Engineering Applications of Artificial Intelligence 41 (2015) 128–138

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

# Regularisation methods for neural network model averaging

Nico Didcock [a,*], Stefan Jakubek [a], Hans-Michael Kögeler [b]

[a] Christian Doppler Laboratory for Model Based Calibration Methodologies, Vienna University of Technology, Getreidemarkt 9, 1060 Vienna, Austria
[b] AVL List GmbH, Hans List Platz 1, 8020 Graz, Austria

ARTICLE INFO

*Article history:*
Received 19 August 2014
Received in revised form
15 January 2015
Accepted 7 February 2015

*Keywords:*
Neural networks
Nonlinear regression
Model ensembles
Empirical risk
Information criteria

ABSTRACT

Data based modeling applications require the complete automation of model creation of, in general, nonlinear processes. Numerous model architectures are available to approximate complex data structures, however, creating and selecting the best model provides a present challenge for application. Model ensembles provide excellent strategies if the prediction performance of the models can be assessed in some way. Information criteria approaches, trading off goodness of fit criteria against the average variance error of the model structures, are theoretically applicable, but demonstratively perform poorly in settings with sparse data sets and complex model structures, settings which are present for many industrial purposes. This paper proposes and discusses a methodology how to regularise the weight calculation building model ensembles. Thereby, the weights trade off goodness of fit criteria against empirical measures of risk, increasing the model ensemble prediction performance. The methodology is independent of the model class, applicable to all black box models, and improves prediction performance for a variety of data in several industrial application areas.



## 1. Introduction

Demands for automation of statistical model building provide challenges for industrial application purposes. Optimal control of industrial processes relies on the accurate modeling of its components. For example, static emission models for combustion engines are estimated in order to effectively calibrate the control variables, such as injection timing or exhaust gas recirculation, to achieve optimal $NO_x$ and soot emissions during a driving cycle. In general, one is interested in the causal behavior of measured, real valued outputs, $y$, as a function of some $d$-dimensional inputs, $\boldsymbol{u}$, where, for simplicity, we restrict the discussion to one-dimensional outputs. This so-called *deterministic component* of the process, $f(\boldsymbol{u})$, say, is of interest for future predictions and approximated by means of *nonlinear regression models*. The *error component*, $y(\boldsymbol{u})-f(\boldsymbol{u})$, captures those components that are unpredictable by the inputs, especially stochastic components such as measurement noise, $p$, or bias due to explanatory limitations.

This paper investigates the setting when a set of black box models has been trained and it is up to the user to assess their plausibility, but, possibly, little information on the model architecture is available. One may think of a setting, where the user knows the training data and some software builds neural networks of unknown structure approximating the data. The presumably best performing model can then be selected, or a new *model ensemble* may be created as a weighted sum of the trained models, see Buckland et al. (1997), Hansen (2007), Burnham and Anderson (2002), and Claeskens and Hjort (2008). Opposed to the existing literature it will be demonstrated that, for a wide range of physical processes, a set of model ensembles can be created using no information on the structure of the model at all. This can be done building a Pareto frontier trading off training error against *empirical* measures of risk. These model ensembles tend to perform particularly well when the data is *sparsely distributed* in the input space. This is a particular characteristic for industrial modeling, where outputs are to be estimated in areas where data cannot be generated, or when the sample size is small. For example, data allocation for engine emission models is often costly since the number of test beds is limited, moreover, accurate measurements are often time intense. If a statistical model of sufficient prediction quality has been estimated, the measurement procedure may be stopped. In this sense, the amount of necessary data becomes *as small as possible*.

Consider one or several model architectures (or classes), such as neural networks or kriging techniques, to choose from. Within each model class a number of hyperparameters are selected, such as the number of neurons and hidden layers in neural networks, or some kernel function for kriging models. Feedforward neural network structures provide powerful tools since they are able to approximate smooth functions arbitrarily

* Corresponding author. Tel.: +43 1 58801 325532.
*E-mail addresses:* nico.didcock@tuwien.ac.at (N. Didcock), stefan.jakubek@tuwien.ac.at (S. Jakubek), Hans-Michael.Koegeler@avl.com (H.-M. Kögeler).

http://dx.doi.org/10.1016/j.engappai.2015.02.005
0952-1976/

well (Hornik et al., 1989), however, the degree of approximation is subject to the number of neurons in the network. Highly complex structures are prone to overfit small sample data since the parameter variance is large in this setting (Geman et al., 1992). Automatisation of data modeling requires a reliable assessment of each of the candidate model's *prediction performance*. The ensemble weight corresponding to each model is chosen according to its (relative) expected performance. It will be shown that assessment of predictive performance will be inaccurate, if (a) training data is sparse or (b) information on the model structure is not available. It follows, that, the predictive performance of model ensembles will be poor in these settings.

Cross validation (CV) provides a technique to evaluate predictive performance of each estimated model (Stone, 1974). Recently proposed ensemble methods include *bootstrap averaging* (bagging) and *boosting*. These algorithms generate a set of predictors by retraining the models, using random subsets of the training data. For discussion, see e.g. Breiman (1996) and Kotsiantis et al. (2006). These methods require that the models are retrained several times, which might become highly computationally expensive, and bagging therefore has not been considered any further for this reason. Information criteria (IC) approaches provide model quality indicators, and are faster to compute than the CV or bagging procedures. Here, the models are ranked according to IC score values, which combine goodness of fit-indicators with potential risk estimates. Usually, model training minimises a loss function such as the negative log likelihood or the quadratic loss with respect to available parameters. The obtained loss value will be a biased estimator of the true loss, which is the predictive performance if the model was used for future predictions. Standard IC procedures penalise the estimated loss with the *average* loss bias of the model class. Classic model selection criteria include the *Akaike Information Criterion* (AIC) (Akaike, 1973), Mallows $C_p$ (Mallows, 1973), and Schwarz's *Bayesian Information Criterion* (BIC) (Schwarz, 1978). While the first two have been shown to be asymptotically equivalent (Nishii, 1984), and optimal (Shibata, 1981), for linear regression models, the BIC is known to be consistent (Sin and White, 1996). The classic information criteria for linear models penalise the estimated loss with functions of the *number of parameters* that give approximations to the loss bias. The statistical theory has been discussed for general, nonlinear model frameworks in Jones (1983) and for neural networks specifically in Murata et al. (1994), Moody (1994), White (1989), and Geman et al. (1992). Once IC score values have been computed for the candidate models, it is advantageous to average the estimators according to the scores. The two mostly used model averaging methods are the exponential Akaike weights (Burnham and Anderson, 2002; Claeskens and Hjort, 2008) and weights mininising Mallow's equation (Hansen, 2007). The latter weighting scheme performs asymptotically optimal in the sense of Li (1987) for homoscedastic linear regression models. Li's concept of optimality, as well as the Bayesian consistency property, is unquestionably desirable but *asymptotic* properties. It has been demonstrated in Anders and Korn (1999) that IC procedures perform badly compared to statistical testing procedures when the network models are overparameterised. Training a large number of parameters of a complex model structure is particularly problematic if the sample size is small. In this case, high variability of the parameter estimates leads to bad prediction performance and, even if the true model structure is known, it may be desirable to use smaller models for prediction.

This paper investigates methods to rank, compare and average models with as little knowledge on the model structure as possible, and without retraining any network. For industrial modeling, such as chemical reaction models, etc., one can often expect the resulting function of interest to be smooth. The weight calculation then should reflect this aversion to non-smoothness. Not the model class itself should be penalised, but structure of the estimated model. Prediction error estimates, and therefore, ensemble weight calculation, perform badly when the training data is *sparse*, i.e. data is not available in prediction areas. Here, it will be demonstrated that error prediction can be improved, if the empirical discrepancy between the model at the data and the area of interest for prediction is taken into account. These empirical complexity criteria can be computed for any black box model such that the ensembles can be built for arbitrary model classes. It will be demonstrated that the methodologies outperform standard methods, especially the small data setting. The methods are validated rigourously using several, partially publicly available, data sets from various fields, including industrial emission data, chemical, and medical data.

The outline of the paper is as follows: Section 2 revises model selection criteria, including the AIC and Mallow's equation, with emphasis on the application to nonlinear models structures, and the underlying assumptions. Regularised model averaging with empirical complexity measures is introduced and discussed in Section 3. A neural network architecture used in the experimental setup is briefly introduced in Section 4. Section 5 presents results for real data and simulations verifying the proposed concepts on artificial data. Section 6 concludes.

## 2. Model assessment

This section reviews the AIC formula, Mallow's equation, and existing attempts to generalise the formulas to nonlinear model structures. The model estimates will be denoted as $\hat{y}_j(\cdot|\boldsymbol{\theta}_j)$, where $j=1\ldots M$ and $\boldsymbol{\theta}_j$ denotes the model specific parameter vector that is optimised to fit the data, e.g. regression coefficients. The models may be of arbitrary structure. For example, the model $\hat{y}_1(\cdot|\boldsymbol{\theta}_1)$ may denote a polynomial, and $\hat{y}_2(\cdot|\boldsymbol{\theta}_2)$ a feedforward neural network. For the proposed purposes, it will only be assumed that the maps $(\boldsymbol{u},\boldsymbol{\theta}_j)\mapsto\hat{y}_j(\boldsymbol{u}|\boldsymbol{\theta}_j)$ are sufficiently differentiable and all occurring integrals exist. A model ensemble is defined as a weighted combination of the models

$$\hat{y}(\boldsymbol{u})=\sum_{j=1}^{M} w_j\cdot\hat{y}_j(\boldsymbol{u}|\boldsymbol{\theta}_j). \tag{1}$$

The weights $w_j$ reflect relative *expected performance*. Most of the present information criteria are derived in the following manner: the parameters are assumed to have been trained to minimise some loss function for a given set of data $\mathcal{D}=\{[\boldsymbol{u}_1,y_1],\ldots,[\boldsymbol{u}_N,y_N]\}$, such as

$$\ell_N(\boldsymbol{\theta}_j)=\begin{cases}-\log\ P_N(\mathcal{D}|\boldsymbol{\theta}_j) & \text{negative log}-\text{likelihood}\\ \mathrm{MSE}_{\mathcal{D}j}=\sum_{n=1}^{N}\frac{1}{N}(\boldsymbol{y}_n-\hat{\boldsymbol{y}}_j(\boldsymbol{u}_n|\boldsymbol{\theta}_j))^2 & \text{quadratic loss evaluated for training data}\end{cases} \tag{2}$$

For Gaussian, independent and identically distributed (i.i.d.) error components, the likelihood is defined as

$$P_N(\mathcal{D}|\boldsymbol{\theta}_j)=\prod_{n=1}^{N}(2\sigma^2\pi)^{-1/2}\exp\left(-\frac{(y_n-\hat{y}_j(\boldsymbol{u}_n|\boldsymbol{\theta}_j))^2}{2\sigma^2}\right), \tag{3}$$

where $\sigma^2$ denotes the variance of the error component. The loss function is a biased approximation of the (hypothetical) long run performance, $\ell(\cdot)=\lim_{N\to\infty}\ell_N(\cdot)$, which is the actual target function. The training error, $\mathrm{MSE}_{\mathcal{D}j}$, is an approximation of the expected out of sample prediction error, $\mathrm{MSE}_{\mathrm{pred}\,j}$, that can be expressed via the $L^2$ norm, $\|\cdot\|_{\xi}^2$, corresponding to an input