected.

3 The subsampling approximation

The estimation of the quantile of order $1 - \alpha$ is obtained by using the subsampling. The resampling scheme runs as follows. Fix b such that $b < n$ and let $\mathbf{Y}_1, \ldots, \mathbf{Y}_S$ be equal to $S = \begin{pmatrix} n \\ b \end{pmatrix}$ subsets of $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$. Let $\widehat{T}^s_{b,j}$ be the test statistic evaluated at $\mathbf{Y}_s, s = 1, \ldots, S$. Then, for $\mathbf{x} \in \mathbb{R}^d$, the true joint cdf of the test statistics evaluated at $\mathbf{x}$ is given by

$$G_n(\mathbf{x}) = \Pr\left\{\widehat{T}_{n,1} \leq x_1, \widehat{T}_{n,2} \leq x_2 \ldots, \widehat{T}_{n,d} \leq x_d\right\} \tag{7}$$

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

and it can be estimated by the subsampling approximation

$$\widehat{G}_n(\mathbf{x}) = \binom{n}{b}^{-1} \sum_{s=1}^{S} \mathbb{I}\left\{\widehat{T}_{b,1}^s \leq x_1, \widehat{T}_{b,2}^s \leq x_2, \ldots, \widehat{T}_{b,d}^s \leq x_d\right\} \tag{8}$$

where as usual $\mathbb{I}(\cdot)$ denotes the indicator function.

As a consequence, for $D \subset \{1, \ldots, d\}$, the distribution of the maximum of the test statistics, let's say $H_{n,D}(x)$, can be estimated by the empirical distribution function $\widehat{H}_{n,D}(x)$ of the values $\max\left\{\widehat{T}_{b,j}^s, j \in D\right\}$, that is

$$\widehat{H}_{n,D}(x) = \binom{n}{b}^{-1} \sum_{s=1}^{S} \mathbb{I}\left\{\max\left\{\widehat{T}_{b,j}^s, j \in D\right\} \leq x\right\} \tag{9}$$

and the quantile of order $1-\alpha$ can be estimated as

$$\widehat{c}_L(1-\alpha) = \inf\left\{x : \widehat{H}_{n,D}(x) \geq 1-\alpha\right\}. \tag{10}$$

The choice of the subsampling as resampling technique can be justified as follows. Firstly, the method does not require any knowledge of the specific structure of the data and so it is robust against misspecifications, a key property when dealing with artificial neural network models. Moreover, the procedure delivers consistent results under very weak assumptions. In our case, by assuming: (i) $b \to \infty$ in such a way that $\frac{b}{n} \to 0$, as $n \to \infty$, (ii) conditions that guarantee asymptotic normality of $\widehat{\mathbf{w}}_n$ are fulfilled [11], (iii) smoothness conditions on the test statistics $\widehat{T}_{n,j}$ [12], the subsampling approximation is a consistent estimate of the unknown (multivariate) sampling distribution of the test statistics [10]. Observe that, the number of subsets of length b which can be formed out of a sample of size n grows very fast with n. Therefore usually, just B random selected subsets are considered for computing the subsampling approximation.

Clearly, the main issue when applying the subsampling procedure lies in choosing the length of the block, a problem which is common to all blockwise resamplig techniques. Nevertheless, [8] proposed a number of strategies to select b and theorems that ensure that the asymptotic results are still valid for a broad range of choices for the subsample size.

4 Numerical results

To illustrate the performance of the proposed model selection procedure we use simulated data sets generated by models with known structures. The aim is to evaluate the ability of the test procedure to select a proper set of explanatory variables for the given data generating process. For the experimental setup we assume $n = 300$, $b = 100$, $r = 2$, $B = 1000$, $\alpha = 0.05$. The hidden layer size of the neural networks has been determined by using the test procedure proposed by [6] and all neural network models have been estimated by using a square loss function. The simulated data sets have been generated by the following models.

The first model (Model M1) assumes that Y depends on 10 explicative variables $\{X_1, X_2, \ldots, X_{10}\}$ but just variables $\{X_3, X_4, X_5, X_6\}$ are relevant to the model, that is,

$$Y = 3\psi\left(2X_3 + 4X_4 + 3X_5 + 3X_6\right) + 3\psi\left(2X_3 + 4X_4 - 3X_5 - 3X_6\right) + \varepsilon$$

where $\varepsilon \sim N(0, 0.7)$ and ψ is the logistic activation function, $\mathbf{X} = (X_3, X_4, X_5, X_6)^T$ is a vector of multivariate Gaussian random variables with zero mean, unit variance and pairwise correlation equal to 0.5. Clearly, a neural network with logistic activation function, four input neurons and two hidden neurons is a correctly specified model and no misspecification is present.

Table 1: Model M1. Results of the multiple testing procedure ($n = 300$, $b = 100$, $r = 2$, $B = 1000$, $\alpha = 0.05$). Figures in bold refer to the rejection of the corresponding hypotheses H_{r_j}.

j	$\widehat{T}_{n,r_j}$	r_j	$\widehat{T}_{n,r_j} - \widehat{c}_1$	$\widehat{T}_{n,r_j} - \widehat{c}_2$	$\widehat{T}_{n,r_j} - \widehat{c}_3$
1	4.1649	4	**2.8040**	–	–
2	1.0315	5	-0.3295	**0.5303**	–
3	1.0105	3	-0.3505	**0.5092**	–
4	0.9680	6	-0.3930	**0.4667**	–
5	0.0142	8	-1.3468	-0.4871	-0.1836
6	0.0038	7	-1.3571	-0.4975	-0.1940
7	0.0025	9	-1.3585	-0.4988	-0.1952
8	0.0019	10	-1.3590	-0.4993	-0.1958
9	0.0016	2	-1.3594	-0.4997	-0.1962
10	0.0010	1	-1.3599	-0.5002	-0.1967

The results of the multiple testing procedure for variables selection are reported in Table 1. After the first step, the procedure rejects the hypothesis that variable 4 is not relevant and accepts all others hypotheses. At the second step, variables 5, 3 and 6 are recognized as relevant, as well. At the third step, the remaining variables are recognized as not relevant and the procedure stops.

For the second model (Model M2) again, we assume that Y depends on 10 explicative variables $\{X_1, X_2, \ldots, X_{10}\}$ but just variables $\{X_3, X_4, X_5, X_6, X_7\}$ are relevant, that is

$$Y = \left(10\sin\left(\pi X_3 X_4\right) + 20\left(X_5 - 0.5\right)^2 + 10X_6 + 5X_7 + \varepsilon\right)/25$$

where $\mathbf{X} = (X_3, X_4, X_5, X_6, X_7)^T$ is drawn randomly from the unit hypercube.

Again, the procedure is able to correctly identify the set of relevant variables in three steps, as clearly shown in Table 2.

Observe that a multiple step procedure is necessary. In both cases at the first step some variables were incorrectly classified as not relevant to the model.

Table 2: Model M2. Results of the multiple testing procedure ($n = 300$, $b = 100$, $r = 2$, $B = 1000$, $\alpha = 0.05$). Figures in bold refer to the rejection of the corresponding hypotheses H_{r_j}.

j	$\widehat{T}_{n,r_j}$	r_j	$\widehat{T}_{n,r_j} - \widehat{c}_1$	$\widehat{T}_{n,r_j} - \widehat{c}_2$	$\widehat{T}_{n,r_j} - \widehat{c}_3$
1	0.2422	3	**0.1951**	–	–
2	0.2019	4	**0.1548**	–	–
3	0.1750	5	**0.1280**	–	–
4	0.1591	6	**0.1120**	–	–
5	0.0400	7	-0.0070	**0.0354**	–
6	0.0002	1	-0.0470	-0.0045	-0.0020
7	0.0001	2	-0.0470	-0.0045	-0.0020
8	0.0001	8	-0.0470	-0.0045	-0.0020
9	0.00009	10	-0.0470	-0.0045	-0.0020
10	0.00006	9	-0.0470	-0.0045	-0.0020

References

[1] Baxt, W. G., White, H. (1995). Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction, *Neural Computation*, 7, 624–638.

[2] Guyon I., Elisseeff A. (2003) An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3, 1157–1152.

[3] Holm S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, 6, 65–70.

[4] La Rocca M., Perna C. (2005a). Variable selection in neural network regression models with dependent data: a subsampling approach, *Computational Statistics and Data Analysis*, 48, 415–429.

[5] La Rocca M., Perna C. (2005b). Neural network modeling by subsampling in *Computational Intelligence and Bioinspired Systems*, J. Cabestany, A. Prieto and F. Sandoval (Eds.), Lecture Notes in Computer Science 3512, 2005, Springer.

[6] La Rocca M., Perna C. (2006). Resampling techniques and neural networks: some recent developments for model selection, in *Atti della XLIII Riunione Scientifica SIS*, Torino,14–16 giugno 2006, Vol. 1, 231 – 242.

[7] Leray P., Gallinari P. (1999). Feature selection with neural networks, *Behaviormetrika*, 26, 1, 145 – 166

[8] Politis D. N., Romano J. P., Wolf, M. (1999). *Subsampling*, Springer-Verlag, NY.

[9] Priddy K. L., Rogers S. K., Ruck D. W., Tarr, G. L., Kabrisky M. (1993). Bayesian selection of important features for feed-forward neural networks, *Neurocomputing*, 5, pp. 91 – 103

[10] Romano J. P., Wolf M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing, *JASA*, 100, 94–108.

[11] White H. (1989). Learning in artificial neural networks: a statistical perspective, *Neural Computation*, 1, 425–464.

[12] White, H., Racine, J. (2001). Statistical Inference, The Bootstrap, and Neural-Network Modeling with Application to Foreign Exchange Rates, *IEEE Transactions on Neural Networks*, 12, 657–673.