

Interdependencies in Data Pre-processing, Training Methods and Neural Network Topology Generation

Stephan Rudolph* and Steffen Brückner†

Institute for Statics and Dynamics of Aerospace Structures,
Stuttgart University, Stuttgart, Germany

ABSTRACT

Artificial neural networks are adaptive methods which can be trained to approximate a functional relationship implicitly encoded in training data. A large variety of neural network types (e.g. linear versus non-linear) gives rise to principal questions about the appropriateness of data pre-processing techniques, training methodologies, the resulting neural network topology and possible interdependencies thereof. The a posteriori interpretation of the numerical results gives hints for some guidelines for neural network applications in engineering applications. Data pre-processing techniques are a powerful means for pre-structuring the problem setting of function approximation through an adaptive training procedure. Especially integral transforms may change the nature of the training problem significantly without loss of generality if carefully selected and represent an excellent opportunity to incorporate additional knowledge about the process to improve the training and the result interpretation. Some numerical examples from engineering domains are used to illustrate the theoretical arguments in the context of a practical setting.

Keywords: Neural Networks, Data Pre-processing, Topology Generation, Integral Transforms

1. INTRODUCTION

Artificial neural networks (ANNs) are a mathematical tool for function approximation mapping input vectors \mathbf{x} onto output vectors \mathbf{y} . The networks are constructed and represented by several layers of simple information processing units, the neurons, and weighted connections between these neurons. Different classes of ANNs are known, such as the ANNs described in the following:

- linear and nonlinear ANNs
depending on the kind of activation functions used in the neurons,
- feed-forward and recurrent ANNs
depending on the connections between the neurons, recurrent networks incorporate feedback
feed-forward ANN: static model,
recurrent ANN: dynamic model.

2. DATA PRE-PROCESSING

The training algorithms used for ANNs usually belong to the group of optimization algorithms. This kind of iterative methods is extremely sensitive to the training data. Therefore the careful pre-processing of the raw data gained e.g. from experiments, is mandatory. Algorithms for data cleansing, the deletion of outliers, and for most ANN types normalization of the input data should be fairly standard. In engineering applications additional knowledge besides the raw data itself are the associated dimensions. Each measurement (“date”) has an assigned dimension, such as length L or time T . This knowledge is often neglected but can be used advantageously to pre-process the data and to impose additional restrictions on the possible ANN topology as will be shown in the following.

* rudolph@isd.uni-stuttgart.de, ISD Uni Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart

† brueckner@isd.uni-stuttgart.de, ISD Uni Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart

3. DIMENSIONAL ANALYSIS

Buckingham's Pi-Theorem¹ states that for each dimensional homogeneous and complete relationship f of n physical variables x_i

$$f(x_1, \dots, x_n) = 0 \quad (1)$$

there exists a corresponding relationship

$$F(\pi_1, \dots, \pi_m) = 0 \quad (2)$$

of only $m \leq n$ dimensionless groups

$$\pi_j = x_{j+r} \prod_{i=1}^r x_i^{-\alpha_{ji}} ; \alpha_{ji} \in \mathbb{R} \quad (3)$$

with $m = n - r$, where r denotes the rank of the dimensional matrix.

The dimensional matrix can be established from the knowledge of the problems' relevance list, which is the list of all relevant parameters and their respective dimensions. The rows of the dimensional matrix correspond to the parameters while the columns correspond to the value of the dimension exponent of the variable. The elements of the dimensional matrix are therefore the exponents of the dimensions of all the relevant parameters.

$$\begin{array}{|c|ccc|} \hline & D_1 & \cdots & D_k \\ \hline x_1 & & & \\ \vdots & & & \\ x_n & & & \\ \hline \end{array} \Rightarrow \begin{array}{|c|ccc|} \hline & \bar{D}_1 & \cdots & \bar{D}_r \\ \hline x_1 & & & I \\ \hline \vdots & & & \\ \hline x_n & & & \alpha_{ji} \\ \hline \end{array} \quad (4)$$

The dimensional matrix \mathbf{D} , shown in eq. (4) on the left, is formed by the relevant parameters x_i as the rows and the corresponding dimensional exponents D_j as the columns of the matrix. Applying rank preserving column operations on the dimensional matrix, the original dimensional matrix is transformed into the matrix shown in eq. (4) on the right side. This matrix consists of an upper square identity matrix \mathbf{I} of size $r \times r$ and a lower sub-matrix \mathbf{A} of size $(n-r) \times r$. The lower sub-matrix \mathbf{A} contains the exponents a_{ji} used in the forming of the dimensionless groups in equation (3).

The rank r of the dimensional matrix \mathbf{D} represents the number of independent base dimensions involved in a given problem. Mechanical problems e.g. can often be written in the two base dimensions force and distance instead of the three SI-unit system base dimensions length, mass, and time. In either case however, the rank of the dimensional matrix will be $r=2$ after execution of the rank preserving dimensional matrix operations.

4. SIMILARITY NETWORKS

For engineering applications, where the principle of dimensional homogeneity holds, the neural network topology can be designed according to the results of dimensional analysis. This leads to the network structure shown in figure 1. The first layer encodes the pi-transform $\boldsymbol{\pi}$ which is a priori knowledge based on the technique of dimensional analysis. The same applies to the last layer, which encodes the back-transform $\boldsymbol{\pi}^{-1}$ from the dimensionless groups π_j to the physical variables x_i using additional shortcut connections from the input layer.

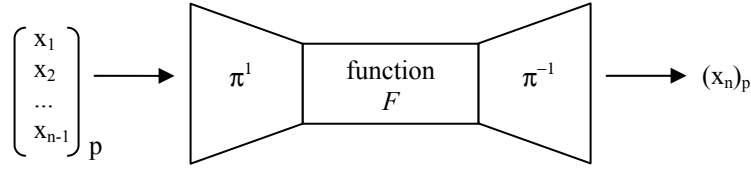


Fig.1 : Topology of a similarity network²

Neural network topology designs according to these rules are called “similarity networks” because they incorporate the principle of physical similarity in terms of the dimensionless groups². Compared to a neural network without this construction principle, a much smaller parameter space of inner neural network weights has to be searched. All neural network states in such a similarity network fulfill the requirement of dimensional homogeneity, the similarity network has in every stage of the training a physically admissible state. Once a training pattern has been correctly learned by the neural network, a similarity network has also learned all completely similar cases. This leads to the two propositions²

- A neural network is **pointwise correct** for an unknown data set $(x_1, \dots, x_n)_q$ if and only if a completely similar training set $(x_1, \dots, x_n)_p$ is approximated without error. This is the mathematically necessary condition for the pointwise correct generalization, since all physically completely similar points $(x_1, \dots, x_n)_q$ are condensed onto one point with $\pi_{j,p} = \text{const}$.
- A neural network is **correct on the complete domain** of the dimensional function f if and only if the network has learned to approximate the true dimensionless function F without error. This is the mathematically *necessary and sufficient* condition for correct generalization, since the dimensionless function F can readily be transformed by inserting the dimensionless groups into the correct dimensional function f .

The use of similarity networks needs the input/output pairs to fulfill the principle of dimensional homogeneity. This means that the original data has to be pre-processed eliminating unnecessary information and emphasizing the useful information in the data. Dimensional analysis, as described, helps to reduce the number of parameters for a problem by exploiting the available information about the physical dimensions of the input-output variables (x_1, \dots, x_n) . These considerations hold for all kinds of functional relationships in the form of eq. (1). For time-series data additional means of pre-processing steps must be found to represent this kind of signals appropriately for a successful neural network training. Some of these pre-processing methods will be described in later sections.

Similarity networks can also be combined with evolutionary techniques to allow an automated topology optimization. This will be shown in the following section.

5. TOPOLOGY GENERATION

Another way of selecting an adequate neural network topology is the use of genetic algorithms in conjunction with the data property of complete similarity. Due to this property of every measurement value pair in the data set, an infinite number of completely similar cases can be generated automatically. The genetic algorithm has a DNA (i.e. a genetic representation) which encodes the number of neural network layers and the activation functions used in these layers. Numerical simulations have shown that for engineering examples, neural network topologies are found which encode the (known) structure of the physical relationship and produce an error in parameters and input/output behavior of less than 0.01% within a range of only 50 – 150 generations.

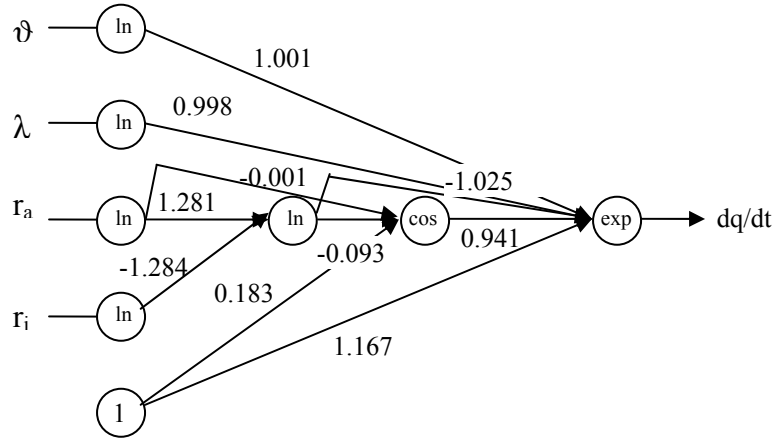


Fig. 8: Neural network topology for heat transfer through a pipe with a generalization error of 0.0362 % after 49 generations with a genetic algorithm³

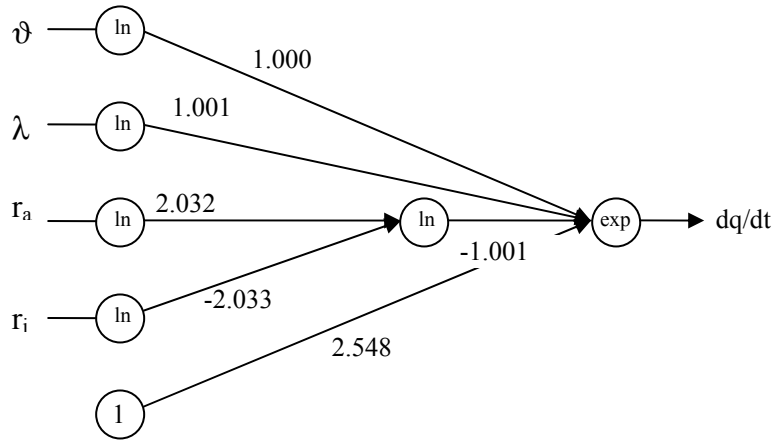


Fig. 9: Neural network topology for heat transfer through a pipe with a generalization error of 0.0011 % after 71 generations with a genetic algorithm³.

The neural networks in shown in figure 8 and 9 have been trained using a genetic algorithm on a grid of up to 5 layers and up to 5 neurons per layer³. The algorithm has a choice of 10 activation functions and the neural networks in each generation were trained using vanilla backpropagation or threshold accepting³. The physical relationship encoded in the neural network in figure 8 can be extracted from the network as³

$$\frac{dq}{dt} = 3.212 \lambda^{1.001} \vartheta^{0.998} \frac{\exp\left(0.941 \cos\left(0.183 - \left(0.093 \ln\left|1.281 \ln|r_a| - 1.284 \ln|r_i|\right|\right) - 0.001 \ln|r_a|\right)\right)}{\left(1.281 \ln|r_a| - 1.284 \ln|r_i|\right)^{1.025}} \quad (5)$$

The function encoded in the neural network in figure 9 is similarly given by

$$\frac{dq}{dt} = 12.782 \cdot \lambda^{1.001} \cdot \vartheta^{1.000} \cdot (2.032 \cdot \ln|r_a| - 2.033 \cdot \ln|r_i|)^{-1.001} \quad (6)$$

Compared to the known analytic formula

$$\frac{dq}{dt} = 2\pi \cdot \lambda^{1.000} \cdot \vartheta^{1.000} \cdot (\ln|r_a| - \ln|r_i|)^{-1} \quad (7)$$

it can be seen, that these formulas (6) and (7) agree up to the second digit. The intermediate neural network after 49 generations gives eq. (5) which is, if evaluated *numerically*, very similar to the analytic formula in eq. (7). The *symbolic* representation however is much more distinct from the known solution and difficult it not impossible to interpret. The results in figure 9 have been achieved without pre-structuring the neural network, but only by using artificially generated completely similar data.

It has been shown in simulations³, that classical (dimensional inhomogeneous) neural networks can perform better at certain points in the domain, but the dimensional homogeneous neural networks had a much better overall performance over the domain.

The use of neural networks in the engineering domain often requires the analysis of time series data. If used without any pre-processing, the raw time series data could be fed into the neural network as a relatively high number of time-value pairs. Alternatively, the very same time series data could be transformed into another, probably more suitable signal representation using much less coefficients. In the following sections different means of transforming time-series data into inputs for neural network classifiers are shown.

6. TIME HARMONIC SIGNALS

An important class of signals is the class of time-harmonic signals, which have time-invariant frequency content. Each time-harmonic signal can be written in terms of a Fourier series

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \sin(k\omega t) + b_k \cos(k\omega t)) \quad (8)$$

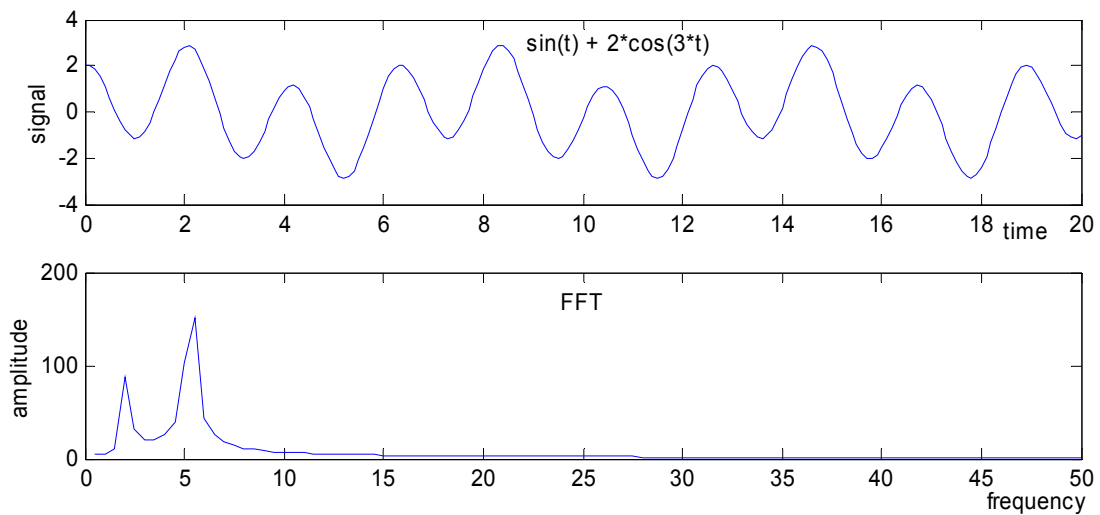
with the coefficients

$$\begin{aligned} a_k &= \frac{2}{T} \int_0^T x(t) \cdot \cos(k\omega t) dt \quad ; k = 0; 1; 2; \dots \\ b_k &= \frac{2}{T} \int_0^T x(t) \cdot \sin(k\omega t) dt \quad ; \omega = \frac{2\pi}{T} \end{aligned} \quad (9)$$

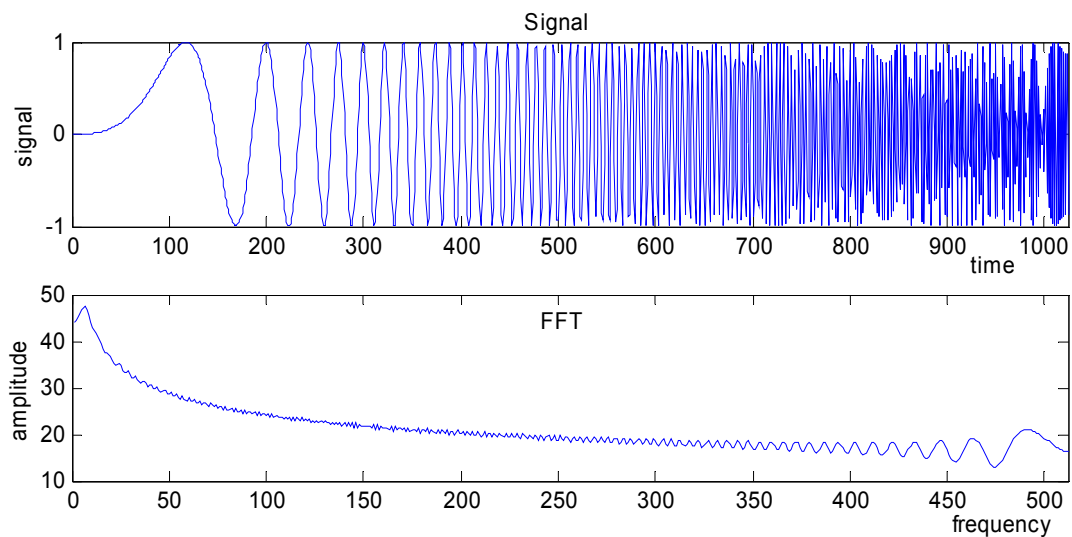
In the case of a continuous signal the continuous Fourier transform is given by the integral equation (10)

$$F(\omega) = \int_{-\infty}^{\infty} x(t) \cdot e^{-j\omega t} dt \quad (10)$$

When working with signals which contain only a few frequencies or a small frequency band, the number of Fourier coefficients needed to describe the signal can be significantly less than the number of sampling points in the time domain. On the other hand, when working with signals with broad-band frequency content Fourier analysis leads to a large number of coefficients.



*Fig. 2: Time-harmonic signal and it's Fourier representation
The two sinusoidal component can be clearly identified in the Fourier transform*



*Fig. 3: Chirp signal and its Fourier transform;
The time-dependency of the chirp signal cannot be identified in the Fourier transform*

The Fourier representations of the time-harmonic and chirp signal in figure 2 and 3 respectively clearly show, that Fourier analysis gives only satisfying results when signals with time-invariant frequency content are analyzed.

7. WAVELET ANALYSIS

For non-stationary signals the multilevel wavelet decomposition provides a good means of characterizing the data. Compared to Fourier analysis the wavelet analysis can resolve frequency dependencies as well as time-dependencies. E.g. for analyzing noise pulses on top of time-harmonic electro-magnetic oscillations wavelet analysis is far more better suited than Fourier analysis. The continuous wavelet transform is well suited as an analysis tool, but it's the lack of reversibility that usually leads to the use of the discrete wavelet transform which is reversible. Mostly, the multilevel discrete wavelet decomposition and reconstruction method are used.

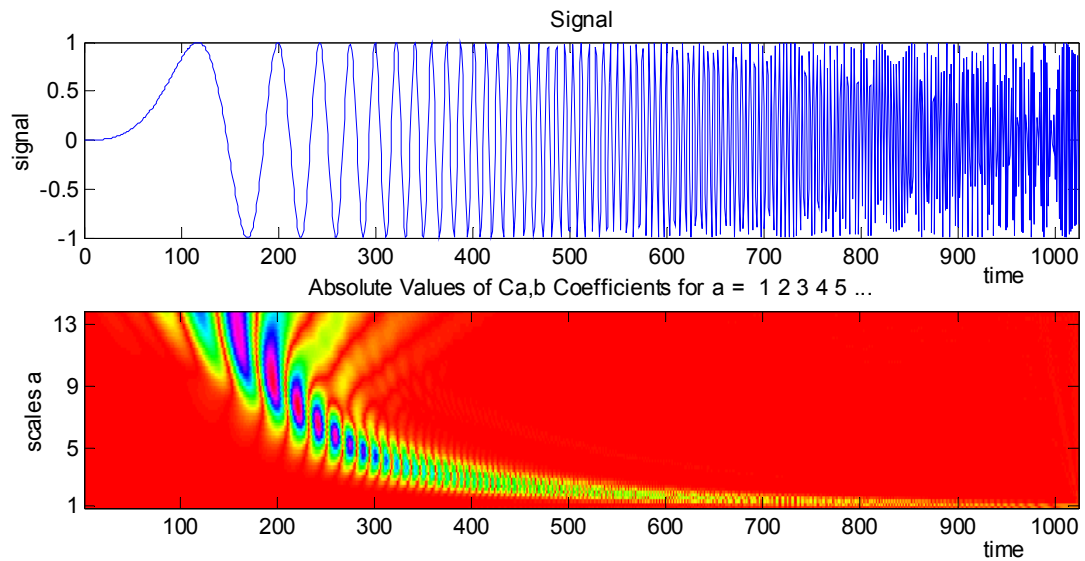


Fig. 4: Chirp signal and the continuous wavelet transform

The signal shown in figure 5 is given as a 201 data point time-series. The wavelet representation shown in the lower part of figure 5 is reconstructed from only 30 wavelets. Especially for non-harmonic time-series the wavelet transform can significantly reduce the number of coefficients needed to represent the signal.

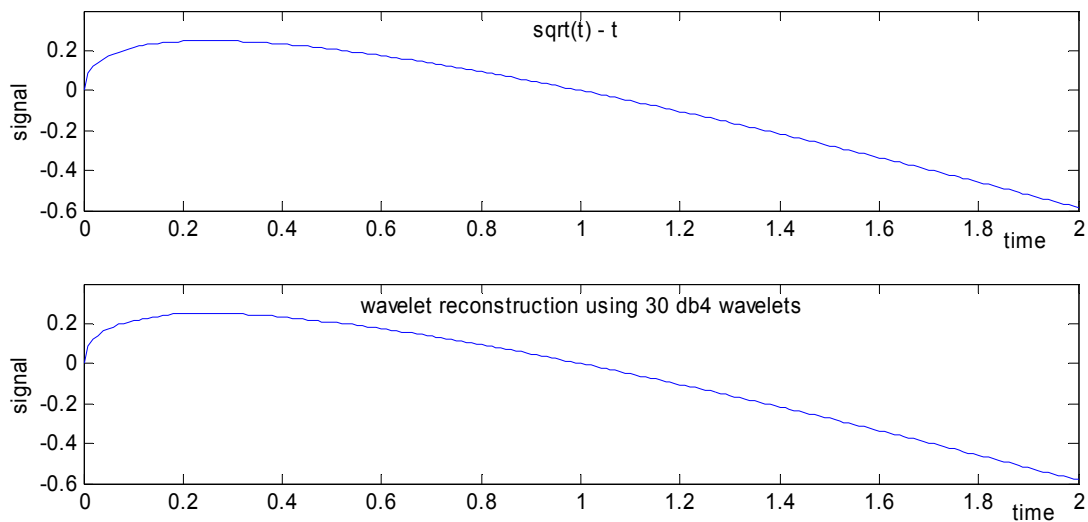


Fig. 5: Wavelet reconstruction of a signal consisting of 201 data points using only 30 wavelet coefficients.

8. ACOUSTIC CLASSIFICATION

The classification of acoustic phenomena is a non-trivial task for scientists and engineers. In order to use a neural network to classify acoustic data, this data has to be pre-processed. In this example, the acoustic data consists of time-series of pressure measurements (taken with a microphone). These time-series are transformed into two-dimensional time-frequency distributions which can then be interpreted as images. From the research in pattern recognition, the method of higher moments is well known⁵. Employing this method on the images of the time-frequency distributions and non-dimensionalizing the moments leads to dimensionless moments of the acoustic data. Where the classical moments hide isotropic scaling from the classifier, the dimensionless moments also hide anisotropic scaling. It can be shown that for some classes of acoustic phenomena (e.g. the class of “whistling” sounds⁴) the dimensionless moments remain constant. Once one data point from this class is classified correctly, all other members of this class are also classified correctly.

This method using dimensionless higher moments follows the general rule that all previous knowledge should be exploited and the neural network should only be used to adjust for the yet unknown part of the mapping.

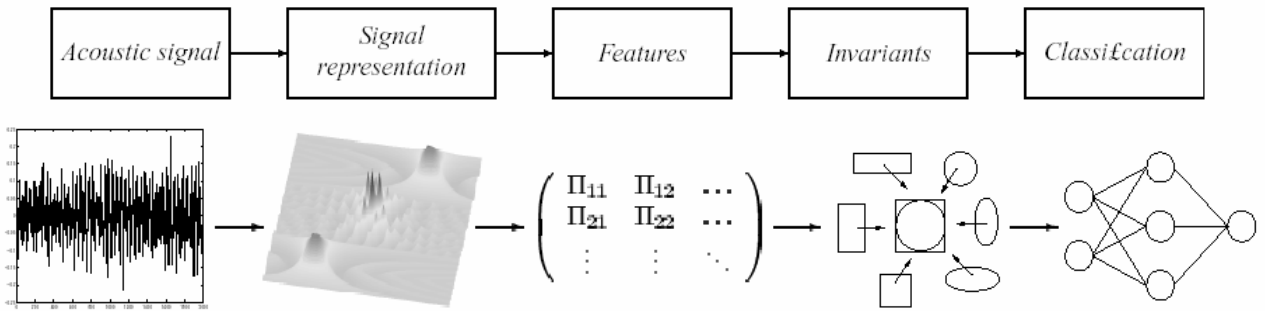


Fig. 6: The signal classification approach⁴

The signal classification approach shown in figure 6 can be adopted for acoustic signals transforming the one-dimensional signal into a two-dimensional time-frequency representation, such as the spectrogram shown in figure 7. A general class of time-frequency distributions is given by the Cohen class

$$C_x(t, \omega) = \iiint x\left(t + \frac{\tau}{2}\right) \cdot x^*\left(t - \frac{\tau}{2}\right) \cdot \Phi(\xi, \tau) \cdot e^{-j\xi t + j\xi u - \omega\tau} du d\tau d\xi \quad (11)$$

where $\Phi(\xi, \tau)$ is a two-dimensional kernel function which can be freely chosen. The selection of this kernel function leads to different time-frequency representations, such as the Spectrogram, the Wigner-Ville distribution, and others. Depending on the source signals, these representations have specific advantages. An optimization algorithm can select the kernel function which gives the best classification results with the given data.

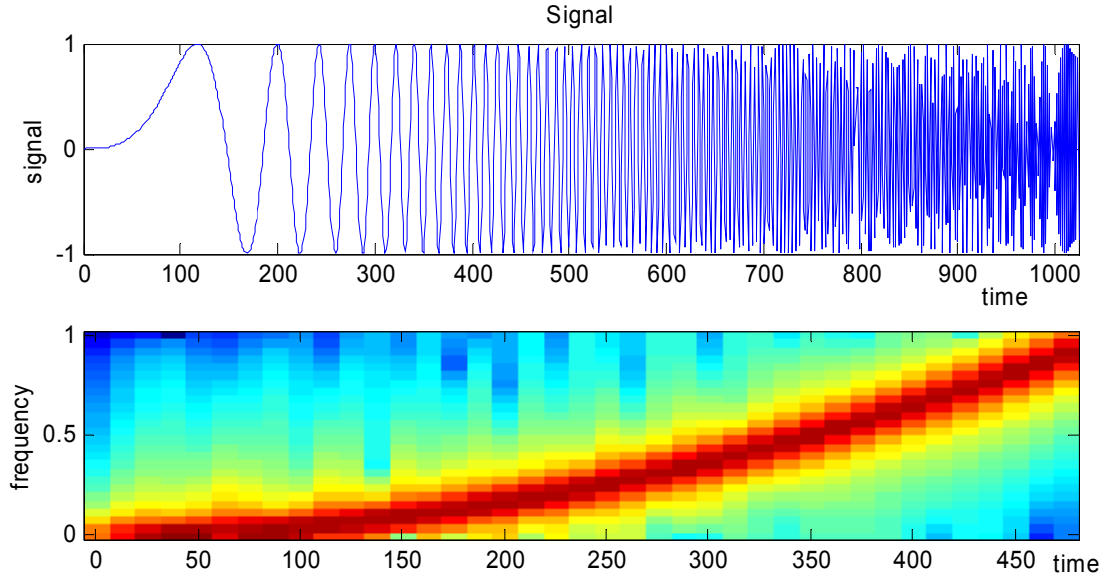


Fig. 7: Chirp signal and spectrogram representation

From the generated images with the image function $f(x,y)$, higher moments

$$m_{pq} = \iint x^p y^q f(x,y) dx dy \quad (12)$$

can be calculated. These moments can be non-dimensionalized to the dimensionless moments Π_{pq} ⁵

$$\Pi_{pq} = \frac{m_{pq}}{\frac{-p+3q+2}{8} \frac{3p-q+2}{8} m_{02} m_{20}} \quad (13)$$

These non-dimensional moments can be used as inputs in a neural network classifier. It has been shown⁴ that these dimensionless moments Π_{pq} are invariant for whistling and booming noise, therefore producing identical inputs to the classifier for all noises belonging to these classes. Once one signal is classified correctly, all other similar signals are also classified correctly.

9. SUMMARY

It has been shown that neural networks for engineering applications can be designed according to the results of dimensional analysis. These similarity networks can be combined with evolutionary techniques to form a method for automated topology generation and network parameter optimization. The resulting networks produce numerically very well behaved and accurate results. The theoretical background of the similarity network provides also a framework for the interpretation of the internal neural network structure. The symbolic equations extracted from these networks come close to the analytically derived formula in the shown test case.

In the second part of this work, different means of data pre-processing for time series data have been shown. The Fourier transform approach is well suited for signals with time-independent frequency content and the wavelet approach is advantageous for signals with multi-scale effects. Another method has been show for acoustic signal classification. Using time-frequency distributions and non-dimensional higher moments provides a pre-classification for certain classes of acoustic signals, supporting therefore the neural network classifier.

These pre-processing methods are especially useful if they support the concept of physical dimensions in their mathematical representation, which allows a fruitful combination with similarity networks used as classifiers or function approximators which exploit that additional dimensional information.

10. REFERENCES

1. Buckingham, E., *On Physically Similar Systems: Illustration of the Use of Dimensional Equations*, Phys. Review 4, 345-376, 1914.
2. Rudolph, S., *On Topology, Size and Generalization in Non-Linear Feed-Forward Neural Networks*, Neurocomputing, Vol. 16, 1, 1-22, July 1997.
3. Rudolph, S., *On a Genetic Algorithm for the Selection of Optimally Generalizing Neural Network Topologies*, In: I.C. Parmee (ed), Second International Conference on Adaptive Computing in Engineering Design and Control '96, University of Plymouth, March 26-28, 79-86, 1996.
4. Till, M. and Rudolph, S., *Optimized time-frequency distributions for signal classification with feed-forward neural networks*, in Proceedings SPIE Aerosense 2000 Conference On Applications and Science of Computational Intelligence III, SPIE, (Orlando, USA), 24-28th April 2000.
5. Melan, A. and Rudolph, S., *An analytical approach to classification by object reconstruction from features*, Proceedings SPIE Aerosense 2000 Conference On Signal Processing, Sensor Fusion and Target Recognition IX, Orlando, Florida, April 24-28th, 2000.