Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 30–42
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Selection of artificial neural network models for survival analysis with Genetic Algorithms

Federico Ambrogi[a,*], Nicola Lama[a], Patrizia Boracchi[a], Elia Biganzoli[b]

[a] *Istituto di Statistica Medica e Biometria "G.A. Maccacaro", University of Milano, via Vanzetti 5, 20133 Milano, Italy*
[b] *Divisione di Statistica e Biometria, National Cancer Institute, via Vanzetti 5, 20133 Milano, Italy*

Available online 8 May 2007

Abstract

In follow-up clinical studies, the main time end-point is the failure from a specific starting point (e.g. treatment, surgery). A deeper investigation concerns the causes of failure. Statistical analysis typically focuses on the study of the cause specific hazard functions of possibly censored survival data. In the framework of discrete time models and competing risks, a multilayer perceptron was already proposed as an extension of generalized linear models with multinomial errors using a non-linear predictor (PLANNCR). According to standard practice, weight-decay was adopted to modulate model complexity. A Genetic Algorithm is considered for the complexity control of PLANNCR allowing to regularize independently each parameter of the model. The ICOMP information criterion is used as fitness function. To demonstrate the criticality and the benefits of the technique an application to a case series of 1793 women with primary breast cancer without axillary lymph node involvement is presented.

Keywords: Regularization; Genetic Algorithms; Competing risks; Neural networks

1. Introduction

In recent years, oncological researchers have focused on patient and disease-specific characteristics to model the dynamic of the cancer and determine the optimal treatment. The quantity of interest in such clinical studies is often the time to failure of a given therapeutic strategy, considering a number of clinically relevant risk factors. A further analysis concerns the impact of the risk factors on the different causes of failure. From a statistical viewpoint this implies the study of the cause specific hazard (CSH) functions for possibly censored survival data (Marubini and Valsecchi, 1995). The estimate of the CSH can be useful to explore the disease dynamics, and possibly to identify patients at different risk of failure.

Such kind of analysis may involve the assessment of non-linear effects on time and covariates and possible high order interactions, thus calling for suitable modeling techniques. Following the above perspective, in the context of discrete time models, a multilayer perceptron (MLP) artificial neural network (ANN) was recently proposed as extension of generalized linear models (GLM) (McCullagh and Nelder, 1989) with multinomial errors using a non-linear predictor (PLANNCR, (Biganzoli et al., 2006)) in analogy with the model for single risk with binary error (Biganzoli et al.,

* Corresponding author. Tel.: +39 02 23903282; fax: +39 02 50320866.
E-mail address: federico.ambrogi@unimi.it (F. Ambrogi).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.05.001

1998, 2002; Efron, 1988). According to standard practice, PLANNCR addresses complexity control through regularized scaled deviance of the GLM model, $\xi(\boldsymbol{\beta})$, with a weight-decay term (Ripley, 1996):

$$\xi(\boldsymbol{\beta}) + \boldsymbol{\beta}^{\mathrm{T}} \Lambda \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the D-dimensional column vector of the MLP parameters, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$, is the matrix of the regularization parameters.

In a non-Bayesian perspective, the standard model selection technique consists in minimizing an estimate of the prediction error obtained through information criteria (Murata et al., 1994), cross validation or bootstrap techniques as functions of a single regularization parameter common to all MLP parameters. Although alternative regularization schemes may be considered (Ripley, 1996), such a technique proved to be effective in most applications. However, in applications requiring the modulation of the smoothing effects over different outcomes, as in the case of competing risks, the adoption of a single regularization parameter could be limiting. In fact, cause-specific risk patterns over time could be different since the different dynamics for events like distant metastases or local recurrences. Therefore, independent regularization across different CSH and cause-specific effects of covariates might be useful.

The objective of this work is the development and assessment of an optimization algorithm to control the complexity of PLANNCR with a set of regularization parameters, one associated with each MLP parameter. In particular, optimization through Genetic Algorithms (GA) is considered. The information complexity criterion ICOMP (Urmanov et al., 2002) is considered as suitable fitness function.

The regularization parameters need to be treated as free parameters of a stochastic optimization problem. This refers to the complex problem of minimization of a function with a random component. In fact, the criterion used for MLP model selection is a random variable since it depends on the sample used and on the multiple local minima of the error surface (Bishop, 1995, p. 260). Parameter estimates might change depending on the initialization values of the training algorithm. A trade-off between the number of retraining for each MLP and the number of iteration of the optimization algorithm is to be faced.

An application to breast cancer patients is provided in order to show the important issues to be considered when determining the best search strategy and the benefits of the GA approach over standard model regularization.

The present paper is organized as follows. In Section 2 the PLANNCR model used to estimate the hazard function for right censored survival data is described. In Section 3 the optimization problem for the selection of the optimal size of the regularization terms is considered. In Section 4 the results from the application of the optimization method to breast cancer survival data, in competing risks framework, is reported.

2. PLANNCR for censored survival data

Given a sample of $i = 1, 2, \ldots, N$ patients, in presence of R competing failure causes, each one with $T^1, \ldots, T^R$ potential times of occurrence which end individual observation time, right censored data are represented by the random vector $(T_i, \delta_i, \delta_i \rho_i)$, where $T_i = \min(C_i, T_i^1, \ldots, T_i^R)$ and ρ_i has realizations $r_i = 1, \ldots, R$. Sample data are therefore represented by vectors $(t_i, d_i, r_i d_i)$ (Marubini and Valsecchi, 1995). By partitioning the time axis into $l = 1, 2, \ldots, L$ disjoint intervals $A_l = (\tau_{l-1}, \tau_l]$, for each lth interval, observed times are grouped on a single point (τ_l). Concomitant information on demographic, clinical, pathological and biological characteristics of the patient or the disease is represented by the covariates vector $\mathbf{x}_i$. Discrete CSHs for competing risks are defined as conditional failure probabilities:

$$\tilde{h}_{lr}(\mathbf{x}_i) = P(T \in A_l, \ \rho = r \mid T > \tau_{l-1}, \mathbf{x}_i). \tag{1}$$

According to the relationship between GLM and grouped time models (Efron, 1988), PLANNCR was introduced as extension, with a non-linear predictor, of GLM with multinomial error (Biganzoli et al., 2006) for the flexible modeling of the CSH functions.

The model can be implemented by an augmented data matrix. This is built by the replication of each subject for each lth time interval in which $t_i \geqslant \tau_l$. The time interval τ_l is included in the covariate vector $\mathbf{x}_i$. The vector of the $R + 1$ response variables is $d_{ilr} = 1$ if the rth event is observed in the lth interval and $d_{ilr} = 0$ otherwise. The $R + 1$ response, d_{ilR+1}, is 1 if the subject is not failed in the interval and 0 otherwise.

The fitting is obtained through the minimization of half the scaled deviance function

$$\xi = \log(L_0/L),$$

where L is the model likelihood and L_0 the value achieved by the perfect fit of the observations (McCullagh and Nelder, 1989).

If subjects are grouped into K cells with equal covariate $\mathbf{x}_k$ vectors, the error function is given by

$$\xi = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{r=1}^{R} \left\{ p_{klr} \log \left[\frac{p_{klr}}{\tilde{h}_{lr}(\mathbf{x}_k)} \right] \right\} \cdot n_{kl},$$

where $p_{klr} = d_{klr}/n_{kl}$ with $d_{klr} = \sum_{i \in k} d_{ilr}$. In this case ξ amounts to half the deviance of a model of $K \times L$ multinomial random variables p_{klr}.

PLANNCR uses multiple output units, one for each cause of failure: $\eta_{lr}(\mathbf{x}_k, \boldsymbol{\beta})$. The multinomial inverse logit link (also called softmax function) is used to estimate $\tilde{h}_{lr}(\mathbf{x}_k)$, as follows:

$$\tilde{h}_{lr}(\mathbf{x}_k) = \frac{\exp(\eta_{lr}(\mathbf{x}_k, \boldsymbol{\beta}))}{\sum_{r=1}^{R+1} \exp(\eta_{lr}(\mathbf{x}_k, \boldsymbol{\beta}))}.$$

In particular a three layer perceptron model with H hidden units was adopted

$$\eta_{lr}(\mathbf{x}_k, \boldsymbol{\beta}) = \beta_0^r + \sum_{h=1}^{H} \beta_h^{\alpha\, r} \alpha_h(\beta_{0h} + \boldsymbol{\beta}_h^{\mathrm{T}} \mathbf{v}_{kl}),$$

where $\boldsymbol{\beta}_h^{\mathrm{T}}$ indicates the transpose of the vector of the parameters connecting the inputs to the hth hidden unit, the additional input for the time interval is included in $\mathbf{v}_{kl} = (\mathbf{x}_k, \tau_l)$ and α_h is the activation function, typically the logistic $\alpha_h(\mathbf{x}_k, \boldsymbol{\beta}_h) = \exp(\beta_{0h} + \boldsymbol{\beta}_h^{\mathrm{T}} \mathbf{x}_k)/(1 + \exp(\beta_{0h} + \boldsymbol{\beta}_h^{\mathrm{T}} \mathbf{x}_k))$.

Model complexity is modulated by means of a weight decay term added to the scaled deviance ξ:

$$\xi^*(\boldsymbol{\Lambda}) = \xi + \sum_{i=1}^{D} \lambda_i \boldsymbol{\beta}_{\mathbf{i}}^2,$$

where D is the number of parameters of the MLP. The MLP parameters are estimated through the minimization of the penalized scaled deviance, $\xi^*(\boldsymbol{\Lambda})$. The minimization of the error function is attained using Broyden, Fletcher, Goldfarb and Shanno BFGS Quasi-Newton algorithm (`nnet`, Venables and Ripley, 2002) using R (R Development Core Team, 2006).

PLANNCR model accounts for complex non-monotonic and non-additive effects, allowing for the joint flexible modeling of the CSH as function of time and covariates. Model results can be represented by two- or three-dimensional conditional surface plots of estimated discrete CSHs as a function of time and of the covariates of interest, after fixing other covariates to pre-specified values (for example, their median values).

3. Choice of the model complexity

In this section the model selection criteria and the optimization heuristics for model selection will be described.

The model complexity choice can be formalized as the following constrained minimization problem:

$$\Lambda_{\mathrm{opt}} = \underset{\substack{\lambda_i \in [a,b] \\ \forall i = 1,\ldots,D}}{\operatorname{argmin}}\ E(\boldsymbol{\Lambda}), \quad E(\boldsymbol{\Lambda}) : [a,b]^D \to \mathbb{R}, \quad a > 0, \tag{2}$$

where $E(\boldsymbol{\Lambda})$ is a model selection criterion. In order to avoid convergence problems a is generally set between 10^{-4} and 10^{-3}. Moreover, if the variables are rescaled between 0 and 1 the conservative upper limit $b = 1$ can be introduced following the suggestions reported in Ripley (1996, p. 159–163).

In applications, the standard approach is to consider a single regularization parameter for all the MLP parameters:

$$\lambda_i = \mu, \quad \forall i = 1, \ldots, D, \quad \mu \in [a, b].$$

In this case the function to be minimized, $E(\mu)$, is simply from $[a, b]$ to $\mathbb{R}$. In order to find the minimum of $E(\mu)$ a simple univariate grid search or Brent's (1973) method can be used.

In order to account for the multiple local minima of the error surface, for each Λ, different MLP parameter initializations for the BFGS Quasi-Newton algorithm were adopted (Ripley, 1996, p. 159) in the MLP training. The mean value of the model selection criterion over the MLP parameter initializations, $\bar{E}(\mathbf{\Lambda})$, is used for the fitness evaluation of $\mathbf{\Lambda}$.

The full optimization in (2), with one regularization parameter for each MLP parameter, was preliminary addressed using a direct search approach with the the down-hill simplex method of Nelder and Mead (1965) (implemented in `Optim` function of R). This is to illustrate that the algorithm does not find the global minimum nor good local minima for arbitrarily chosen starting points. The regularization parameter constraints were enforced making the model selection evaluation function return a 'very big value' when parameters do not satisfy the bounds, thus discouraging the simplex method to explore undesired portions of the parameters space.

In the next subsection the model selection criteria investigated in this work are presented, while in the successive subsection the GA technique is introduced.

3.1. Model selection criteria

The model selection criteria considered in this work rely on information theory and cross validation.

The information criterion approach to model selection was pioneered by Akaike (1973). The derivation of AIC is based on the method of maximum likelihood. When the penalized scaled deviance is used as the fitting criterion, the assumptions on which AIC is derived (the true model is included in the set of candidate models) are no longer tenable and a criterion able to take into account model misspecification has to be used. In particular, the network information criterion (NIC) was proposed (Murata et al., 1994). NIC is computed as

$$\mathrm{NIC} = 2\,\xi^* + 2\,p_{eff},$$

where $p_{\mathrm{eff}} = \mathrm{trace}(GQ^{-1})$, $Q = -E[\partial^2 \xi^*(\mathbf{X}_i, \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\partial\beta^{\mathrm{T}}]$ is the expected Hessian matrix of the log-likelihood and $G = Var[\partial\xi^*(\mathbf{X}_i, \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}]$ is the expected value of the outer product of the score functions.

The matrices Q and G are estimated as training sample averages and evaluated at the fitted values $\boldsymbol{\beta_0}$ (Ripley, 1996). The quantity p_{eff} is familiar in statistical literature as the Lagrange-Multiplier test statistic (Bozdogan, 2000) and can be derived in a very general framework for different error functions (Linhart and Zucchini, 1986). When the model is true and the penalization is not used (AIC hypothesis), the matrices G and Q coincide with the Fisher information matrix, then $\mathrm{trace}(GQ^{-1}) = D$, where D is the number of parameters of the network. Weight decay has the effect of reducing p_{eff} with respect to D (Moody, 1992). According to the original AIC theory, NIC provides an estimate of the expected penalized deviance on new data (generalization error) and is obtained through the Taylor expansion of the generalization error at its minimum. Hence NIC theory relies on a single minimum for the deviance function (Ripley, 1996) and can be unreliable when the regularization levels are small (small λ_i).

The ICOMP criterion (Bozdogan, 1987) provides an additional penalization, useful in the presence of small regularization, based on the interdependences between the parameter estimates. ICOMP was extended to penalized likelihood by Urmanov et al. (2002). The criterion is computed adding to NIC the term:

$$C_1(Q^{-1}) = D \log\left(\frac{v_a}{v_b}\right),$$

where $v_a = (1/D)\sum_{i=1}^{D} v_i$, $v_b = (\prod_{i=1}^{D} v_i)^{1/D}$ and v_i are the singular values of Q^{-1}. In order to compute $C_1(Q^{-1})$, it is necessary to control the positive definiteness of the Hessian matrix. When negative eigenvalues are present (Bishop, 1995, p. 410) a further restart of the training algorithm is attempted. ICOMP was originally proposed in order to select the optimum value of $\mathbf{\Lambda}$ but it is not clear if it can be used to discriminate among different MLP structures (hidden node number).

A different approach to the estimation of the generalization error is cross validation (CV). CV was systematized by Stone (1974) and Geisser (1975). Since the training error functions of non-linear models often contain many local

minima, a refinement of the classic CV procedure must be adopted. The v-fold NCV procedure (Moody, 1994), adopted in this work, follows these steps:

(1) ANN is trained on the entire data set, with estimated parameters $\boldsymbol{\beta}_0$.
(2) Divide the original data set into $v = 1, 2, \ldots, V$ partitions of approximately equal size. Each partition is excluded in turn from the training data.
(3) ANN is retrained using the reduced data set after the exclusion of the v partition, with $\boldsymbol{\beta}_0$ as starting values.
(4) Compute ξ_v on the excluded vth partition on the basis of the new estimates $\boldsymbol{\beta}_0^{(-v)}$ after retraining.
(5) Estimate the generalization error as $\xi_{NCV} = \sum_{v=1}^{V} \xi_v$.

In this work V is set equal to 5. The main disadvantage of NCV is the time required to compute ξ_{NCV} that is generally higher than the time required for the computation of the information criteria.

3.2. GA description

GA (Holland, 1975; Mitchell, 1996) are heuristic optimization algorithms inspired by the mechanism of evolution in nature and are generally described using a biological terminology. The algorithm mimics the evolution of a population of computer representations of the solutions by iteratively applying genetic operators, such as recombination and mutation, to the solutions that have the highest fitness in the population. The selection of the fittest solutions makes the fitness distribution more peaked and skewed from generation to generation (exploitation of the search space), while genetic recombination and mutation contribute to increase the heterogeneity of the fitnesses of the population (exploration of the search space). The proper balance between exploration and exploitation of the search space is crucial for obtaining good performances. In particular, the dimension of the population, N, the selective pressure of the selection mechanism, the number of crossover points, the probability of crossover (p_{cross}) and of mutation (p_{mut}) are the parameters to be controlled to modulate the trade-off exploration–exploitation (Shapiro et al., 1994). The adopted GA implementation is described in the following.

Let $\bar{E}(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}) = (1/S)\sum_{i=1}^{S} E(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{\mathbf{i}})$, where $\alpha \in \{1, \ldots, N\}$, S is the number of different MLP parameter initializations and $E(\boldsymbol{\Lambda})$ the model selection criterion adopted for the model choice.

Algorithm 1. Pseudo-code for GA.

```
Generate P = {Λ_1, ..., Λ_N}
Compute Ē(Λ_1), ..., Ē(Λ_N)
while stopping criteria not met do
  Select P' = argmin_{Λ_i ∈ P} Ē(Λ_i)
  for i = 2 to N do
    Select Λ_a and Λ_b, from P
    Apply cross-over to Λ_a and Λ_b to produce Λ_Child with probability p_cross
    Apply Mutation to Λ_Child with probability p_mut
    P'' = P'' ∪ Λ_Child
  end for
  P = P' ∪ P''
end while
```

3.2.1. Solution coding

The problem is discretized, on a D-dimensional lattice (Chatterjee et al., 1996), following Davis (1991) for the choice of the lattice step. In particular, a string of $22 \times D$ bits is converted in D real numbers in the range $[a, b]$ in the following way:

(1) The string of $22 \times D$ bits is converted to D integers, each between 0 and $2^{22} - 1$.
(2) The integers are converted to real numbers between a and b by multiplying them by $(b-a)/(2^{22}-1)$ and adding a.

The value of a is to be selected in order to avoid convergence problem of the MLP as discussed previously.

3.2.2. GA operators

The selection operator is based on the multinomial probability, where the probability of the solution $\Lambda_\alpha, \alpha \in \{1, \ldots, N\}$, is determined on the basis of the rank of $\bar{E}(\mathbf{\Lambda_\alpha})$ in the population (Mitchell, 1996), i.e. the solution with lowest rank has the highest probability to be selected.

The Gibbs probability proposed by Shapiro et al. (1994) is also adopted:

$$p(\mathbf{\Lambda_\alpha}) = \frac{\exp(-\beta \bar{E}(\mathbf{\Lambda_\alpha}))}{Z}$$

and

$$Z = \sum_{\alpha=1}^{D} \exp(-\beta \bar{E}(\mathbf{\Lambda_\alpha})).$$

The Gibbs probability is unchanged by adding a constant to the fitness function. This is important as the fitness values could be very large. Moreover, the parameter β can be used to control the selection strength. The parameter β is set iteratively equal to the inverse of the standard deviation of the fitness distribution of the set of solutions (Shapiro et al., 1994). In this way, as the standard deviation of the fitness distribution decreases along the iterations, the selective pressure increases.

An implementation of the selection operator that takes into account the variability of the model selection criterion was also investigated. It consists of a weighted version of the rank selection approach, where the weight assigned to each Λ_α is the inverse of

$$Var(E(\mathbf{\Lambda_\alpha})) = \frac{1}{S-1} \sum_{i=1}^{S} (E(\mathbf{\Lambda_\alpha^i}) - \bar{E}(\mathbf{\Lambda_\alpha}))^2.$$

The weighted ranks are computed rescaling the weights so that their sum is equal to the maximum rank (i.e. N) and cumulating them along ordered Λ_α values. The weighted rank of Λ_α is defined as the cumulative weight until Λ_α plus half times the rescaled weight of Λ_α, minus 1. Intuitively, the weighted rank values associated to the Λ_α with small variance (small relatively to the variance of the other Λ_α) tends to be "stretched away" from ranks of contiguous Λ_α values. Conversely, large variance of Λ_α leads to "shrink" ranks values to those of the contiguous Λ_α values. The computation of weighted ranks is implemented in the R function `wtd.rank` in the Hmisc package (Harrell, 2006). The selective pressure obtained with the weighted ranks is stronger than the one obtained with non-weighted ranks.

Elitism is adopted: the best solution of the current population is always put in the successive population.

Two types of crossover operators are tested: uniform (Syswerda, 1989), and the standard single point. The two crossover operators differ in the way they preserve and combine the schemata of the strings encoding the solutions. Uniform crossover can be useful in particular when there are complex search spaces with respect to the dimension of the population (De Jong and Spears, 1990).

The mutation operator can perturb every regularization parameter simultaneously ("macro evolution" strategy, Chatterjee et al., 1996), although the probability of this to happen is negligible considered the values adopted for p_{mut}.

3.2.3. GA parameters setting

A fraction of the initial population was generated at random. The rest of the population was generated by setting $\lambda_i = \mu$, $\forall i = 1, \ldots, D$, where μ is selected in a interval around the optimal value found by the univariate minimization.

The setting of the parameters (N, p_{mut}, p_{cross}) is problem dependent. An optimal implementation would require a careful investigation for the search of the optimal parameters. As the algorithm is computationally expensive, general guidelines from Mitchell (1996) were followed. Experiments were conducted to analyze the behaviour of the GA and verify that our choice of GA operating parameter settings was reasonable. Various GA parameter combinations were tested and the results were compared.

The choice between uniform and single crossover, and between rank and Gibbs selection was performed considering the results obtained from a restart of the algorithm with $N = 50$, $p_{\text{cross}} = 0.7$ and $p_{\text{mut}} \in \{0.003, 0.006, 0.009\}$.

The obtained $\mathbf{\Lambda_{Optim}}$ were used to evaluate the model selection criterion over 50 different MLP parameters initializations. The ICOMP distributions were compared with the Wilcoxon rank test, stratified over the mutation probabilities, paired over the MLP parameters initializations and accounting for multiple testing with the Bonferroni correction.

Subsequently, an analysis of the GA with $N = 50$, uniform crossover set to 0.6, 0.7, 0.8 and mutation probability set to 0.001, 0.003, 0.006, 0.009 was performed. In order to compare the different settings of p_{mut} and p_{cross}, the ICOMP distributions were compared with the Wilcoxon rank test as described above.

According to Schaffer et al. (1989), GA runs with $N = 20$ were implemented. The reduction of the size of the population allows, ceteris paribus, a more precise $\bar{E}(\mathbf{\Lambda})$ estimate by increasing the number of the different MLP parameters initializations. This particular setting was also adopted for the selection based on the weighted ranks. In fact, the variability of $\bar{E}(\mathbf{\Lambda})$ needs to be computed over the greatest possible number of different parameters initializations since the selection procedure depends on this information. Reduction of N generally requires an increase of p_{mut} and p_{cross} (Schaffer et al., 1989) for maintaining a balance between exploration and exploitation of the search space.

The GA and the function for the ICOMP computation were written in R language.

4. Application to breast cancer data

To demonstrate the model selection issues involved with PLANNCR model and the benefits that can be derived from the use of GA, an application in breast cancer is presented and the results are compared with the conventional approach. One thousand and ninety three women with primary resectable invasive breast cancer without axillary lymph node involvement, no radiological or clinical evidence of distant metastasis or synchronous bilateral tumour or a second primary tumour, and without adjuvant treatment were considered. A median follow-up of 127 months and 602 neoplastic events were observed. The cases with these clinicopathological features were selected from some 7000 women with an operable tumour, consecutive with respect to ER and PgR determination who underwent surgery at the Istituto Nazionale Tumori of Milan between January 1981 and December 1986. End points of the analysis were: 183 loco-regional tumour recurrences (LR), 225 distant metastases (DM), 119 contralateral breast cancer (CL) and 75 other second primary tumours. In this application, only the competing risk analysis, focused on the first three types of relapse, was considered. In Biganzoli et al. (2003) a more detailed description of the case series is reported. The input of the model are of three types:

(1) Continuous variables: Age (in years), tumour size (in mm), receptors content for ER and PgR (in fmol/mg cytosol prot.).
(2) Categorical variable: Histologic type: (IDC, ILC, IDC + ILC, other);
(3) Time intervals.

The continuous variables were rescaled between 0 and 1. The value of the covariates were grouped according to the procedure proposed by Gray (1996).

4.1. Standard approach: univariate minimization

In this section the standard approach to modulate model complexity is shown. A unique regularization parameter is considered.

For values $\mu < 10^{-3}$ the training of PLANNCR had convergence problems: the Hessian matrix of the log-likelihood presented negative eigenvalues for a very large number of different MLP parameter initializations. For the full optimization, the lower value of the regularization parameters, a, was then set to 10^{-3}.

PLANNCR models with 8, 10 and 12 hidden units and with different $\mu \in \{0.010, 0.025, 0.05, 0.075, 0.100, 0.250, 0.500, 0.750, 1.000\}$ were evaluated according to the three different model selection criteria considered. The search was then refined in the range [0.100, 0.500]. To account for the multiple local minima of the error surface 10 different MLP parameters initializations, for the BFGS algorithm, were adopted. In Fig. 1 the mean values of NIC, ICOMP and of ξ_{NCV} are reported as a function of the regularization parameter.

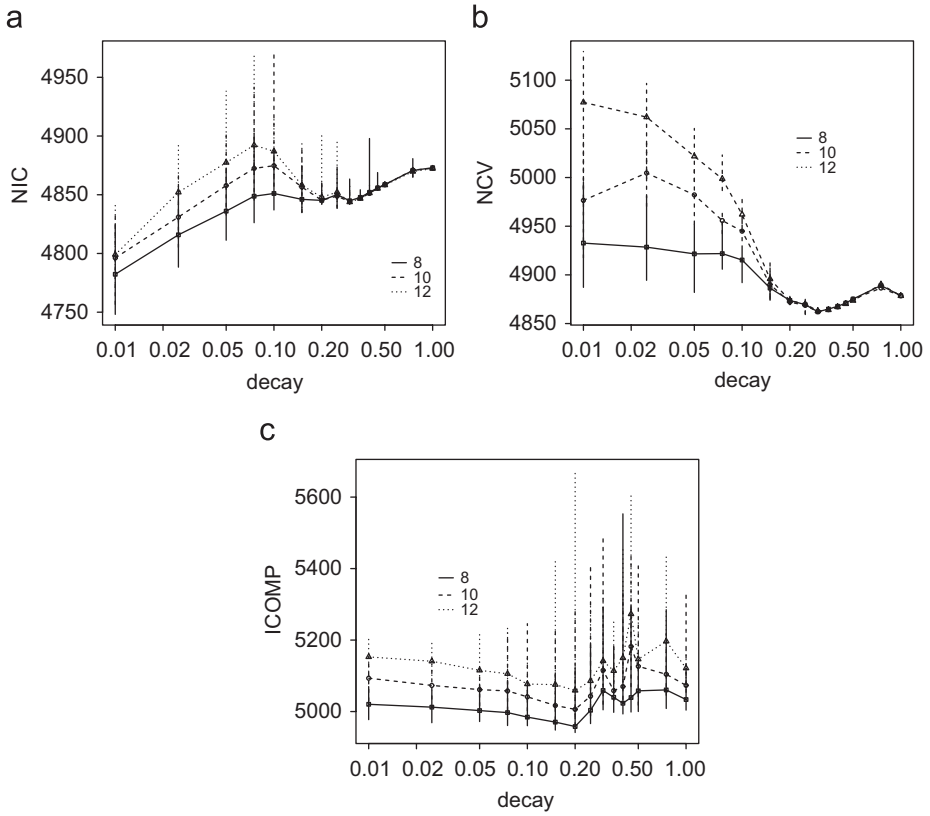

Fig. 1. Univariate Optimization: NIC, ξ_{NCV} and ICOMP criteria as a function of $\lambda \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 1\}$ in log scale.

NIC increases until about 0.100 then decreases until a minimum between 0.200 and 0.300. However, as previously reported, NIC can be not reliable for small λ (Urmanov et al., 2002). Actually, the use NIC as fitness criterion in the GA is difficult as there is no guideline in order to set automatically its limits of validity. For $\lambda > 0.200$ the values of NIC and ξ_{NCV} are similar for the three ANN models. As expected, the complexity of the three ANN architectures appears controlled by the penalty factor. For $\lambda > 0.075$ the minimum of NIC is at 0.300. The minimum (global) is attained at 0.300 also by ξ_{NCV}. The minimum (global) ICOMP value for PLANNCR model is attained at $\lambda = 0.200$ for the three MLP architectures. The computational cost of ξ_{NCV} appeared greater than the costs for computing NIC and ICOMP. The application of the Brent's (1973) method gave the minimum at $\lambda = 0.188$ with a ICOMP value of 4954.88. At last, ICOMP was used as fitness criterion for GA.

PLANNCR with eight hidden units were considered.

In the next subsection the multivariate minimization will be described.

4.2. Multivariate minimization

4.2.1. Nelder–Mead algorithm

The Nelder–Mead algorithm was initialized with $\lambda_i = 0.200$, $\forall i = 1, \ldots, D$. Two different MLP parameter initializations were adopted for the mean evaluation of ICOMP. No solution improvement was registered after about 2500 iterations. The minimum mean value obtained was 4941.381, with a starting value of 4948.770. The optimized penalty vector was practically equal to the starting value: $\lambda_i \approx 0.200$, $\forall i = 1, \ldots, D$. Further parameters starting values, using a Beta distribution centered on 0.200 with variance 0.120, were attempted. The Beta distribution takes values in the range of interest (between 0 and 1) and allows to exploit the prior information gathered from the univariate minimization by setting first and second moments. In particular the variance of the Beta was deter-

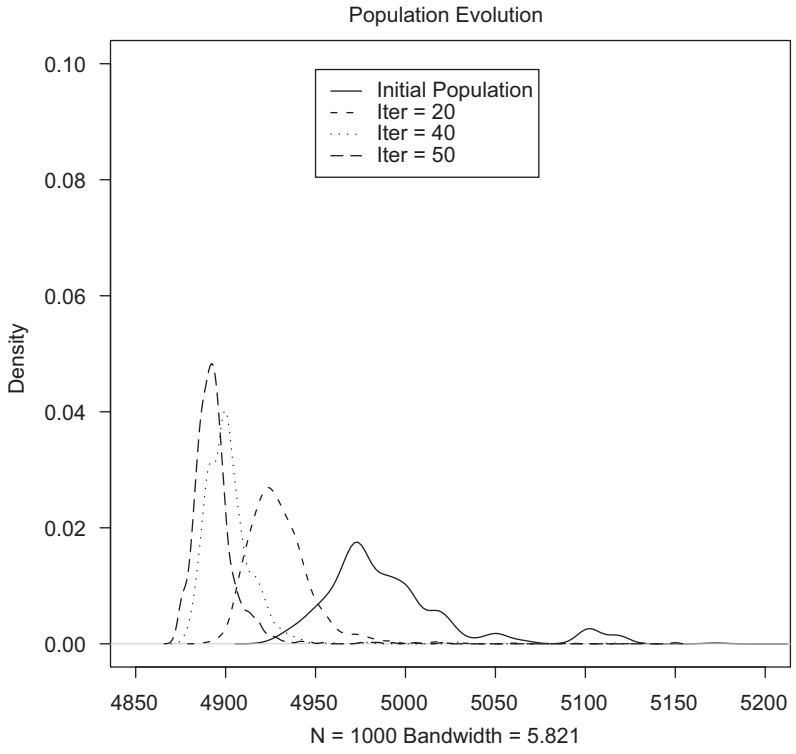

Fig. 2. Distribution of ICOMP values in the GA population of solutions, averaged over 20 restarts of the algorithm, at iterations 0, 20, 40 and 50.

mined in order to have a unimodal distribution. Improvements in the solutions were not evident after about 2500 iterations.

4.2.2. GA algorithm

Improvements of the best solution were generally not observed beyond 50 iterations of the algorithm, with $N = 50$. The GA maximum number of iterations was set to 50 (stopping criterion).

This amounts to 5000 MLP training if two different MLP parameter initializations are adopted for the computation of the mean ICOMP. The time required for training two times, with two different seeds, a MLP and to compute the ICOMP value is generally greater than 60 s on a Pentium IV with 1 GB of RAM memory.

The first experiment with the GA was conducted with rank selection, uniform crossover, $N=50$, $p_{\text{mut}}=0.006$, $p_{\text{cross}}=0.7$ and two different MLP parameter initializations for the BFGS training algorithm. In each restart of the GA, the obtained $\mathbf{\Lambda_{Optim}}$ were used to compute again the mean ICOMP value over 50 different MLP parameter initializations of the BFGS training algorithm. The overall mean value was 4903.892 (1st quartile = 4888.903, 3rd quartile = 4914.234). The inspection of ICOMP values distributions, associated with a particular set of regularization parameters, generally exhibited a positive skewness. The different GA operators were compared using Wilcoxon rank test. Uniform crossover and rank selection were preferred to single crossover and Gibbs selection.

Different setting of p_{mut} and p_{cross} were also tested. It appeared that $p_{\text{mut}} = 0.001$ and $p_{\text{cross}} = 0.7$ outperformed the other settings.

Twenty restarts of the GA with: $N = 50$, $p_{\text{mut}} = 0.001$, $p_{\text{cross}} = 0.7$, rank selection and uniform crossover, were then performed. In Fig. 2 the density estimate (Gaussian Kernel) of the mean ICOMP values in the populations at iterations 0, 20, 40 and 50, computed over the 20 restarts is reported.

The overall mean value is 4899.021 (median = 4892.509, min =4870.972, 1st quartile = 4886.649, 3rd quartile = 4906.303, max =5501.985).

Five restarts of the GA with a reduced population and increased crossover and mutation probabilities ($N = 20$, $p_{\text{mut}} = 0.003$, $p_{\text{cross}} = 0.8$, rank selection and uniform crossover) were performed. The computational time saved by

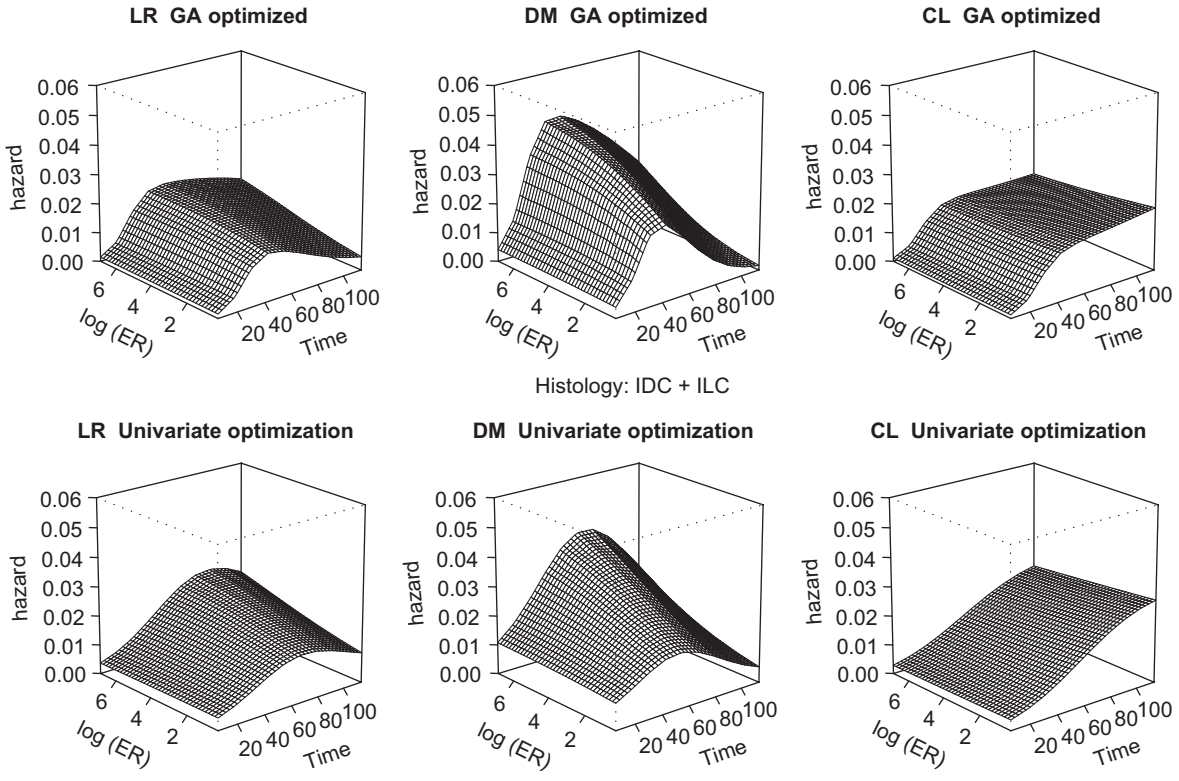

Fig. 3. Estimated CSH as function of $\log(ER)$ and time, after conditioning the value of the other continuous covariates to their median values and histology to IDC + ILC.

the reduced population was allocated to perform five different MLP parameter initializations. The obtained optimum mean ICOMP values did not appear better than those reported with the previous settings.

Selection based on weighted ranks was used to compute other five restarts of the GA. As selection based on ranks realize a greater selective pressure than the standard rank selection, the mutation probability was increased, $p_{\text{mut}}=0.009$, in order to preserve population heterogeneity. Again the results did not appear better than those reported earlier. This might be due to the complexity of the search space (the string has 2376 bits). In this situation, the increase of the crossover and mutation probabilities does not sufficiently increase the exploration of the search space (De Jong and Spears, 1990) to compensate the reduction of N.

A graphical comparison of the estimated CSH obtained with $\lambda_i = 0.188,\ \forall i = 1, \ldots, D$ and with the $\mathbf{\Lambda}_{\text{Optim}}$ obtained in one of the restarts of the *GA* with rank selection and uniform crossover was performed. In Fig. 3 the estimated CSH are reported as function of $\log(ER)$ and time after conditioning the value of the other continuous covariates to their median values and Histology to IDC + ILC.

It is to note that for LR and CL risks the CSH appears increasing at earlier times in the case optimized with GA with respect to the case with $\mu = 0.188$, suggesting an increasing risk after about 2 years from surgery. Moreover, as expected, the risk of distant metastases appears to increase at very early follow-up times using GA optimized parameters.

In Fig. 4 the estimated CSH are reported as function of SIZE and time after conditioning the value of the other continuous covariates to their median values and histology to ILC. The differences in the shape of the CSH for DM in the two cases appears particularly interesting. The effect of SIZE, in the case optimized with GA, appears to have a saturation effect for values greater than 40 mm. This is in accordance with the previous knowledge about the marginal effect of such a covariate. Despite the intrinsic stability of the major pattern across the estimated CSHs in the two cases, it is to note the improved modulation of the shape of the CSHs over time achieved by the GA optimized model.

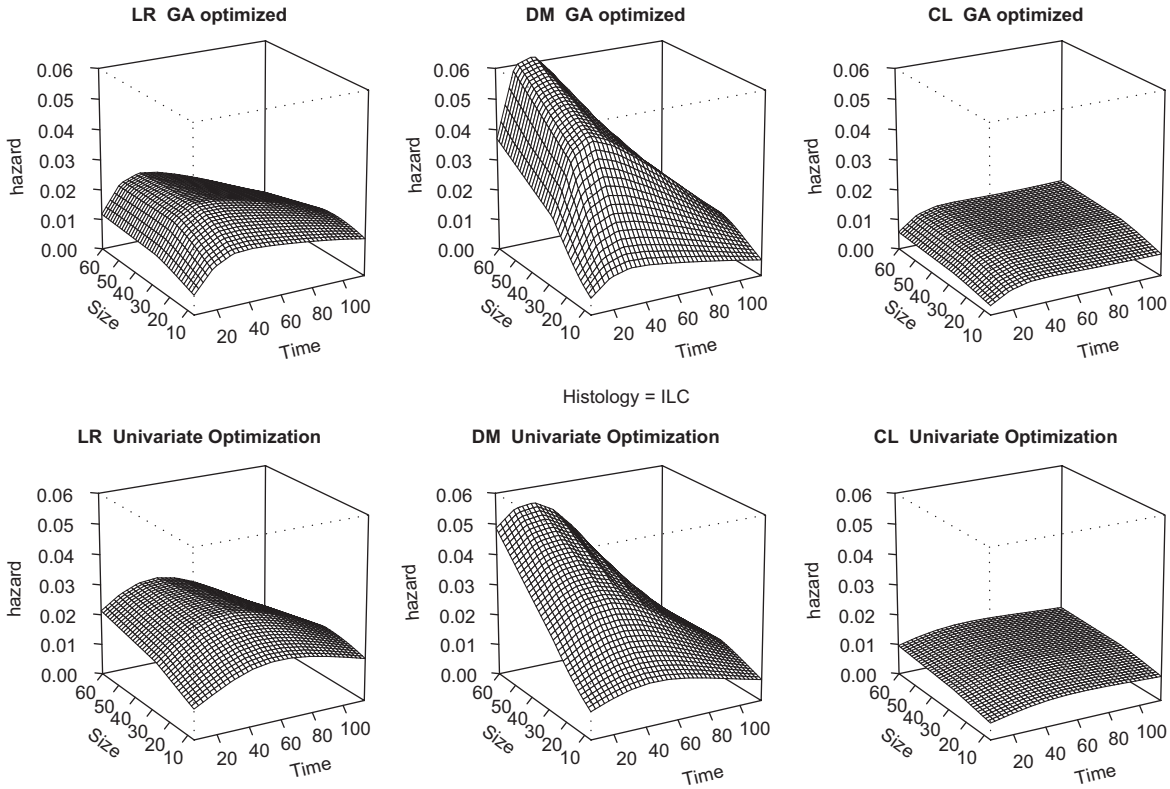

Fig. 4. Estimated CSH as function of SIZE and time, after conditioning the value of the other continuous covariates to their median values and histology to ILC.

5. Conclusions

Model selection is a difficult issue when considering non-linear models involving a large number of parameters, such as MLP.

PLANNCR is implemented using weight-decay regularization in order to modulate model complexity.

Following a Bayesian perspective, the weight-decay term results by assuming the MLP parameter distribution as Gaussian with 0 mean and variance connected to the regularization parameters. In the ARD Bayesian framework (Bishop, 1995), the estimate of the regularization parameters (hyper-parameters) can be obtained resorting to empirical maximum likelihood II (ML-II) approach (Berger, 1985). This is applied in PLANN-ARD (Lisboa et al., 2003), where a common penalty for groups of model parameters corresponding to attributes from a single field or variable is considered. The elegant Bayesian results are obtained at the cost of hyper-parameter distributional assumptions and integral approximations.

In a classical/frequentist view the use of a regularization term can be justified on the basis of the bias-variance trade-off (Geman et al., 1992), or, more generally, in the context of ill-posed problems (Wahba, 1987; Vapnik, 1998).

This work proposes a frequentist model selection technique based on the optimal choice of the regularization parameters by optimizing the ICOMP information criterion through a GA.

Considering the computational cost of an optimization heuristic for PLANNCR, the simplified problem with a unique regularization parameter can be preliminarily explored as usually done in applications.

However, in the presence of complex models, for example in the case of competing risks, the approximation introduced considering a unique regularization parameter may limit the benefits of using a flexible non-linear model. In this case it is decisive to resort to optimization heuristics, when adopting a non-Bayesian perspective, as demonstrated in the presented application. In fact, standard optimization techniques, such as the Nelder–Mead algorithm, cannot be expected

to provide better results than those obtained from the use of a unique regularization parameter. On the contrary, GA optimization appeared effective in minimizing the ICOMP value.

As far as the comparison of different information criteria, ICOMP has the advantage of being computationally less intensive than NCV and is specifically designed to be used also in presence of small regularization (Urmanov et al., 2002). It is interesting to note that NIC (considering $\mu > 0.075$) and NCV show a shared minimum value at about 0.300, while ICOMP reveals a minimum at about 0.200 thus suggesting a less smoothed (biased) model with respect to the former criteria.

As far as the tuning of the optimization heuristics, the optimal balance between exploration and exploitation of the search space is of crucial importance. The superiority, although slight, of uniform crossover and of selection schemes with poor selective pressure is in accordance with the need to explore extensively the search space. This is probably due to the limited size of the population (Chatterjee et al., 1996). Considering the cost of the computation of a model selection criterion, the possibility of trading the dimension of the population with an increase in the exploration of the algorithm seems a good compromise. This is another aspect of the versatility of optimization heuristics with respect to traditional optimization. The adoption of elitism appears crucial in such circumstances.

The adopted GA implementation, with binary strings, has the advantage of treating in a very natural way the constraints of the optimization problem and of a very easy implementation of the mutation operator.

The presented approach is of importance when considering model selection approaches as alternative to the Bayesian one. The use of flexible modeling tools allows the researcher to explore the effect of the covariates on the disease dynamics showing possibly non-linear and non-additive effects that should be prespecified with traditional modelling tools. However, flexible modelling tools can be confidently applied if a fine-tuning of the model complexity is available as in the present context.

Acknowledgements

The authors would like to acknowledge two anonymous referees for their helpful and critical comments and suggestions. This work was partially funded by the Biopattern Network of Excellence FP6/2002/IST/1; proposal N. IST- 2002-508803; Project full title: Computational Intelligence for Biopattern Analysis in Support of Healthcare; URL: www.biopattern.org.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Biganzoli, E., Boracchi, P., Mariani, L., Marubini, E., 1998. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Statist. Med. 17 (10), 1169–1186.

Biganzoli, E., Boracchi, P., Marubini, E., 2002. A general framework for neural network models on censored survival data. Neural Networks 15 (2), 209–218.

Biganzoli, E., Boracchi, P., Coradini, D., Daidone, M.G., Marubini, E., 2003. Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors. Ann. Oncol. 14, 1484–1493.

Biganzoli, E., Boracchi, P., Ambrogi, F., Marubini, E., 2006. Artificial neural network models for the joint modeling of discrete cause specific hazards. Artificial Intelligence Med. 37, 119–130.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52 (3), 345–370.

Bozdogan, H., 2000. Akaike's information criterion and recent developments in information complexity. J. Math. Psych. 44, 62–91.

Brent, R., 1973. Algorithms for Minimization without Derivatives. Prentice-Hall, Englewood Cliffs, NJ.

Chatterjee, S., Laudato, M., Lucy, A.L., 1996. Genetic algorithms and their statistical applications: an introduction. Comput. Statist. Data Anal. 22, 633–651.

Davis, L., 1991. Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York, NY.

De Jong, K.A., Spears, W., 1990. An analysis of the interacting roles of population size and crossover in genetic algorithms. Proceedings of the First International Conference on Parallel Problem Solving from Nature. Morgan Kaufman, Los Altos, CA.

Efron, B., 1988. Logistic regression, survival analysis and the Kaplan–Meyer curve. J. Amer. Statist. Assoc. 83, 414–425.

Geisser, S., 1975. The Predictive Sample Reuse Method with Applications. J. Amer. Statist. Assoc. 50, 320–328.

Geman, S., Bienenstock, E., Doursat, D., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4, 1–58.

Gray, R.J., 1996. Hazard rate regression using ordinary nonparametric regression smoothers. J. Comput. Graph. Statist. 5, 190–207.

Harrell, Jr., F.E., and with contributions from many other users, 2006. Hmisc: Harrell Miscellaneous. R package version 3.1-2. ⟨http://biostat.mc.vanderbilt.edu/s/Hmisc⟩, ⟨http://biostat.mc.vanderbilt.edu/twiki/pub/Main/RS/sintro.pdf⟩, ⟨http://biostat.mc.vanderbilt.edu/twiki/pub/Main/StatReport/summary.pdf⟩.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, MI, USA.

Linhart, H., Zucchini, W., 1986. Model Selection. Wiley, New York.

Lisboa, P.J., Wong, H., Harris, P., Swindell, R., 2003. A Bayesian neural network approach for modeling censored data with an application to prognosis after surgery for breast cancer. Artificial Intelligence Med. 28 (1), 1–25.

Marubini, E., Valsecchi, M.G., 1995. Analysing Survival Data from Clinical Trials and Observational Studies. Wiley, Chichester.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London, pp. 23–25.

Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge.

Moody, J., 1994. Prediction risk and architecture selection for neural networks. In: Cherkassky, V. et al. (Eds.), From Statistics to Neural Networks, NATO ASI Series F. Springer, Berlin.

Moody, J.E., 1992. The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In: Moody, J.E., Hanson, S.J., Lippmann, R.P. (Eds.), Advances in Neural Information Processing Systems, vol. 4, Morgan Kaufmann, San Mateo, CA, pp. 847–854.

Murata, N., Yoshizawa, S., Amari, S., 1994. Network information criterion—-determining the number of hidden units for artificial neural network models. IEEE Trans. Neural Networks 5, 865–872.

Nelder, J.A., Mead, R., 1965. A simplex algorithm for function minimization. Comput. J. 7, 308–313.

R Development Core Team, 2006. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨http://www.R-project.org⟩.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Schaffer, J.D., Caruana, R.A., Eshelman, L.J., Das, R., 1989. A study of control parameters affecting online performance of genetic algorithms for function optimization. In: Schaffer, J.D. (Ed.), Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann, Los Altos, CA.

Shapiro, J., Prügel-Bennett, A., Rattray, M., 1994. A statistical mechanical formulation of the dynamics of genetic algorithms. In: Lecture Notes in Computer Science, vol. 865, pp. 17–27.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. B 36, 111–147.

Syswerda, G., 1989. Uniform crossover in genetic algorithms. In: Schaffer, H. (Ed.), Third International Conference on Genetic Algorithms, vols. 2–9, Morgan Kaufmann, San Mateo.

Urmanov, A.M., Gribok, A.V., Hines, J.W., Uhrig, R.E., 2002. An information approach to regularization parameter selection under model misspecification. Inverse Problems 18, 1207–1228.

Vapnik, V.N., 1998. The Nature of Statistical Learning. Springer, New York.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. fourth ed. Springer, Berlin.

Wahba, G., 1987. Three topics in ill posed inverse problems. In: Engl, M., Groetsch, G. (Eds.), Inverse and Ill-Posed Problems. Academic Press, New York.