ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 439-444

Nonlinear, statistical Data-Analysis for the optimal Construction of Neural-Network Inputs with the Concept of a Mutual Information

Frank Heister, Gregor Schock

{frank.heister | gregor.schock}@daimlerchrysler.com
DaimlerChrysler Research and Technology (FT2/EA)
HPC T721
D-70546 Stuttgart
Germany

Abstract. In this article we focus on a statistical method for nonlinear time series analysis of data-sets used in supervised neural network training. A new method for identifying a minimal neural input-vector with maximum information content is proposed. Further, we demonstrate the capability of the mutual information for nonlinear time series analysis of real measurement data. From the viewpoint of information theory this approach provides optimal solutions for a large variety of problems.

1. Introduction

One of the basic postulates of information theory is that information can be treated like a measurable physical quantity, such as density or mass. Whenever entities of the real world are interacting, an abstract flow of information occurs. Quantifying this flow of information is of vital interest for the determination of implicit causalities and hence for the construction of proper training-sets for neural networks. The mutual information is capable of identifying arbitrary dependencies, while the coefficient of correlation (CoC) fails to detect nonlinear dependencies between arbitrarily distributed random variables. Another advantage of this quantity is its applicability to multi-dimensional time series.

2. Mutual Information

The mutual information $I(\xi, \eta)$ can be interpreted as the quantity of information obtainable about a random variable η, from the prior knowledge of another variable ξ. Two, possibly multivariable, signals $\{x_n\}$ and $\{y_n\}$ can now be interpreted as realizations of the random processes $\{\xi_n\}$ and $\{\eta_n\}$. In this case,

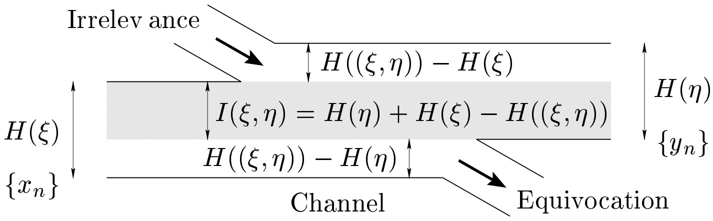

Figure 1: Shannon's model for the flow of information through an abstract, symmetric message-channel contaminated with noise.

the mutual information is used to measure the flow of information and thus the degree of statistical dependency between both random variables.

Shannon introduced the concept of mutual information for the quantitative description of abstract message-channels. Figure 1 depicts Shannon's model for the flow of information through an abstract, symmetric message-channel contaminated with noise. The received information $H(\eta)$ is comprised of the mutual information $I(\xi,\eta)$ and the irrelevance $H((\xi,\eta)) - H(\xi)$ resulting from disturbances. $H((\xi,\eta)) - H(\eta)$ describes the equivocation, i.e. the information which is actually lost by the channel and is never received.

This pragmatic methodology is based upon the introduction of an adequate entropy-measure H. For the discrete random variable ξ with a probability distribution $\{p_m\}$, the entropy-measure H is defined as

$$H_\alpha(\xi) := H_\alpha(\{p_m\}) = \begin{cases} \frac{1}{1-\alpha} \log_2 \sum\limits_{m=1}^{M} p_m^\alpha & : \quad \alpha \geq 0, \alpha \neq 1 \\ -\sum\limits_{m=1}^{M} p_m \log_2 p_m & : \quad \alpha = 1 \end{cases} . \tag{1}$$

Later, CHINTSCHIN [2] and FADDEJEW [3] formulated an axiomatic characterization of H. The mutual information for two discrete random variables ξ and η is then defined as $I(\xi,\eta) := H(\eta) - [H((\xi,\eta)) - H(\xi)]$, where $H(\eta)$ is the A-priori uncertainness of η and $H((\xi,\eta)) - H(\xi)$ is its remaining A-posteriori uncertainness, if ξ is known.
Let $P = \{p_m\}_{m=1}^{M}$, $Q = \{q_n\}_{n=1}^{N}$ and $S = \{s_{m,n}\}_{m=1,n=1}^{M,N}$ be probability distributions of the random variables ξ, η and (ξ,η), respectively. Let further η be uniformly distributed, with $q_n = N^{-1}$ for $n = 1, ..., N$. The mutual information $I_2(\xi,\eta) = H_2(\xi) + H_2(\eta) - H_2((\xi,\eta))$ has the following properties:

1. Symmetry: $I_2(\xi,\eta) = I_2(\eta,\xi)$
2. Limitation: $0 \leq I_2(\xi,\eta) \leq min(H_2(\xi), H_2(\eta))$
3. Independency: $I_2(\xi,\eta) = 0 \iff \xi$ and η are statistical independent.
4. Determination: $I_2(\xi,\eta) = H_2(\eta) \iff \eta$ is a function of ξ,
   $I_2(\xi,\eta) = H_2(\xi) \iff \xi$ is a function of η.

For arbitrary random variables ξ and η, $0 \leq I_2(\xi,\eta)$ holds if and only if at least η is uniformly distributed.

3. Matrix Calculus

In order to formulate the estimation algorithm for the mutual information, an appropriate matrix calculus [5] is presented. Since the algorithm requires uniformly distributed time series, the original measurement-data is transformed to absolute rank numbers.

$$\{\bar{R}(k)\}_{k=1}^{K} \equiv \begin{pmatrix} R_0(1) & R_0(2) & \cdots & R_0(K) \\ \vdots & \vdots & \ddots & \vdots \\ R_D(1) & R_D(2) & \cdots & R_D(K) \end{pmatrix}. \qquad (2)$$

The matrix in Eqn. 2 depicts the transformed $(D+1)$-dimensional measurement data. Each time series $R_d(k)\}_{k=1}^{K}$ is comprised of K sample-points. For the computation of the required entropy measure $H_2(\cdot)$, every row-vector $\{R_d(k)\}_{k=1}^{K}$ of the above matrix has to be considered separately. This is done by computing *rank-distance* matrices

$$\delta_{d,(j,k)} \equiv \| R_d(j) - R_d(k) \|, \quad j,k = 1,...,K \qquad (3)$$

for each $d = 0,...,D$. The K^2 entries of the matrix $\delta_{d,(j,k)}$ represent the distances between pairs of absolute rank-numbers in a particular sequence $\{R_d(k)\}_{k=1}^{K}$. The entries in $\delta_{d,(j,k)}$ are further used for computing *binary rank-distance* matrices

$$B_d = \begin{pmatrix} b_{d,(1,1)} & b_{d,(1,2)} & \dots & b_{d,(1,K)} \\ \vdots & \vdots & \ddots & \vdots \\ b_{d,(K,1)} & b_{d,(K,2)} & \dots & b_{d,(K,K)} \end{pmatrix}. \qquad (4)$$

For a predefined distance parameterϵ_d, with $0 \ll \epsilon_d \ll K$. The entries of B_d are set according to the rule

$$b_{d,(i,j)} := \begin{cases} 1 & : \quad \delta_{d,(i,j)} < \epsilon_d \\ 0 & : \quad \perp . \end{cases} \qquad (5)$$

Figure 2 shows the binary rank-distance matrix of a measurement signal. The binary matrices still contain all information about the statistic dependencies of the underlying time series, with respect to a predefined coarseness level ϵ_d.

Let $B_0, ..., B_D$, $D \in I\!N$ be binary matrices as introduced in Eqn. 4. The conjunction of multiple binary matrices is defined as follows:

$$\bigwedge_{d=0}^{D} B_d := (b_{0,(i,j)} \wedge \dots \wedge b_{D,(i,j)}), \quad i,j = 1,...,K. \qquad (6)$$

The correlation integral $C_{\bar{\epsilon},K}$ is now obtained as the relative weight of the resulting binary matrix:

$$C_{\bar{\epsilon},K} = \frac{1}{K^2} \sum_{i,j=1}^{K} \left(\bigwedge_{d=0}^{D} B_d \right) = \frac{1}{K} + \left[\frac{2}{K^2} \sum_{i<j} \left(\bigwedge_{d=0}^{D} B_d \right) \right]. \qquad (7)$$

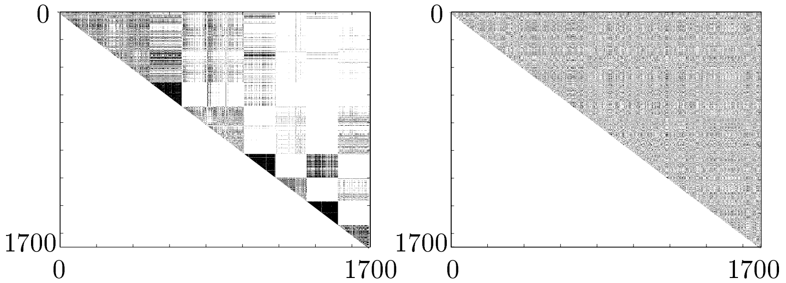

Figure 2: Left: Binary rank-distance matrix B of a measurement signal. Right: Binary matrix of a pure stochastic process, showing no statistic dependencies. Binary rank-distance matrices of signals without statistic dependencies are equally gray. The gray-level increases if the distance parameterϵ decreases.

Finally, an approximation for the entropy measure $H_2(\xi)$ of a (D+1)-dimensional random variable with its realizations $\{\bar{x}(k)\}_{k=1}^{K}$ is

$$H_2(\xi) \approx -\log_2(C_{\bar{\epsilon},K}) = -\log_2\left(\frac{1}{K} + \left[\frac{2}{K^2}\sum_{i<j}\left(\bigwedge_{d=0}^{D} B_d\right)\right]\right). \qquad (8)$$

This approximation is further used to obtain a measure for the mutual information $I_2(\xi,\eta) = H_2(\eta) + H_2(\xi) - H_2((\xi,\eta))$.

4. Nonlinear Time Series Analysis of Combustion Pressure Data

An example for the demonstration of the introduced method is taken from the field of automotive engineering. One particular point of interest from the viewpoint of combustion-control, is the determination of the 50% energy-conversion-point (ECP) solely from combustion-pressure data[4]. The 50%-ECP is defined as the crank angle position at which 50% of the fuel-mass in the cylinder has chemically reacted during the course of combustion [1].

Figure 3(a) depicts a small portion of the set of in-cylinder curves used for the calculation of the GMIFs. In Fig. 3(b), multiple iterations of the General Mutual Information between the in-cylinder pressure and the 50%-ECP are plotted against the crank-angle. Considering one particular iteration, each value of the GMIF is computed from all measurements of the in-cylinder pressure curve at a certain crank-angle position. For instance,at position +21° the first iteration of the GMIF reaches its maximum. This position is considered the most relevant point of the in-cylinder pressure with respect to the determination of the 50%-ECP. For the computation of further points, this particular position is assumed to be known, while an additional input is taken from the set of sample-points and varied. Hence, the next input-sequence is now a realization of the two-dimensional random variable $(\xi_{x_i}, \xi_{x_{i_1}})$. Continuing with this

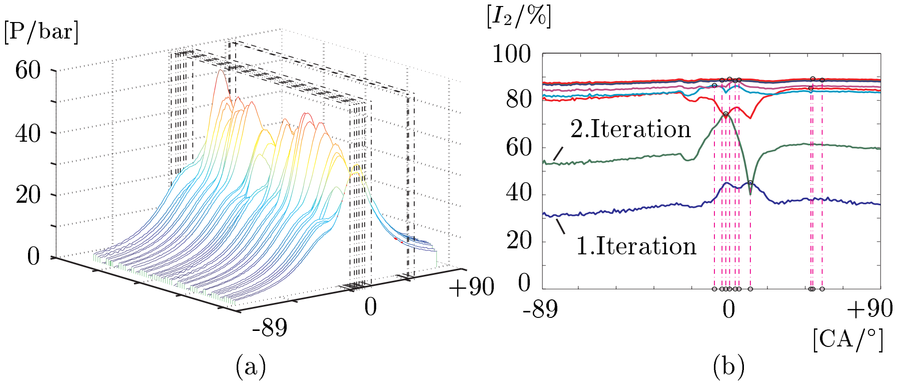

Figure 3: Identification of relevant points from a set of in-cylinder pressure curves.

strategy , an ordered sequence of points with the highest information content, can be successively identified. As depicted in Fig. 3(b), the normalized mutual information $I_2(\xi, \eta)$ is converging to its upper limit 1.0 when considering larger sets of sample-points. This procedure terminates when the maximum gain of mutual information between successive iterations drops below a predefined threshold.

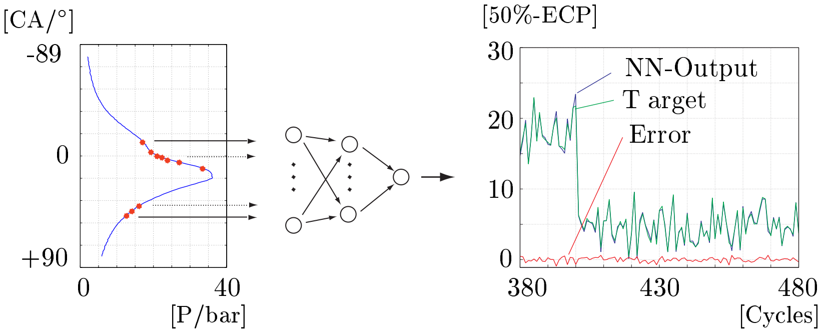

Figure 4: Presentation of the input-vector with the highest information content to a Neural Network.The neural structure has previously been trained with the identified input-vector taken from a separate training-set to determine the 50%-ECP.

Figure 4 depicts the presentation of the in-cylinder pressure curve at the identified crank-angle positions to a previously trained neural network. It can be observed, that the selected points are not equally distributed. Hence, we can draw the conclusion that training a neural structure with equally spaced input-vectors is not a reasonable choice for the investigated problem.

5. Conclusion

In this paper we demonstrated the application of the general mutual information as a quantitative method for measuring the information content of sample-points of the in-cylinder pressure curve with respect its according 50%-ECP. This general approach has been employed for identifying a sequence of most relevant sample-points in the pressure-signal of a combustion-engine.

The concept of mutual information can be utilized for the reduction of the input dimension of neural networks. In our case, we were able to reduce the size of the input-vector from 34 to ten sample-points. This maintained the quality of the solution and increased the computational performance.

Due to the generality of this approach, the concept of a mutual information can be applied for the nonlinear analysis of arbitrary data-sets. The introduced method represents a general and constructive framework for the preprocessing of neural network training-data.

References

[1] M. Bargende. Verbrennungs- und Ladungswechselanalyse. In *Stuttgarter Symposium Kraftfahrwesen und Verbrennungsmotoren*, pages M8.1–M8.16, Stuttgart, 1995.

[2] A. J. Chintschin. Der Begriff der Entropie in der Wahrscheinlichkeitsrechnung (russ. Orginalarbeit). In *Arbeiten zur Informationstheorie I*, volume 2, pages 7–29. Deutscher Verlag der Wissenschaften, Berlin, 1961.

[3] D.K. Faddejew. Zum Begriff der Entropie eines endlichen Wahrscheinlichkeitsraumes (russ. Orginalarbeit). In *Arbeiten zur Informationstheorie I*, volume 2, pages 86–90. Deutscher Verlag der Wissenschaften, Berlin, 1956.

[4] R. O. Müller. *Modernes Motormanagement mit Neuronalen Netzen*. PhD thesis, University of Würzburg, Department of Computer Science, 1998.

[5] B. Pompe. Measuring statistical dependencies in a time seris. *J Stat. Phys.*, 73:587–610, 1993.