MODEL SELECTION FOR SYSTEM IDENTIFICATION BY MEANS OF ARTIFICIAL NEURAL NETWORKS

H. Neuner

Geodetic Institute, Leibniz University Hanover
Nienburger Straße 1, 30167 Hanover, Germany - (neuner@gih.uni-hannover.de)

KEY WORDS: Artificial Neural Networks, Cross-validation, Saliency of weights, Model capacity

ABSTRACT:

System identification is one main task in modern deformation analysis. If the physical structure of the monitored object is unknown or not accessible the system identification is performed in a behavioural framework. Therein the relations between input and output signals are formulated on the basis of regression models. Artificial neural networks (ANN) are a very flexible tool for modelling especially non-linear relationships between the input and the output measures. The universal approximation theorem ensures that every continuous relation can be modelled with this approach. However, some structural aspects of the ANN-based models, like the number of hidden nodes or the number of data needed to obtain a good generalisation, remain unspecified in the theorem. Therefore, one faces a model selection problem.

In this article the methodology of modelling the deformations of a lock occurring due to water level and temperature changes is described. We emphasize the aspect of model selection, by presenting and discussing the results of various approaches for the determination of the number of hidden nodes. The first one is cross-validation. The second one is a weight deletion technique based on the exact computation of the Hessian matrix. Finally, the third method has a rigorous theoretical background and is based on the capacity concept of a model structure. The results of these methods are compared from the viewpoint of generalisation.

1. INTRODUCTION

In a modern approach the monitored object is regarded as a system (Welsch et al., 2000). The loads acting on the object represent the input to the system and its reaction by deformation the output. The dynamic deformation modelling aims at the time-related description of the causal chain consisting of the input, the system and the output.

In engineering geodesy the properties of the monitored system are typically derived from synchronous observations of the input and output measures in the framework of the so called experimental system identification. The modelling equations describing the system properties can be formulated based on physical principles or in a purely mathematical way. This paper addresses the latter, the so called behavioural approach. Therein the relationship between the systems input and output is expressed by mathematical functions. The coefficients of these functions describe the object properties but have no or very restricted physical meaning. The generality of the models structure makes it applicable to a large number of different object types.

The traditional way of system identification in the behavioural approach is to model the actual deformation state y_k as a linear combination of present and past values of the loads $x_k,\dots, x_{k-q}$ respectively, and past deformation states y_{k-p}:

$$\begin{aligned} y_k = a_0 + a_1 y_{k-1} + a_2 y_{k-1} + \dots + a_p y_{k-p} + \\ + b_0 x_k + b_1 x_{k-1} + b_2 x_{k-2} + \dots + b_q x_{k-q} \end{aligned} \qquad (1)$$

The characteristics of the monitored object are captured by the coefficients a_i and b_j, with i = 0,…, p, j = 1…q.

In every model structure derived from (1) the parameters p and q that define the order of the model need to be chosen prior to the estimation of the coefficients. Therefore one faces at first a problem of model selection.

This paper treats the problem of model selection in the more general framework of nonlinear model structures defined by Artificial Neural Networks (ANN). This model structure is used here because it includes, as shown by Neuner and Kutterer (2010), all the traditional modelling strategies, like in eq. 1.

The main aspects related to system identification by means of ANN are presented in the chapter 2. Three strategies for solving the model selection task will be presented in the chapter 3: the cross-validation approach, the saliency of weights method and a theoretically founded approach based on the concept of model capacity. The results obtained for the modelling of the deformations of a lock due to influences of water level changes and temperature with these strategies will be described in the last chapter of the paper.

2. ARTIFICIAL NEURAL NETWORKS

ANN is a model structure with processing units, so called nodes, organised in layers. The minimal configuration of a meaningful ANN consists of an input and an output layer. The number of nodes in these layers corresponds to the number of observed acting loads and to the number of the deformation components respectively. Therefore, these layers can be considered as fixed with respect to the model's structure. Further processing units can be included in the model in so called hidden layers. The number of hidden layers and the number of units contained in them are variable with respect to the model's structure and have to be set-up in accordance with the modelling task.

The nodes of subsequent layers are connected. The strength of the connection between the kth node in layer L and the ith node from the previous layer (L-1) is expressed by the weights $w_{ki}^{(L)}$. These are the unknown parameters of the model that need to be estimated from the observed data.

In this paper only structures of ANN are considered where the information processing is done in just one direction through the network. Such structures are called feed-forward networks. The use of such network architectures implies that phase differences between the input and deformation measures are calculated prior to the modelling with ANN and therefore, that the input and output data series are aligned in time.

The model structure described by an ANN is exemplified in eq. 2. The output $y_1^{(2)}$ from a network consisting of $N_I = 2$ input units, N_H units organised in one hidden layer and $N_O = 1$ output unit results according to:

$$\begin{aligned} y_1^{(2)} &= \varphi^{(2)}\left(\sum_{j=1}^{N_H} w_{1j}^{(1)} y_j^{(1)} + b_2\right) \\ &= \varphi^{(2)}\left\{\sum_{j=1}^{N_H} w_{1j}^{(1)} \varphi^{(1)}\left[\sum_{i=1}^{2} w_{ji}^{(0)} x_i + b_{i1}\right] + b_2\right\} \end{aligned} \qquad (2)$$

with:

- $\varphi^{(L)}$ – the activation function of the units in the Lth-layer, L = 1, 2,
- b_{iL} – the bias term of the ith unit in the Lth layer,
- $y_i^{(L)}$ – the output from the ith unit in the Lth layer,
- x_i – the ith observed acting load.

Various activation functions are available from technical literature (i.e. Haykin, 1999). In the present study we have considered one of the most common ones: the tanh-function and the linear function. While the first one is a non-linear sigmoidal function that maps its arguments into the domain [-1, 1], the linear function leaves the arguments unchanged.

The main reason for treating the model selection problem in the framework of ANN is given by the theorem of universal approximation. This theorem states that networks with one hidden layer, sigmoidal activation of the units in that hidden layer and linear activation of the output layer units are able to approximate every continuous function from one finite dimensional space to another to any desired degree of accuracy, provided a sufficient number of hidden units (Hornik et al., 1989). Therefore, we have restricted this study to network architectures that contain only one hidden layer and meet the activation conditions of the universal approximation theorem. The only task remained to be solved for the definition of the model structure is the specification of the number of hidden units. We will address this problem separately in the next chapter.

For now, let's assume a fixed model structure and focus on the estimation of the unknown weights $w_{ki}^{(L)}$ and biases b_{iL}. This is done on the basis of the observed input and output data by minimising a loss function E_{av}. Typically, this loss function is chosen to be the sum of squared differences between computed and observed outputs over all samples N and all output units N_O:

$$E_{av} = \frac{1}{N}\sum_{i=1}^{N} E_i = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\sum_{k=1}^{N_O}\left(y_{ANN,i}^{(k)} - y_{obs,i}^{(k)}\right)^2\right] \qquad (3)$$

Due to the high non-linearity of the ANN model (see eq. 2) the equations obtained from differentiation of the loss function (3) with respect to the unknown weights are still non-linear. Therefore, the estimation problem cannot be solved with closed formulas. Several methods are available for solving the minimisation problem. Upon them the gradient-based steepest descend method is widespread. This method uses only the 1st order derivative of the loss function to approach the minimising solution. In this study, we use the Levenberg-Marquardt algorithm (LM-algorithm) to obtain the estimates of the unknown parameters. Compared to the aforementioned method the LM-algorithm includes a 2nd order approximation to the loss function. It has therefore higher convergence rates and leads to more precise results. The second order approximation requires the computation of the Hesse matrix **H**. This is especially for large network structures a cumbersome computational task. Therefore, the LM-algorithm uses an approximation of the Hessian based on the Jacobi matrix **J**. In an iterative process the weights are changed in accordance to the rule:

$$\Delta\mathbf{w} = \left(\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}\right)^{-1} \cdot \nabla\mathbf{E}(\mathbf{w}) \qquad (4)$$

with:

- μ - regularisation parameter that insures the regularity of the approximation,
- **I** - identity matrix and
- ∇**E**(**w**) - the gradient of the loss function.

In this study we have chosen a value of 0.01 for the regularisation parameter. This small value gives a high contribution of the quadratic form - the approximation of the second order - to the change of the weights.

The aspects briefly presented in this chapter point out that there are theoretically well-founded methods for solving ANN models, provided that the model structure is fixed. While the theorem of universal approximation states the sufficiency of one hidden layer for good approximation properties of the network, it doesn't specify the exact number of the units contained in that hidden layer. It not even guarantees that this number is finite. Therefore, one still faces a problem of model selection prior to the estimation task. The solution to this problem is object of numerous research activities of the last years (Anthony and Bartlett, 2009). However, a unique solution has not been given yet. We will address this problem in the next chapter and present available methods to find the suitable number of nodes in the hidden layer.

3. MODEL SELECTION

The theorem of universal approximation guarantees the possibility of an exact projection of the input data onto the deformation signals. In practical applications the observed data is noisy. Therefore, perfect approximation of the deformations from the influencing factors cannot be the main scope of a modelling activity. Such an approximation would assume the storage of noise in the model coefficients. Rather than this, a model should capture the real functional relationship existing between the acting loads and the deformation signals on the basis of the observed data. Thus, the model including the estimated parameters will approximate well the observed deformation and will perform reasonable well on new data, i.e. for prediction tasks, at the same time. The latter aspect is called the generalisation property of a model. Obviously, an appropriate model is characterised by its ability to approximate and to generalise well.

3.1 Cross-validation

A good approximation property doesn't imply also a good generalisation property. A model structure that is chosen to be too complex in relation to the real functional relationship captures in its free coefficients the noise contained in the data. This occurrence is called overfitting. Such a model will perform well in approximating the data used for the estimation of its parameters but extremely poor on new data.

These facts are basic for the cross-validation, an empirically motivated method for selecting a suitable number of hidden nodes in an ANN. The methodical approach is to divide the available data set into two subsets: The first data set is called the training set and is used for estimating the weights of a certain model structure with a specified number of hidden units. Subsequent, the input data of the second subset, called the test data, is fed into the ANN which runs in the prediction mode. The discrepancy between the computed and the observed data of the second subset, expressed as mean square error (mse), is a measure of the generalisation property of the network. The relationship between training and test data is usually chosen to be 70% to 30%.

In the application of the cross-validation method one starts with a small model structure which is stepwise increased by adding hidden units (s. figure 1). For every structure the approximation and generalisation errors are computed as mse on the basis of the training and the test data respectively. It is expected, that the approximation error decreases continuously with increasing complexity of the model. At first, the generalisation will also improve with increasing number of hidden nodes. However, beyond a certain complexity the model will have a poor generalisation performance although the number of hidden nodes is still increasing. This point marks the structure where the model begins to overfit the training data. The number of hidden units corresponding to the minimum of the generalisation error determines the optimal model structure and represents the solution to the model selection problem.

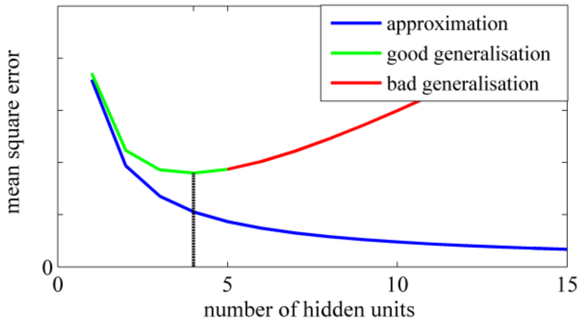

Figure 1. Cross-validation

3.2 Saliency of weights

In linear regression problems one faces a model selection problem as well. A typical approach to this task is to start with a large model structure which is sequentially thinned out by assessing the significance of single regression coefficients. This is done by a significance test that uses in case of normally distributed data the ratio between the model coefficient and its variance as a test value. In the linear case there is no problem to build this test value, because the variance of the coefficients results from the estimation process. This approach cannot be applied unaltered to ANN models due to their high degree of non-linearity. However, the concepts used in the linear case can be transferred to the ANN case in order to evaluate the relative importance – the saliency – of single weights of the ANN.

As in the linear case one starts with a relatively large network and removes sequentially weak connections between the units until a reasonable network architecture is obtained. The main characteristic of the method is the evaluation of the relative importance of the single weights. Due to the fact that the ANN is trained by minimising a loss function (s. chapter 2) it is a natural way to use this function for the definition of the relative importance of the weights. The saliency of a weight is defined as the change in the loss function induced due to its removal from the model structure.

For the direct identification of the weight with the lowest saliency the parameters of the complete model structure have to be estimated first. Then, in a second step each weight is removed temporarily from the model, the reduced model is trained and the corresponding change in the loss function is stored. The weight with the lowest contribution to the change of the loss function is removed permanently. However, this direct approach is computation intensive and thus, especially for large networks very time consuming.

Therefore, a different way of evaluating the saliency of weights is presented in Bishop (2008). The change of the loss function ΔE_{av} as consequence from the removal of a weight w_i is given in a second order approximation by:

$$\Delta E_{av,\, w_i} = \sum_i \frac{\partial E_{av}}{\partial w_i} \Delta w_i + \frac{1}{2} \sum_i \sum_j H_{ij} \Delta w_i \Delta w_j + \ldots \qquad (5)$$

In eq. (5) H_{ij} denotes the elements of the Hesse-matrix **H**. The removal of a weight is formal equivalent to a change of that weight $\Delta w_i = -w_i$. For a trained network the first summand in eq. (7) can be neglected. Thus, the variation of the loss function is determined mainly by the elements of the Hessian. If the non-diagonal terms of **H** are discarded, a usual procedure in evaluating the Hessian, the change of the loss function simplifies to:

$$\Delta E_{av,\, w_i} \approx \frac{1}{2} \sum_i H_{ii}^2 \Delta w_i^2 \qquad (6)$$

If a weight w_i is removed from the model structure the loss function increases according to eq. (6) approximately with $H_{ii} \cdot w_i^2 / 2$. Thus, this quantity measures the saliency of the specific weight w_i.

3.3 Model capacity

From functional point of view the model structure defined by an ANN (see i.e. eq. 2) corresponds to a family of functions f($\mathbf{x}$, $\boldsymbol{\omega}$) that maps the input $\mathbf{x}$ with an unknown distribution P($\mathbf{x}$) onto the output $\mathbf{y}$. The true conditional distribution P(y | $\mathbf{x}$) is also unknown. Therefore, the parameters $\boldsymbol{\omega}$ need to be estimated from observed pairs of data ($\mathbf{x}_i$, y_i), with i = 1,..., N. The estimation process is based on the minimisation of an empirical loss function:

$$R_{emp}=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-f\left(\mathbf{x}_i,\boldsymbol{\omega}\right)\right)^2 \tag{7}$$

This general structure of the loss function remains unaltered with respect to the one used for the approximation in LM-algorithm (eq. 3). The measure described by eq. 7 is the empirical risk. Good generalisation properties are obtained for a model if it minimises not only the empirical risk but also the true risk, defined as the expected value of the loss function:

$$R=E\left[\left(y-f\left(\mathbf{x},\boldsymbol{\omega}\right)\right)^2\right] \tag{8}$$

Eq. 8 cannot be evaluated because the joint distribution P($\mathbf{x}$, y) is unknown and only a finite sample of data is available from the observations. However, there are some upper bounds on true risk available for different model structures. All of them are based on the work of Vapnik and Chervonenkis (1971). In case of an unbounded nonnegative loss-function and some weak assumptions on the data distribution, like that of small probability for observing large values of the loss function, the following bound on the true risk holds with probability (1-η):

$$R\le R_{emp}\left[1-\sqrt{\frac{h}{N}\left(\frac{N}{h}+1\right)-\frac{1}{N}\ln\left(\frac{\eta}{4}\right)}\right]^{-1} \tag{9}$$

In eq. 9 h denotes the Vapnik-Chervonenkis dimension (VCdim) of the model structure. This measure is characteristic for a certain model structure. Its nature will be briefly explained in the following. This can be done more comprehensive if we leave for short the framework of regression and address problems of classification. In consequence, the family of functions f($\mathbf{x}$, $\boldsymbol{\omega}$), with $\boldsymbol{\omega} \in \Omega$, is a set of indicator functions that map the input onto {0, 1}. A sample x_1, x_2, ...,x_n is said to be shattered by the function set f($\mathbf{x}$, $\boldsymbol{\omega}$) if for every possible separation there exist an $\boldsymbol{\omega}^* \in \Omega$ such that the function f($\mathbf{x}$, $\boldsymbol{\omega}^*$) classifies error-free the data. The maximum number of samples that are shattered by a function set f($\mathbf{x}$, $\boldsymbol{\omega}$) defines the VCdim of the function set. In a binary classification task n samples can be separated in 2^n ways. Therefore, h equals max(n) for which the function set separates error-free the samples. The VCdim is a measure of the capacity of a model.

For exemplification the VCdim of the linear discriminator function set is analysed in figure 2 for a two dimensional input space.

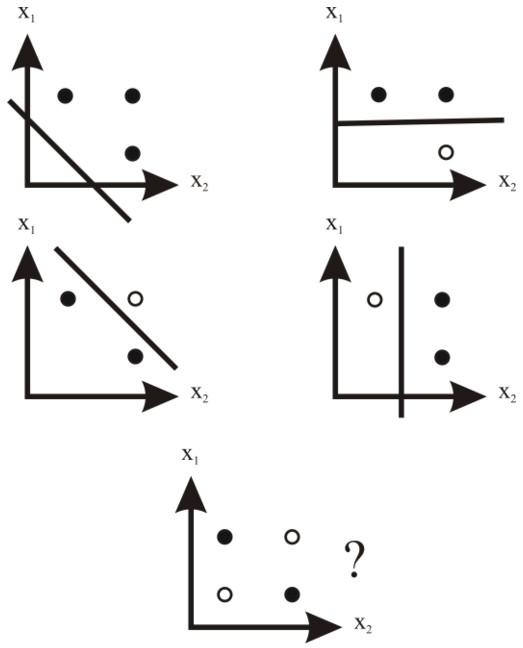

Figure 2. VCdim of the linear discriminator

As can be seen from the first 2 rows in figure 2 the linear discriminator performs all possible classifications of a sample size of N = 3. Only 4 cases are exemplified in figure 2. The remained 4 cases are obtained by switching the white and black circles. For a sample size N = 4 the linear discriminator is not able anymore to perform all possible separations. The 3rd row in figure 2 illustrates a case when the linear discriminator fails. Therefore its VCdim is h = 3.

Returning to the regression framework one faces the problem of extending the concept of VCdim, which is obviously a combinatorial measure, to sets of real functions f($\mathbf{x}$, $\boldsymbol{\omega}$). This is accomplished by conversion of the real functions into indicator functions. For each $\mathbf{x}$ these indicator functions specify whether f($\mathbf{x}$, $\boldsymbol{\omega}$) exceeds or is below a certain level β:

$$f\left(\mathbf{x},\boldsymbol{\omega}\right)\rightarrow I\left[y,f\left(\mathbf{x},\boldsymbol{\omega}\right),\beta\right]=I\left[\left\|y-f\left(\mathbf{x},\boldsymbol{\omega}\right)\right\|^2-\beta>0\right] \tag{10}$$

The exact VCdim is known exactly only for a few ANN structures (see i.e. figure 2). Vapnik et al. (1994) proposed a method for estimating the VCdim from experimental data. This method is based on the maximum deviation of error rates observed on two independent data sets ($\mathbf{x}$, $y)_1$ and ($\mathbf{x}$, $y)_2$. Closed formulas were derived for the expectation of this maximum deviation. The latter is a function of VCdim and the data length n. Therefore, calculating the maximum deviation of error rates on two data sets of different lengths n_1, n_2, ..., n_k = N, one is able to chose an appropriate value of h such that the empirical variation of the maximum difference fits optimally, i.e. in the L_2-norm sense, the derived closed formula.

With the VCdim of the model structure estimated in accordance to the method outlined above and the empirical risk R_{emp} obtained from the network training the upper bound (9) for the true risk can be evaluated. It is expected, that the ANN structure that leads to the least upper bound (9) will exhibit the best generalisation properties from the set of possible model structures.

4. EXAMPLES OF MODEL SELECTION

The methodologies presented in the former chapters were used to perform a system identification of a lock situated in Uelzen. This lock was object of numerous research works; especially at the Geodetic Institute of Hanover (i.e.: Neuner, 2008; Boehm and Kutterer, 2006). The deformation model of the lock was already discussed in a large number of publications related to this research work. Therefore it will not be repeated here.

The deformation measurements were performed with an inductive measuring plummet system that was mounted in the tower of the tail-bay. The sampling interval is 10 min. The analysed data covers a time span of 4 days. The main acting loads causing the deformation: the change of water level in the chamber due to the activity of the lock and the temperature were recorded synchronous to the deformation. The analysed data is presented in figure 3.

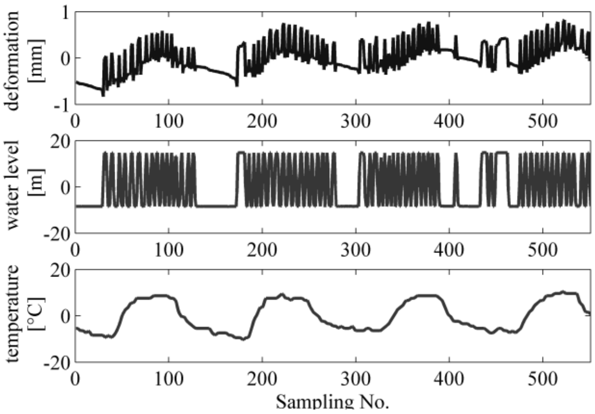

Figure 3. Analysed data

The phase differences between the deformation and the acting loads were computed with methods of time series analysis. Afterwards the time series were aligned in time prior to their modelling with ANN. The phase difference between the time series in figure 3 is already removed.

The causal relationship between the system's input and output is obvious from figure 3. Although one might assume a linear relationship between the two components of the causal chain this is not the case. As described in Neuner (2008) the direction of the deformation changes during the filling and the emptying of the lock. Furthermore, the system's reaction to the changes of water level is also temperature dependent. Therefore, a non-linear model structure based on ANN is chosen for the system identification. The feed-forward ANN used here contains two input units corresponding to the abovementioned acting loads. The single output unit corresponds to the deformation. The network weights were estimated using the LM-algorithm.

The main task of this study is to assess the number of units in the hidden layer that leads to a model structure with good generalisation properties. For this purpose the 3 methods described in chapter 3 were used.

For cross-validation the data set was separated into two subsets: the training data set covering a time span of three days and the test data set covering one more day. The training and testing procedures were performed on model structures with 1, 2, 3, 5 and 10 hidden units. The maximum number of 10 was set in accordance with the number of samples of the training set such that the weights of the resulting model can still be determined. The results obtained from the training and the testing of the ANNs with the LM-algorithm are shown in figure 4.

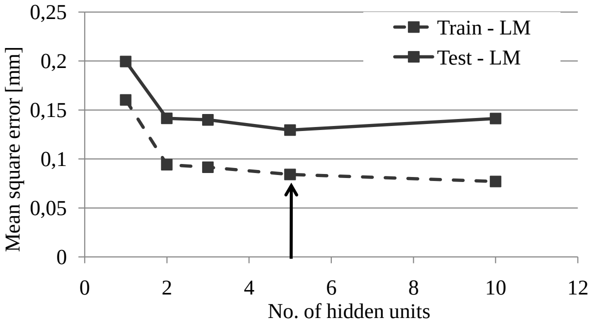

Figure 4. Results of cross-validation procedure

The empirical variation of the mse with respect to the number of hidden units agrees well with the expected one described in chapter 3.1. The mse computed from the training data is smaller than the one calculated from the test data and decreases at the highest rate between the structures with 1 and 2 hidden units. The minimum is attained in case of the structure with 5 hidden units. Therefore, as a result of cross-validation it is expected that this model has the best generalisation properties.

The saliency of weights method requires as described in chapter 3.2 the computation of the Hessian matrix **H**. This contains the 2nd order derivations of the loss function E_{av} with respect to the weights. Several elements of the Hessian were rigorously calculated by differentiation of E_{av} given in eq. 2 with $\varphi^{(1)}$ as the tanh-function and $\varphi^{(2)}$ as the linear function. This was accomplished for all model structures with 1, 2, 3, 5 and 10 hidden nodes. All training samples are included in the numerical computation of the elements.

For comparison reasons with the cross-validation the properties of the Hessian and the saliencies of weights for the structure with 5 hidden nodes are presented here. The condition of the resulting Hessian can be assessed in figure 5 that contains the eigenvalues obtained from the single value decomposition of **H**.

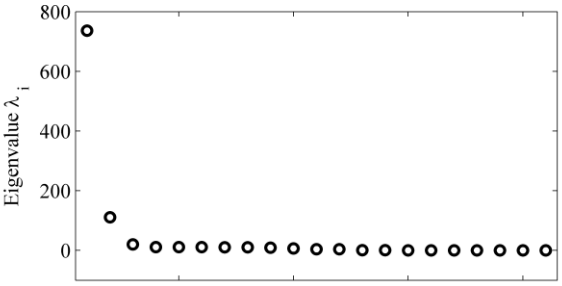

Figure 5. Eigenvalues of the Hesse matrix **H**

The Hessian has one dominant eigenvalue. The relation between the maximum and the minimum eigenvalue, the conditional number of the matrix, is $2{,}77 \cdot 10^{-5}$. This doesn't allow a sharp classification of the model as good or bad conditioned. However, in further studies performed with the estimated Hesse matrix in spite of the small conditional number a stable calculation of the inverse was possible. Thus, a stable solution to the approximation problem could be obtained with the Newton method. From this point of view the model choice after cross-validation seems to be appropriate.

The saliency of the weighting coefficients of the ANN structure with 5 hidden nodes was calculated due to the availability of the entire Hessian according to the last summand in eq. 5.The use of the approximating eq. 6 produces only small differences between the saliency coefficients.

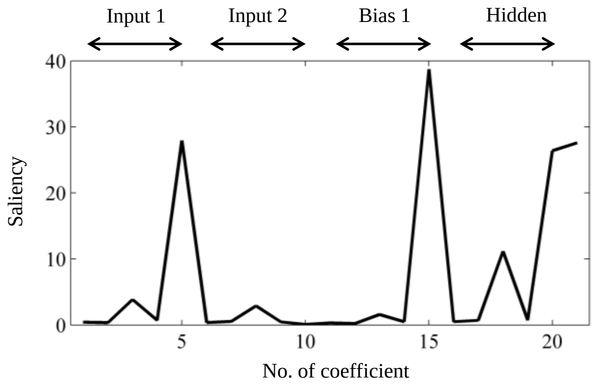

Figure 6. Saliency of weights

The obtained saliency coefficients are plotted in figure 6. There is a total amount of 21 coefficients each of them corresponding to a weight of the ANN. The first 3 groups of 5 coefficients correspond to the weights of the connection between one input node and the 5 hidden units. The last 6 coefficients refer to the connections of the output node with the 5 hidden nodes and the bias node respectively. At first, the focus of the analysis lies on the last 6 coefficients. Deleting one of these connections means that the respective hidden node will have no contribution to the computed output from the network and thus, it can be removed from the structure. Figure 6 reveals that at least 3 connections, including the one to the bias, have a high saliency. The other three connections are characterised by relatively small but not necessary neglectable coefficients. Therefore, the method of saliency of weights leads to a plausible result that confirms the one obtained by cross-validation. Though, notice that the saliency of weights method refers only to the approximation capabilities of the model structure. It doesn't refer to its generalisation properties.

The generalisation property of a model structure is optimised using the theoretical concept of model capacity. Using the method outlined in chapter 3.3 the following values were obtained for the VCdim parameter h in case of network structures with 2, 3, 5 and 10 nodes:

Number of hidden nodes	2	3	5	10
Estimated h	4	5	8	12

Table 1. Estimated values for h

Using the values of the minimum empirical risk resulting from the training process with the LM-algorithm the upper bound for the true risk was calculated for the 4 structures of the ANN given in Table 1 using eq. 9 with a probability error of $\eta = 5\%$. The lowest bound was obtained for the ANN with 5 hidden nodes. This result agrees very well with the one empirically obtained by cross-validation.

5. SUMMARY

This paper refers to the non-linear system identification by means of ANN. It is focused on the model selection task that consists in the specification of an adequate number of hidden nodes that lead to models with good approximation and generalisation properties as well. Three methods for model selection were presented: the empirical motivated cross-validation, the saliency of weights method and the theoretically well founded one based on the capacity of the model described by the VCdim measure. The 3 methods were applied for the system identification of a lock and leaded to a good agreement between the obtained results. The identified model structure consists of one hidden layer containing 5 nodes.

References

Anthony, M., Bartlett, P. L. 2009. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, UK.

Bishop, C. M. 2008. *Neural Networks for Pattern Recognition.* Oxford Press, UK.

Haykin, S. 1999. *Neural Networks: a comprehensive foundation*. 2nd edition, Pearson Education, Singapore.

Hornik, K., Stinchcombe, M., White, H. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, Vol. 2, pp. 359 – 366, Pergamon Press.

Neuner, H. 2008. *Zur Modellierung und Analyse instationärer Deformationsprozesse*. Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover, Nr. 269.

Neuner, H., Kutterer, H. 2010. *Modellselektion in der ingenieurgeodätischen Deformationsanalyse*. In: Wunderlich, Th. A. (Ed.): „Beiträge zum 16. Internationalen Ingenieurvermessungskurs, München, 2010“, pp. 199 – 210, Wichmann, Berlin.

Vapnik, V. N., Chervonenkis, A. J. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Apl.* (16), pp. 264 – 280.

Vapnik, V. N., Levin, E., LeCun Y. 1994. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6, pp. 851 – 876.

Welsch, W., Heunecke, O., Kuhlmann, H. 2000. *Auswertung geodätischer Überwachungsmessungen*. In the series: Möser, M., Müller, G., Schlemmer, H., Werner, H. (Eds.): Handbuch Ingenieurgeodäsie. Wichmann Verlag, Heidelberg.