VARIABLE SELECTION BASED ON MUTUAL INFORMATION FOR SOFT SENSORS APPLICATIONS

Francisco Souza, Rui Araújo, Symone Soares, and Jerôme Mendes *,1

* *Institute for Systems and Robotics (ISR-UC), and*
Department of Electrical and Computer Engineering (DEEC-UC),
University of Coimbra, Pólo II, PT-3030-290 Coimbra
fasouza@isr.uc.pt, rui@isr.uc.pt, symonesoares@isr.uc.pt,
jermendes@isr.uc.pt

Abstract: This paper cover a new method for variable selection in Soft Sensors design for industrial applications. We propose the use of Mutual Information for variable selection and to exclude redundant variables. As evaluation of quality model, we use a new criterion of tracking precision called relative variance tracking precision in parallel with the root mean square criterion. The proposed methodology was successfully applied to infer the total nitrogenous T_N in a wastewater treatment system Benchmark.

Keywords: Soft Sensors, Variable Selection, Mutual Information, Neural Networks

1. INTRODUCTION

Data-driven Soft Sensors (DDSS) are inferential models that use available on-line sensor measures for on-line estimation of process variables which cannot be automatically measured at all, or can only be measured at high cost, sporadically, or with high delays (e.g. laboratory analysis). The construction of DDSS models is based on measurements which are recorded and provided as historical data. The approximation methods that serve as the basis for SS are empirical predictive models such as Multi-Layer Perceptron (MLP), Support Vector Machines (SVM). DDSS are valuable tools to many industrial applications such as refineries, pulp and paper mills, wastewater treatment systems, just to give few examples (Fortuna *et al.*, 2006).

Pre-processing is an essential step for a correct development of the DDSS, the goal of this stage is to treat the data in such a way, that it can be effectively processed by the model. The usual steps in this phase are the handling of missing data, outliers detection and replacement, input variable selection (IVS) or feature selection, handling of drifting data and detection of delays between the particular variables (Kadlec *et al.*, 2009). When data is efficiently treated and input variables are correctly selected, the DDSS model can reproduce correctly the input-output relationship. Methods for selection of input variables can be classified in two classes: filter methods and the wrapper methods (Kohavi and John, 1997). Filter methods use a statistical measure to classify the variables, according to their influence and relevance on the target variable. On the other hand, wrapper methods use the learning model as the basis for selection. Often wrapper method may achieve more accurate prediction results because variables selection will take into account the approximation model. Thus, variables selection will be performed such that prediction error is minimized. However, filter methods are more generic since they only select variables for the prediction setting, not only making them suitable to understand the process, but also to leaving open the options for choosing the

[1] This work was supported by Mais Centro Operacional Program, financed by European Regional Development Fund (ERDF), and Agência de Inovação (AdI) under Project SInCACI/3120/2009;

Francisco Souza is supported by Fundação para a Ciência e a Tecnologia (FCT) under PhD Fellowship SFRH/BD/63454/2009. Jérôme Mendes is supported by Fundação para a Ciência e a Tecnologia (FCT) under PhD Fellowship SFRH/BD/63383/2009.

type of approximation model that will be subsequently used to construct the SS.

Mutual information (MI), is a high order statistical parameter that is having growing application in the design of filter methods for variable selection. The main advantages of MI are capacity to measure the dependency between variables, including nonlinear dependency, the robustness to noise and invariance to nonlinear data transformation (Chow and Huang, 2005). However, it is difficult to use MI in high dimensional spaces as required for analysis and modeling in problems involving a large of variables. This is due to the difficult of estimating the probability density (PDF), that is the basis of MI.

To solve this problem, some authors have developed variable selection algorithms, for classification problems, that can be extended for regression problems, based on two-dimensional (2D) approach where variable selection is performed by analyzing MI of pairs of variables (Battiti, 1994; Peng *et al.*, 2005; Chow and Huang, 2005; Kwak and Choi, 2002). A greedy selection approach was proposed by (Battiti, 1994), selecting relevant variables and at the same time excluding redundant variables according a pre-determined factor. Variants of Battiti's Mutual Information Feature Selector (MIFS) are: MIFS with Uniform Distribution (MIFS-U) (Kwak and Choi, 2002), the min-redundancy max-relevance (mRMR) criterion (Peng *et al.*, 2005). But these methods fail to detect the best subset, because do not take into consideration the relation among variables.

The common approaches in variable selection algorithm perform a forward search procedure using a certain criterion as stop, like maximum mutual information criterion, but these approaches can be expensive computationally when input space is large, making the search hard expensive. A idea is that before apply a forward search procedure, be applied a pre-processing step that remove redundant variables and remains that are essential for the model.

This paper proposes a input variable selection (IVS) algorithm based on mutual information. The algorithm is divided in two steps. First, it is performed the exclusion of redundant variables using mutual information criterion and after is performed a forward search procedure based on high dimensional mutual information.

Moreover, the paper proposes and demonstrates the use of the VS method to estimate the total nitrogen T_N in the effluent in a benchmark for watstewater treatment system.

This paper is organized as follows. Section 2 gives the mathematical definition of the problem of input variable selection. Section 3 presents the mathematical definition of mutual information and presents how to estimate it through a K-Nearest Neighbor method. The new variable selection algorithm proposed in this paper is presented in Section 4. Section 5 presents the RVTP evaluation criterion. Section 6 presents experimental results. Finally, Section 7 gives concluding remarks.

2. INPUT VARIABLE SELECTION

The problem of IVS consists on analyzing all available the inputs for a system and choose a subset of variables that are adequate to be used for inputs in order to develop a model of the system for purposes such as classification, prediction, or control. If unnecessary variables are kept in the model, noise may be introduced into the model and the overall results may be poorer than if only the required inputs are used. Moreover, if irrelevant variables are deleted from the model, the soft sensor accuracy can be improved (Qin, 1997). Therefore, it is important to select a subset of process variables that are truly relevant to the predicted variables. Below, the mathematical definition of the IVS problem is given.

2.1 *IVS Problem Statement*

The IVS problem can be described mathematically as follows. For any set of elements $A = \{a_1, \ldots, a_n\}$, define the ν operator that transforms A into vector $\mathbf{a} = \nu(A) = [a_1, \ldots, a_n]^T$. Only ordered sets will be considered in this paper. Conversely, $A = \nu^{-1}(\mathbf{a})$. A function G receives input from variables belonging to set $U = \{u_1(t), u_2(t), \ldots, u_p(t)\}$,

$$y(t) = G(\mathbf{u}), \tag{1}$$

where $\mathbf{u} = \nu(U)$. It is assumed that G can be a linear or nonlinear mapping. To estimate G, it is assumed that a set $X = \{x_1(t), x_2(t), \ldots, x_n(t)\}$ of measurement variables is available. It is assumed that the most appropriate x_i variables can be selected during the IVS design. It is assumed that:

$$U \subseteq X. \tag{2}$$

The goal of IVS is to select the best subset of variables

$$S \subseteq X, \tag{3}$$

that most adequately represent the information contained in the real input variables from U. Hence, an approximation model for G (1) can be written as:

$$\widehat{y}(t) = F(\mathbf{s}; \theta), \tag{4}$$

where F is a functional mapping parameterized by θ, and $\mathbf{s} = \nu(S)$.

In the approach proposed in this paper, the variable selection is performed as the first step excluding redundant variables. Then, in a second step, a high dimensional mutual information criterion is employed in a forward search procedure to select the best variables. Often, mutual information analysis in variable selection methods is performed between pairs of variables. The approach followed in this paper constitutes

a more realistic analysis alternative to mutual information since the analysis is performed directly in the high-dimensional space of the set of all relevant variables that are candidate for the selection procedure.

3. MUTUAL INFORMATION

Process engineers are often eager to find the optimal levels of process variables that make the key quality variable as close to its target as possible (Jun *et al.*, 2009). Some studies have used techniques based on variance such as principal component analysis (PCA) to select these variables (Warne *et al.*, 2004). These methods are designed for linear models, so they can not be the best choice for non-linear modeling. A PLS method is applied in (Jun *et al.*, 2009; Fortuna *et al.*, 2006) to select best variables. This method shows good results when the model used is linear and the data is multicolinear and noised. Recently the use of MI for variable selection in non-linear problems is growing. Mutual Information is a general correlation measure that unlike the correlation coefficient can be generalized to all kinds of probability distributions.

Below, it is given the mathematical definition of multi-dimensional MI. Also, it is discussed how MI can be estimated in high dimensional spaces using a K-nearest neighbors (KNN) based approach.

3.1 *Mutual Information*

Mutual Information is a non-linear measure of dependency between variables. It can be calculated through entropy measurements (Cover and Thomas, 1991). Let $x_1, \ldots, x_n, y$ be random variables, $\mathbf{x} = [x_1 \ldots x_n]^T$. In this paper, it will be assumed that $x_1, \ldots, x_n$ are inputs and y is an output of a system. The Mutual Information of y and $\mathbf{x}$ is defined as the amount of information that input $\mathbf{x}$ contains about output y, and can be calculated as:

$$\begin{aligned} I(\mathbf{x}; y) = H(y) + H(x_1) + \ldots + H(x_n) \\ - H(y, x_1, \ldots, x_n). \end{aligned} \tag{5}$$

$H(y)$, $H(x_1)$, ..., $H(x_n)$ and $H(y, \mathbf{x})$ are the Shannon entropy (Cover and Thomas, 1991). The Shannon entropy of a random variable z_1 is defined as:

$$H(z_1) = -\int_{z_1} f(z_1) \log[f(z_1)] dz_1. \tag{6}$$

In an n-dimensional space of n random variables $z_1, \ldots, z_n$, the multi-dimensional entropy is defined as:

$$\begin{aligned} H(z_1, \ldots, z_n) = -\int_{z_1} \ldots \int_{z_n} f(z_1, \ldots, z_n) \\ \times \log[f(z_1, \ldots, z_n)] dz_1 \ldots dz_n, \end{aligned} \tag{7}$$

where $f(z_1)$ and $f(z_1, \ldots, z_n)$ are the probability density function (PDF) of z_1, and the joint PDF of $z_1, \ldots, z_n$, respectively. The base of the logarithm determines the units in which information is measured. Natural logarithms will be used in the sequel, so that entropy will be measured in nats.

3.2 *Mutual Information Estimation*

The PDF estimation is generally performed using an histogram approach (Ludwig. *et al.*, 2009), but this is very computational expensive and non reliable, generating large errors in high dimensional problems. An alternate approach is to use Parzen window methods, but for high dimensional space such methods become computational expensive, and it becomes difficult choose the correct window size.

In the method proposed in this paper, PDF estimation is performed using the KNN approach proposed by (Kraskov *et al.*, 2004). Assume that a set $\mathcal{Z} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, m\}$ of m samples $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ of random variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is available, where $\mathbf{x}_i$, and $\mathbf{y}_i$ may be vectors or scalars, as special case. Define the norm of $||(\mathbf{u}, \mathbf{v})||_{\max} = \max\{||\mathbf{u}||, ||\mathbf{v}||\}$, where $||\cdot||$ denotes the euclidean norm. Let $N_k(i)$ be the set of k nearest neighbor samples of $(\mathbf{x}_i, \mathbf{y}_i)$ with respect to the norm $||\cdot||_{\max}$, and let:

$$\varepsilon_{\mathbf{x}}(i) = \max\{\| \mathbf{x}_i - \mathbf{x}'_i \| \mid (\mathbf{x}'_i, \mathbf{y}'_i) \in N_k(i)\}, \tag{8}$$

$$\varepsilon_{\mathbf{y}}(i) = \max\{\| \mathbf{y}_i - \mathbf{y}'_i \| \mid (\mathbf{x}'_i, \mathbf{y}'_i) \in N_k(i)\}, \tag{9}$$

$$\varepsilon(i) = ||\varepsilon_{\mathbf{y}}(i), \varepsilon_{\mathbf{x}}(i)||_{\max}, \tag{10}$$

where $\mathbf{z}' = (\mathbf{x}'_i, \mathbf{y}'_i)$ is the k^{th} nearest neighbour of $\mathbf{z}_i$, according with the maximum norm. Taking into account (10) we can count the number of $n_{\mathbf{x}}$ points whose distance from $\mathbf{x}_i$ is strictly less than ε, and similarly the number of points, $n_{\mathbf{y}}$, whose distance from $\mathbf{y}_i$ is strictly less than ε.

In this way, was shown by (Kraskov *et al.*, 2004) that MI can be estimated by:

$$\begin{aligned} &\widehat{I}(\mathbf{x}; \mathbf{y}) = \\ &\psi(k) + \psi(m) - \frac{1}{k} - \frac{1}{m} \sum_{i=1}^{m} \{\psi[n_{\mathbf{x}}(i)] + \psi[n_{\mathbf{y}}(i)]\}, \end{aligned} \tag{11}$$

where ψ is *digamma* function (Kraskov *et al.*, 2004). For n random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}, \mathbf{y}$ the following extension of (11) holds:

$$\begin{aligned} &\widehat{I}(\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_{n-1}; \mathbf{y}) = \\ &\psi(k) + (n-1)\psi(m) - \frac{1}{m} \sum_{i=1}^{m} \big\{\psi[n_{\mathbf{x}_1}(i)] \\ &+\psi[n_{\mathbf{x}_2}(i)] + \ldots + \psi[n_{\mathbf{x}_{n-1}}(i)] + \psi[n_{\mathbf{y}}(i)]\big\}. \end{aligned} \tag{12}$$

An advantage of the above KNN-based method is that it does not simply replace entropies in (5) with their estimates, but it is designed to cancel errors of individual entropy estimates. A practical drawback of the KNN-based approach is that the estimation accuracy depends on the value of k and there seems no systematic strategy to choose the value of k appropriately

(Suzuki *et al.*, 2008). With a small value for k, the estimator has a large variance and a small bias, whereas a large value of k leads to small variance and large bias. In this paper, a range of $k = 6, \ldots, 20$ has been used, and the final MI estimate results from the mean of the estimates for all the values of k in this range.

4. VARIABLE SELECTION ALGORITHM

This section proposes the new input selection algorithm based on high dimensional mutual information by minimal redundancy (HDMIVSR). Assume there is available a data-set $\mathcal{D} = \{(\mathbf{x}(t), y(t)) : t = 1, \ldots, N\}$ of measurements of input and output variables for N time instants $t = 1, \ldots, N$, where $\mathbf{x} = \nu(X) = [x_1, \ldots, x_n]^T$, and $X = \{x_1, \ldots, x_n\}$ is the set of input variables that will undergo the selection procedure for which there is available the measurements included in $\mathcal{D}$. ν is an operator that transforms a set of variables X into the vector whose components are the variables of the set.

The variable selection algorithm proposed here can be divided in two main steps: (1) detection and exclusion of redundant variables, and (2) a forward search procedure, using MI (5) as criterion of selection, until a stop condition is meet. To estimate MI the multidimensional estimator (12) will be employed.

In this paper the redundancy between two random variables x_i and x_j is defined as the following redundancy coefficient:

$$R(x_i, x_j) = \frac{I(x_i; x_j)}{H(x_i) + H(x_j)}. \tag{13}$$

$R(x_i, x_j)$ takes values between 0 and a maximum value of

$$R_{\max}(x_i, x_j) = \frac{\min[H(x_i), H(x_j)]}{H(x_i) + H(x_j)}, \tag{14}$$

where $R(x_i, x_j) = R_{\max}(x_i, x_j)$ corresponds to a high redundant variable and a $R(x_i, x_j) = 0$ means that the two variables are independent. Using (13), and (14), a normalized version of (13) that can take values between a minimum of 0 and a maximum of 1, can be written as follows:

$$\widehat{R}(x_i, x_j) = \frac{R(x_i, x_j)}{R_{\max}(x_i, x_j)} = \frac{I(x_i; x_j)}{\min[H(x_i), H(x_j)]}. \tag{15}$$

The algorithm for removal of redundant variables works as follows. For every pair of input variables (x_i, x_j), such that $\rho(x_i, x_j) > K$, the variable of the pair that has the lowest influence on the output, y, is removed. K is a free parameter. A parameter K can variate between 0 and 1, a typical adequate value for K is 0.4. The degrees of influence of x_i and x_j on the output are estimated using (15) as $\widehat{R}(x_i, y)$, and $\widehat{R}(x_j, y)$, respectively. The HDMIVSR algorithm is described as follows:

I) (Initialization) Set $X \leftarrow$ "Initial set of n variables", and $S \leftarrow$ "empty set"; set $k = 1$.

II) (Detect redundant variables) Set $X_R \leftarrow$ "Set of selected redundant variables to be excluded"; set $X \leftarrow X \backslash X_R$.

III) (Computation of MI with the output variable) for each variable $x \in X$ compute $I(x; y)$.

IV) (Choice of first feature) Find the variable x for which $I(x; y)$ is maximum, i.e. $x = \underset{u \in X}{\text{argmax}} \{I(u; y)\}$; set $I^k_{\max} = I(x; y) = \max_{u \in X} \{I(u; y)\}$; set $I^{k-1}_{\max} = I^k_{\max}/2$; set $X \leftarrow X \backslash \{x\}$; set $S \leftarrow \{x\}$;

V) (Forward Selection) repeat until $I^{k-1}_{\max} > I^k_{\max}$

  i) $k = k + 1$.

  ii) (Computation of MI) Find variable $x \in X$ that, when incrementally added to the set of selected variables, S, maximizes the high-dimensional MI $I[\nu(S \cup \{x\}); y]$, i.e. $x = \underset{u \in X}{\text{argmax}} \{I[\nu(S \cup \{u\}); y]\}$;

  iii) set $I^k_{\max} = I[\nu(S \cup \{x\}); y]$; set $X \leftarrow X \backslash \{x\}$; set $S \leftarrow S \cup \{x\}$;

VI) Output the set S containing the selected variables.

5. EVALUATION CRITERION

The most common indicator of quality models is the root mean square error (MSE), but the MSE does not measure the tracking precision. (Li *et al.*, 2009) proposed the use of relative variance tracking precision, RVTP, for soft sensors applications:

$$RVTP = 1 - \frac{\sigma^2_{\text{error}}}{\sigma^2_{\text{measurement}}}, \tag{16}$$

where σ^2_{error} is the variance of the prediction error (difference between the model prediction and the measurement value), and $\sigma^2_{\text{measurement}}$ is the output measurement variance, both computed considering all the samples of the complete test set. RVTP (16) indicates the tracking precision between output and the model when the output changes. It is a measure of how precisely the SS output remains with enough precision when the value of the output changes. When RVTP is less than zero, the precision of SS is very low. The closer RVTP approaches 1, the more accurately the SS tracks the real process (Li *et al.*, 2009).

6. EXPERIMENTAL RESULTS

This section presents experimental results of a case study concerning the estimation of the total nitrogen T_N at the effluent of a wastewater treatment plant (WWTP). The WWTP study was conducted using the Benchmark Simulation Model No. 2 (BSM2) (Jeppsson *et al.*, 2006). BSM2 is a platform-independent WWTP simulation environment defining a plant layout, a process model, influent data, test

Table 1.

Variables Description	
Dissolved Oxigen in effluent	u_1
Dissolved Oxigen in reactor 1	u_2
Dissolved Oxigen in reactor 2	u_3
Dissolved Oxigen in reactor 3	u_4
Dissolved Oxigen in reactor 4	u_5
Dissolved Oxigen in reactor 5	u_6
T_{SS} (Suspended solids) in influent	u_7
T_{SS} (Suspended solids) in effluent	u_8
T_{SS} (Suspended solids) in reactor 1	u_9
T_{SS} (Suspended solids) in reactor 2	u_{10}
T_{SS} (Suspended solids) in reactor 3	u_{11}
T_{SS} (Suspended solids) in reactor 4	u_{12}
T_{SS} (Suspended solids) in reactor 5	u_{13}
T_{SS} (Suspended solids) in Aerated Settler	u_{14}
T_{SS} (Suspended solids) in settler	u_{15}
Settler sludge heigth	u_{16}
Q_{in} (Influent flow rate)	u_{17}
Q_e (Effluent flow rate)	u_{18}
Q_{pu} (Primary settler underflow flow rate)	u_{19}
Influent temperature	u_{20}
Effluent temperature	u_{21}
Instantaneous sludge wastage rate	u_{22}
Instantaneous methane production	u_{23}
Instantaneous total gas flow normalized to P-atm	u_{24}
Instantaneous pumping energy	u_{25}

Table 2.

Selected variables for different values of K		
	Group	Selected Variables
$K = 1.00$	G_1	u_{24},u_{23},u_1,u_2,u_3,u_6,u_5,u_4,u_{14}, u_{11},u_{10},u_9,u_{12},u_{13},u_{20},u_{21}, u_{15},u_8,u_{25},u_{22},u_{17},u_{18},u_{19},u_7,u_{16}
$K = 0.90$	G_2	u_{23},u_{21},u_{20},u_{13},u_{12},u_{11},u_{14},u_{15}, u_8,u_1,u_7,u_6,u_5,u_{10},u_9, u_{22},u_{19},u_{18},u_2,u_3,u_4,u_{16}
$K = 0.80$	G_3	u_{23},u_{21},u_{20},u_{13},u_{12},u_{11},u_{14},u_{15}, u_8,u_1,u_7,u_6,u_5,u_{10},u_{22}, u_{18},u_2,u_3,u_4,u_{16}
$K = 0.70$	G_4	u_{23},u_{21},u_{20},u_{13},u_{12},u_{14},u_{15},u_8, u_1,u_7,u_6,u_5,u_{10},u_{22},u_{18}, u_2,u_4,u_{16}
$K = 0.60$	G_5	u_{23},u_{21},u_{20},u_{13},u_{14},u_{15},u_1,u_7, u_6,u_2,u_{22},u_{10},u_4
$K = 0.50$	G_6	u_{23},u_{21},u_{20},u_{13},u_{14},u_1,u_7,u_6,u_2,u_{22}
$K = 0.40$	G_7	u_{23},u_{21},u_{20},u_{13},u_7,u_1,u_6,u_2,u_{22}
$K = 0.30$	G_8	u_{23},u_{21},u_{20},u_{13},u_7,u_{22},u_2
$K = 0.20$	G_9	u_{23},u_{21},u_{13}
$K = 0.10$	G_{10}	u_{13}

procedures and evaluation criteria. The benchmark is evaluated for two years with acquisition data for the variables being available with a 1 hour sampling interval. There are 25 input variables in the data-set which are candidates for the variables and delay selection problem. We define the set of input variables as: $U = [u_1, u_2, \ldots, u_{25}]^T$ and the output, T_N, as y. The description of each variable, in U is given by Table 1.

The HDMIVSR algorithm proposed in Section 4 was applied to estimate T_N. Different values for K were tested to identify what is the best value to perform the prediction. The selecteds inputs variables are shown in Table 2.

Table 3.

Prediction results for each group				
Group	RMSE Train	RMSE Evaluation	RVTP Train	RVTP Evaluation
G_1	0.0100	0.0323	0.91	0.62
G_2	0.0116	0.0175	0.79	0.65
G_3	0.0132	0.0322	0.76	0.67
G_4	0.0129	0.0312	0.76	0.60
G_5	0.0139	0.0244	0.74	0.63
G_6	0.0206	0.0620	0.72	0.54
$\mathbf{G_7}$	0.0172	0.0471	0.65	0.66
G_8	0.0246	0.0335	0.55	0.56
G_9	0.0319	0.0289	0.42	0.57
G_{10}	0.0329	0.0433	0.40	0.37

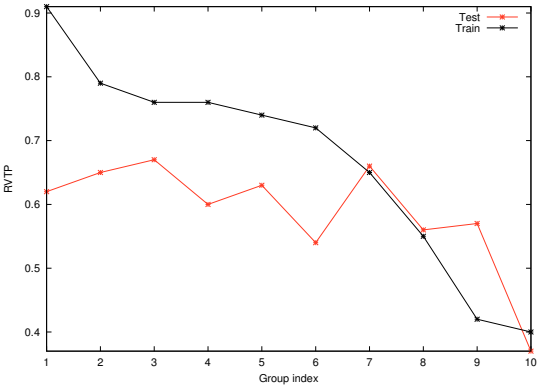

Figure 1. RVTP value for each group (see Table 2 for the corresponding values of K).

Then it was applied a MLP model with architecture $N \times 5 \times 1$, where N is the number of input variables and 5 is the number of hidden nodes. Sigmoidal activation functions were used in the hidden layer, and linear activation functions were used in the output layer. The training process uses the Nguyen Window algorithm for weights initialization, and the Levenberg-Marquardt training method with Mean Squared Error (MSE) as performance function, and minimum gradient threshold, $T = 10^{-10}$, as stop criterion. As evaluation criterion will be used root mean square error $RMSE$ in train and test set. The results are shown in Table 3.

Analyzing the Figs. 1 and 2 is possible to conclude that the G_7 group ($K = 0.4$) achieves the best trade-off between the number of selected input variables, and the train and validation accuracy. Thus, it is possible to conclude that with $K = 0.4$ the model achieves the best results. Fig. 3 presents the prediction results.

7. CONCLUSION

This paper proposed an algorithm to perform variable selection in prediction settings. The main advantage of this algorithm in comparison with previously proposed methods is the exclusion of redundant variables before applying the forward variable search procedure. Thus, by reducing the search space, the forward search procedure becomes faster. The new method was

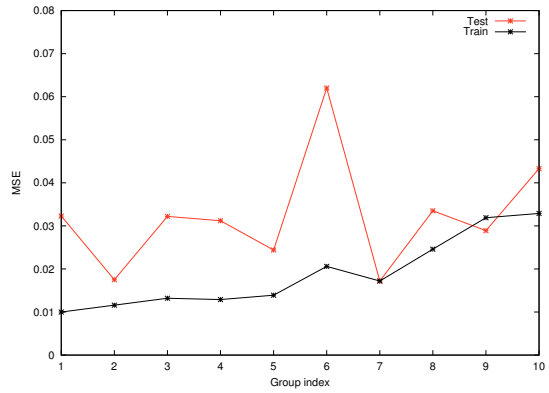

Figure 2. MSE value for each group (see Table 2 for the corresponding values of K).

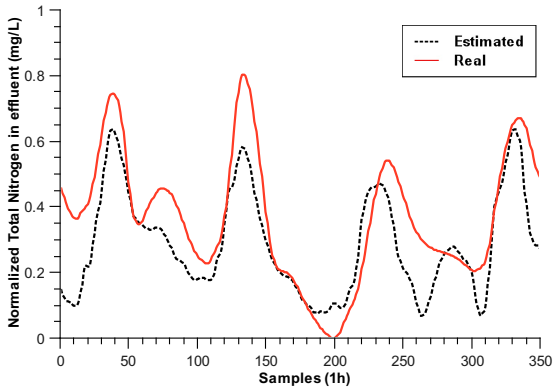

Figure 3. T_N estimation using the variables selected by HDMIVSR algorithm for the G_7 group ($K = 0.4$).

applied for a waste water treatment plant data-driven soft sensor application. The results of applying the selected variables in the prediction setting have shown good prediction accuracy. In this case study, it was concluded that removing irrelevant variables with the factor $K = 0.4$ is a good choice.

8. REFERENCES

Battiti, Roberto (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks* **5**(4), 537–550.

Chow, Tommy W. S. and D. Huang (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural Networks* **16**(1), 213–224.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing.

Fortuna, L. S. Graziani, A. Rizzo and M. G. Xibilia (2006). *Soft Sensors for Monitoring and Control of Industrial Processes*. Vol. 78 of *Advances in Industrial Control*. 1 ed.. Springer.

Jeppsson, U., C. Rosen, J. Alex, J. Copp, K.V. Gernaey, M. N. Pons and P.A. Vanrolleghem (2006). Towards a benchmark simulation model for plant-wide control strategy performance evaluation of wwtps. *Water Science and Technology* **53**(1), 287 – 295.

Jun, Chi-Hyuck, Sang-Ho Lee, Hae-Sang Park and Jeong-Hwa Lee (2009). Use of partial least squares regression for variable selection and quality prediction. In: *Proc. International Conference on Computers Industrial Engineering (ICCIE 2009)*. Troyes, France. pp. 1302–1307.

Kadlec, Petr, Bogdan Gabrys and Sibylle Strandt (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering* **33**(4), 795–814.

Kohavi, Ron and George H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence Archive* **97**, 273–324.

Kraskov, Alexander, Harald Stögbauer and Peter Grassberger (2004). Estimating mutual information. *Phys. Rev. E* **69**(6), 066138.

Kwak, Nojun and Chong-Ho Choi (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks* **13**(1), 143–159.

Li, Xiuliang, Hongye Su and Jian Chu (2009). Multiple model soft sensor based on affinity propagation, gaussian process and bayesian committee machine. *Chinese Journal of Chemical Engineering* **17**(1), 95–99.

Ludwig. Oswaldo, Urbano Nunes, Rui Araújo, Leizer Schnitman and Herman Augusto Lepikson (2009). Applications of information theory, genetic agorithms and neural models to predict oil flow. *Communications in Nonlinear Science and Numerical Simulation* **14**, 52 – 67.

Peng, Hanchuan, Fuhui Long and Chris Ding (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(8), 1226–1238.

Qin, S. Joe (1997). Neural networks for intelligent sensors and control - practical issues and some solutions. In: *Neural Systems for Control* (Omid Omidvar and David L. Elliott, Eds.). pp. 215 – 236. Lecture Notes in Control and Information Sciences. Academic Press.

Suzuki, Taiji, Masashi Sugiyama, Jun Sese and Takafumi Kanamori (2008). Approximating mutual information by maximum likelihood density ratio estimation. *JMLR: Workshop and Conference Proceedings* **4**, 5–20.

Warne, K. G. Prasad, S. Rezvani and L. Maguire (2004). Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. *Engineering Applications of Artificial Intelligence* **17**(8), 871–885.