Variable and Time-Lag Selection using Empirical Data

Francisco Souza and Rui Araújo
Institute of 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

Abstract

The paper proposes a method to select the best variables and respective time-lags for industrial applications when the objective is the estimation of a target variable using the information content of empirical data. No further information is assumed about the process. The problem of jointly selecting the best variables and the respective time-lags is treated as a variable selection problem. This assumption implies an increase of input dimensionality and multicollinearity into input space. Then, a multidimensional mutual information estimator based on the l-nearest neighbor algorithm is used in a forward search procedure to select the best variables and and respective time-lags. To verify the performance of selected variables and delays, the method was successfully applied in two data sets. A least squares support vector machine was used as the main model for the soft sensor in both cases.

1 Introduction

Most industrial processes are equipped with online process sensors for an online supervision, monitoring, and control. However, there are critical variables that can not be measured with physical sensors, but only by laboratory analysis, thus leading to the lack of enough information about what occurs in real time. A real time estimator of these critical variables can lead to a better understanding about the operation of the process, allowing a fast decision making when necessary. This type of estimator, refereed in the literature as soft sensor, in many cases can be build using these online available sensor measurements. Thus, soft sensors are inferential models that use on-line available sensor measures for on-line estimation of 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), they are important tools for many industrial processes. However, to build a soft sensor model, it is not necessarily true that all the recorded variables are relevant or useful. Generally, the selection of the best variables for the prediction setting is done by manual selection, by system experts, and in few cases, only time lag selection is performed through automatic methods [15, 3]. For physically large and highly integrated processes, enumeration of candidate variables based on process insight may not be feasible [15]. In these cases it is necessary to perform automatic methods for variable and time lag selection, without help of system experts.

From the machine learning point of view, the input variable selection objective is to find the optimal subset of the whole set of possible input variables [4], so that the variables of this subset can be used as inputs in a prediction setting to correctly predict the output using a suitable model. However, for complex industrial processes, in most of the cases, there is also the issue of selecting the most adequate time-lag for each input variable. Thus, in this context the variable selection problem includes two sub-problems. The first one is the selection of best input variables, and the second one is to find respective time lags for each variable.

Methods for input variable selection can be classified into two classes: filter methods and the wrapper methods [6]. Filter methods use statistical measures 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 methods may achieve more accurate prediction results because variables selection will take into account the approximation model. Thus, variable selection will be performed in such a way 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 type of approximation model that will be subsequently used to construct the model.

In the literature, two different approaches for selection of variables and respective time-lags using filter and wrapper methods have been addressed. In the first approach, it is assumed that the best variables are known, remaining the selection of the best time-lag for each variable [3, 11]; generally the variables are selected by an expert, this is a good way for selection, but for generic applications or complex processes this analysis can be very complicated,

expensive and/or inaccurate. In the second approach, it is assumed that both the best variables and respective time-lag are unknown [2, 9, 12].

In [3] different methods for time-lag selection are used: the simple correlation method, the partial correlation method, Mallows Coefficient with linear and nonlinear models [15] and PLS based methods. However, most of these methods are designed for linear models, so they can not be the best choice for non-linear modeling. The Mallows Coefficient method with non-linear models is computationally expensive when the input dimensionality high. A linear time-lag model optimized by a genetic algorithm was proposed by [11] to perform delay selection, and it had a good performance to predicting the nitrogen oxides NO_X and oxygen O_2 in the combustion operation in industrial boilers. In [2] the proposed method first selects the best input variables by means of self-organizing map (SOM) and then selects the best time-lags using the Lipschitz quotients. This method fails in some aspects: the first is the selection of the number of neurons in Kohonen maps, that can bring different results, and the second is the time-lag selection after input variable selection which can bring wrong results because a variable with the correct time-lag can contain more information about the output than a variable with the incorrect time-lag [12].

Performing the selection of the time-lags of each variable before variable selection improves the final prediction performance. In [12] the time-lag of each variable was selected using the mutual information measure ([1], Chapter 2) and then a multilayer perceptron model was used to select the best set of input variables (with each variable having the best time-lags selected before). Some works have used the mutual information measure to select best input variables and best time-lags. In [9] a multidimensional mutual information measure, based on histogram estimator, and a multidimensional estimator approach based on the "maximum relevance minimum redundancy" (MRmR) principle [10] was used; And it has been shown that the method outperforms the linear methods based on correlation coefficient in the cases of study. However, this method assumes that a variable can appear just once and the best time-lag is selected for that variable.

For a better understanding about an industrial process (i.e. the knowledge about the best variables that affect a target variable), the best choice to select best variables and respective time-lags is by means of filter methods [3, 15, 12, 9], because they do not take into consideration the model used, but use only the information content of data. Moreover, some complex systems involve nonlinear interactions among variables, making not suitable the use of linear filter methods, that generally are simpler in terms of implementation and computational cost, but fail to measure non-linear interactions. Mutual information, that can be considered as a generalization of the linear correlation value, can measure nonlinear interaction among candidate of input variables and the target output, using information content into input-output data. However, the main difficulties are in the estimation of the mutual information in multidimensional spaces. The most commonly used method is the histogram approach [9, 12], but this method suffers from the problem of high computational costs and low estimation accuracy. The Parzen window estimator is used in [10], but the generalization for multidimensional space is computationally expensive. A suitable estimator to multidimensional mutual information, with low computational cost and one free parameter was proposed recently by [7], and it is based on the l-nearest neighbor algorithm.

The contributions of this work are: the problem of jointly selecting the best variables and the respective time-lags is treated as a single variable selection problem, where all variables and respective time-lags are ensembled in a single set of candidate variables. This assumption implies an increase of input dimensionality and multicollinearity into input space. The variable selection problem is solved with a forward search procedure, with the multidimensional mutual information measure, estimated by the l-nearest neighbor algorithm. Moreover, for this work it is not assumed any knowledge about the process. To verify the performance of selected variables and their respective time-lags, the method was successfully applied in two data sets. A least squares support vector machine was used as the main approximation model for the soft sensor in both cases.

The paper is organized as follows. Section 2 gives the motivation and mathematical definition of the problem of selection of input variables and respective time-lags. Section 3 presents the mathematical definition of mutual information and presents how to estimate it through a l-nearest neighbor method. The proposed procedure to select the variables and respective time-lags is presented in Section 4. Section 6 presents experimental results on two estimation problems. Finally, Section 8 gives concluding remarks.

2 Problem Formulation

2.1 Motivation

As discussed above the problem of selecting variables for prediction settings in complex industrial processes can be divided into two sub-problems. The first, is the problem of selecting the best subset of variables. The second problem is to find respective time lags for each variable. Assuming this formulation, the variables with the respective time-lags can be ensembled into a single set, and the variable selection algorithm chooses the best subset from the whole set. This paper adopts this approach: the problem of jointly selecting variables and the respective time-lags is transformed into a single unified variable selection problem. This assumption implies an increase of input dimensionality and multicollinearity into input space.

The following notation is used in the paper. For any set of elements (variables or constants) $A = \{a_1, \ldots, a_n\}$, ($a_i \in \mathbb{R}$, $i = 1, \ldots, n$), define the ν operator that trans-

forms A into vector $\mathbf{a} = \nu(A) = [a_1, \ldots, a_n]^T$. Only ordered sets will be considered in this paper. Conversely, $A = \nu^{-1}(\mathbf{a})$. Then, for any input vector, $\mathbf{x}$, the corresponding set of elements is $X = \nu^{-1}(\mathbf{x})$.

2.2 Mathematical Formulation

Assume that there is an original training dataset given by a set of exemplars $\mathcal{F}^* = \{(\mathbf{u}(k), y_d(k)) \mid k = 1, \ldots, m\}$, where $\mathbf{u}(k) = [u_1(k), \ldots, u_q(k)]^T \in \mathbb{R}^q$ and $y_d(k) \in \mathbb{R}$ are the vector of input variables and the output target at instant k, respectively. The set of input variables referred to instant k is given by $U^{(k)} = \nu^{-1}(\mathbf{u}(k)) = \{u_1(k), \ldots, u_q(k)\}$. Without loss of generality, the set $X^{(k)} = \{x_1(k), \ldots, x_n(k)\} = \{u_1(k), u_1(k - d_1^{(1)}), \ldots, u_1(k - d_{n_d}^{(1)}), \ldots, u_q(k), \ldots, u_q(k - d_{n_q}^{(q)})\}$ is defined as the corresponding input set with all variables with all possible time-lags, where $\mathbf{x}(k) = \nu(X^{(k)}) = [x_1(k), \ldots, x_n(k)]^T \in \mathbb{R}^n$, where $d^{(j)} = \{d_1^{(j)}, \ldots, d_{n_j}^{(j)}\}$ are the possible time lags for variable u_j. The modified training dataset is given by $\mathcal{F} = \{(\mathbf{x}(k), y_d(k)) \mid k = 1, \ldots, m\}$, where without loss of generality, it is assumed that the number of exemplars remains the same after inclusion of all possible time-lags for all variables. Without referring to a specific time instant, the following two sets of input variables can be defined: $U = \nu^{-1}(\mathbf{u})$, and $X = \nu^{-1}(\mathbf{x})$. The goal of the variable and delay selection procedure is to select a subset $S \in X$ of the most significant variables for the prediction setting.

Thus, it is possible to verify the increase of input dimensionality $|\mathcal{F}^*| \leq |\mathcal{F}|$ due to the ensemble of best variables and respective time-lags. While this mathematical formulation is restricted by the assumption that the possible delays are known, its practical relevance is very high. If only a maximum delay is known, a set of multiple possible delays up to that maximum can be used. In this definition, with the increase of possible delays and consequently the increase of input dimension, the complexity of the search algorithm will scale up, due the fact that the forward search procedure generally results in $\mathcal{O}(n^2)$ worst case computation costs.

3 Mutual Information

Mutual Information is a non-linear measure of dependency between variables ([1], Chapter 2). Differently from the correlation value (the most common filtering criterion used for variable and time-lag selection), mutual information can be expanded for multidimensional spaces, making it suitable as a variable selection criterion. The following subsections give the mathematical definition for multidimensional mutual information (MI) and discusses a MI estimator based on the k-nearest neighbor estimator.

3.1 Mathematical Definition

Let $p(\mathbf{x})$, $p(y_d)$ and $p(\mathbf{x}, y_d)$ be the probability density functions of the input vector $\mathbf{x}$, output y_d, and the joint probability density function of the input vector and the output, respectively. The information of the set X of input variables is given by the Shannon's entropy ([1], Chapter 2):

$$H(X) = \int_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \, d\mathbf{x}. \quad (1)$$

It is important to note that the more informative the set X is, the higher is its entropy value. The information shared by X and $Y_d = \{y_d\}$, or mutual information of X and Y_d, is given by:

$$I(X; Y_d) = \int_{\mathbf{x}} \int_{y_d} p(\mathbf{x}, y_d) \log \frac{p(\mathbf{x}, y_d)}{p(\mathbf{x}) p(y_d)} \, d\mathbf{x} dy_d. \quad (2)$$

Equation (2) is well analyzed when using the following entropy form ([1], Chapter 2):

$$I(X; Y_d) = H(X) + H(Y_d) - H(X, Y_d). \quad (3)$$

From analysis of Equation (3) it is seen that mutual information is the information shared between variables X and Y_d. 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 l-Nearest Neighbor Estimator

A possible estimator for multidimensional mutual information is based on the l-nearest neighbor algorithm and was proposed by [7]. Define variables $\mathbf{z}(k) = (\mathbf{x}(k), y_d(k)) = (x_1(k), \ldots, x_n(k), y_d(k))$ ($k = 1, \ldots, m$), the distances $\varepsilon(k)/2$ between $\mathbf{z}(k)$ and its l-th neighbor, and the distances $\varepsilon_x(k)/2$, $\varepsilon_{x_1}(k)/2$, ..., $\varepsilon_{x_n}(k)/2$, and $\varepsilon_{y_d}(k)/2$ between the same points projected into the subspaces of $\mathbf{x}$, x_1, ..., x_n, and y_d, respectively. For $k = 1, \ldots, m$, and $h = 1, \ldots, n$ (recall that n is the dimension of $\mathbf{x}(k)$), define $n_{x_h}(k)$ as the number of points $\mathbf{x}(j)$ ($j = 1, \ldots, m$) that obey $||x_h(k) - x_h(j)|| \leq \varepsilon_{x_h}(k)/2$, and define $n_{y_d}(k)$ as the number of points $y_d(j)$ ($j = 1, \ldots, m$) that obey $||y_d(k) - y_d(j)|| \leq \varepsilon_{y_d}(k)/2$.

Then, the mutual information can be estimated by:

$$\hat{I}(X; Y_d) = \psi(l) - (n-1)\left(\frac{1}{k} - \psi(m)\right) - \\ - \frac{1}{m} \sum_{k=1}^{m} \left(\sum_{h=1}^{n} n_{x_h}(k) + n_{y_d}(k) \right) \quad (4)$$

where ψ is the *digamma* function. An advantage of the above l-nearest neighbor method is that it does not simply replace entropy in (3) with their estimates, but it is designed to cancel errors of individual entropy estimates [7]. A practical drawback of the l-nearest neighbor-based approach is that the estimation accuracy depends on the value of l and there seems no systematic strategy to appropriately choose the value of l [14]. With a small value for l, the estimator has a large variance and a small bias, whereas a large value of l leads to small variance and large bias. In this work, the mean value of the mutual information values obtained for $l = 3, \ldots, 6$ has been used.

Input: Input set $\mathcal{F}*$ and $d^{(j)}$, for $j = 1, \ldots, q$;
Output: Selected set S of input variables and respective time-lags;
Create the ensembled set $\mathcal{F}$. (see Section 2.2);
$S := \emptyset$;
$J_{max}(0) := 0$;
$int := 0$;
$stop := 0$
while $stop \neq 1$ and $X \neq \emptyset$ **do**
  $int \leftarrow int + 1$;
  forall $x_i \in X$ **do**
    $J_i(int) := \hat{I}(S \cup \{x_i\}; Y_d)$;
    $selected := \arg\max_i J_i(int)$;
    $J_{max}(int) := J_{selected}(int)$;
  end
  if $J_{max}(int) < J_{max}(int-1)$ **then**
    $stop := 1$;
  else
    $S := S \cup \{x_{selected}\}$;
    $X := X \setminus \{x_{selected}\}$;
  end
end
return S;

Algorithm 1: Steps of the Variable and Time-Lag selection algorithm.

4 Variable and Time-Lag Selection Procedure

This section proposes the variable and time-lag selection based on multidimensional mutual information. The proposed variable selection algorithm can be divided into two main steps: (1) transformation of the set $\mathcal{F}*$ of possible variables and time-lags, into a single set $\mathcal{F}$, and (2) a forward search procedure, using the the MI estimator as cost function.

The advantage of the proposed method is that it does not take into consideration the particular model used for modeling/estimating the target variable, and can appropriately measure nonlinear interactions between the system variables, and select also the delay for each variable. The forward selection procedure has shown to be suitable due to the low computation cost and the necessity of few interactions to find the best solution.

Algorithm 1 describes the proposed procedure for the selection of variables and respective time delays. According to Algorithm 1, the ensembled input set $\mathcal{F}$ is constructed from the original input set $\mathcal{F}^*$ and the possible time-lags $d^{(j)}$ for each variable. This procedure is detailed in Section 2.2. With the set $\mathcal{F}$ constructed, a forward search procedure is applied to find the best variables. The first selected variable is one that has the maximum mutual information value, J_{max}, with the target output y_d. This variable is stored in the set of selected variables S and removed from the set of input variables X. At each iteration, a variable from the input set X is selected to be included in the output set S, until the estimated mutual information starts to decrease. A decrease in the mutual information, means that there are no more candidate input variables that can that can be added to the set S, and make S contain more information with respect to the target variable y_d. When such decrease occurs the procedure is finished. The output set S contains the best variables to predict y_d.

5 Least Squares Support Vector Machine

This section is gives a brief overview about the least squares support vector machine (LS-SVM) model used as the approximation model to implement the soft sensor. The output of the LS-SVM has the following form:

$$y(k) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}(k)) + b + e_t(k), \tag{5}$$

where $y \in \mathbb{R}$ is the estimated output, $b \in \mathbb{R}$ is a bias term, $\mathbf{w} \in \mathbb{R}^{n_h}$ is an unknown coefficient vector, and $\boldsymbol{\varphi}$ is a nonlinear variable mapping which transforms the original input $\mathbf{x} \in \mathbb{R}^n$ into a high-dimensional vector $\boldsymbol{\varphi}(\mathbf{x}) \in \mathbb{R}^{n_h}$. Consider the following constrained optimization problem with a regularized cost function:

$$\begin{aligned} &\min_{\mathbf{w},b,e_t} \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\tfrac{1}{2}\textstyle\sum_{k=1}^{m} e^2(k), \\ &\text{s.t } \; y(k) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}(k)) + b + e_t(k), \\ &\qquad k = 1, \ldots, m, \end{aligned} \tag{6}$$

where γ is a linearization constant and $K(\mathbf{x}(i), \mathbf{x}(j)) = \boldsymbol{\varphi}(\mathbf{x}(i))^T\boldsymbol{\varphi}(\mathbf{x}(j)), i, j = 1, \ldots, m$ is a kernel function. In this work the radial basis function kernel was employed: $K(\mathbf{x}(i), \mathbf{x}(j)) = \exp(-||\mathbf{x}(i) - \mathbf{x}(j)||/\sigma^2)$;

The problem (6) is solved using Lagrange multipliers and the solution is expressed in dual form [13]. The estimated $y(k)$ is given by:

$$y(k) = f(\mathbf{x}(k)) = \sum_{j=1}^{m} \alpha_j K(\mathbf{x}(k), \mathbf{x}(j)) + b \tag{7}$$

where α_j are the Lagrange multipliers. Further details about the model can be found in ([13], Chapter 3). The model was implemented using the toolbox available in the webpage [8].

6 Experiments

This section shows the performance of variable and time-lag selection on two real data-sets. In the first dataset, the objective is to estimate the butane concentration in the bottom flow of a debutanizer column, it was introduced by ([3], Chapter 5) and is available for download in the book's website. In the second experiment the objective is to infer the total chemical oxygen demand COD at the effluent of a wastewater treatment plant (WWTP). For the measurement of approximation quality performance it is used the root mean square error (RMSE), normalized mean square error (NMSE), maximum output error

Table 1: Description of the variables of Debutanizer Column.

Variables Description	
Top temperature at debutanizer column	u_1
Top pressure at debutanizer column	u_2
Reflux flow	u_3
Reflux flow	u_4
Flow to the next process	u_5
Bottom temperature 1 at debutanizer column	u_6
Bottom temperature 2 at debutanizer column	u_7
(C_4) concentration	y

(MAE) and the correlation coefficient between predicted and desired output, in the test data. For comparison purposes, the best variables using correlation coefficient procedures, similarly to the procedure used in ([3], Chapter 5) are also selected. Also, for comparison, the model is evaluated without variable and time lag selection, i.e. using all variables from $\mathcal{F}^*$ in the prediction setting. The type of model used for inference is the least square support vector machine (Sec. 5).

6.1 Debutanizer Column

For the (C_4) concentration estimation in a debutanizer column there are available seven candidates for input variable $U^{(k)} = \{u_1(k), u_2(k), u_3(k), u_4(k), u_5(k), u_6(k), u_7(k)\}$. Table 1 gives a detailed description about input variables. A block scheme of the debutanizer column is given in ([3], Appendix A). The maximum delay chosen for each variable is 8 samples, and the possible delays were spaced by 4 samples and started at zero. Thus, the ensembled candidates input set is given by $X^{(k)} = \{u_1(k), u_1(k-4), u_1(k-8), \ldots, u_7(k), u_7(k-4), u_7(k-8)\}$, which includes 21 candidate variables. The original number of samples is 2394, but due to the consideration of delays, this number is reduced for 2386. To apply the input variable and time-lag selection algorithm, the whole dataset was divided into a training data set of 2088 points, and a test dataset of 298 points.

Two algorithms are compared for variable and delay selection. The first method is based on the correlation coefficient, and the second method consists of Algorithm 1. In the first method the correlation coefficient (CC) was used to select the best variables and respective time-lags. The CC between the real output and each input variable in the training dataset of the variables of $X^{(k)}$ was evaluated. The criterion used to remove the least significant variables was the following: input variables with an absolute value of CC below the threshold of 0.2 were rejected in the selection procedure. Table 2 shows the selected variables and respective time lags for correlation coefficient procedure (marked with '$*$'). Ten input variables were selected, where variables u_4, u_6, and u_7 were not selected for any delay. The final prediction model was trained using the least squares SVM with the selected inputs. Table 3 (line CC+LS-SVM") shows the prediction results of the prediction setting designed with the correlation coefficient based variable selection procedure.

Table 2: Selected variables and respective time lags on the Debutanizer experiment.

Time Lag/Variable	MI ($\circ$), CC ($*$)						
	u_1	u_2	u_3	u_4	u_5	u_6	u_7
(k)	$\circ$	$*$	$\circ *$	$\circ$	$\circ *$		
$(k-4)$	$\circ$	$*$	$*$		$*$		
$(k-8)$	$\circ *$	$*$	$*$	$\circ$	$*$		

Table 3: Performance results on the butane concentration estimation in the Debutanizer experiment.

Method+Model	Performance Results				
	$\lvert S\rvert$	RMSE	NMSE	MAE	Correlation
MI + LS-SVM	7	0.0461	0.0750	0.3627	0.9626
CC + LS-SVM	10	0.0625	0.1516	0.3129	0.9223
No Selection	7	0.0619	0.1499	0.3846	0.9234

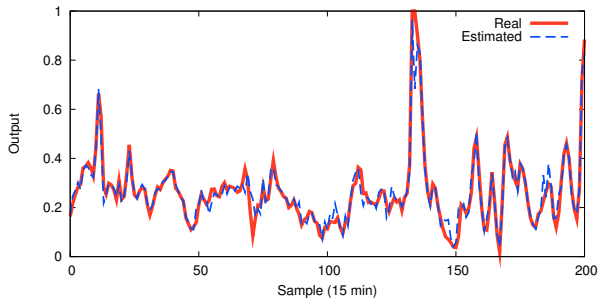

Figure 1: Comparison between the real and predicted C_4 concentration in the Debutanizer experiment, using the variables and delays selection procedure of Algorithm 1.

When the mutual information based forward selection procedure proposed in Algorithm 1 is employed, seven variables are selected. The selected best input variables are indicated in Table 2 (marked with '$\circ$'). Variables u_2, u_6 and u_7 were not selected as inputs, for any delay. Table 3 (line "MI+LS-SVM") shows the prediction results of the prediction setting designed with the variable selection procedure of Algorithm 1. The final prediction, for the test set, is depicted in Figure 1. Table 2, in the line "No Selection", presents the results of prediction performance obtained when variables and delays selection is not performed.

In this case study, it is possible to conclude that the mutual information criterion can adequately handle nonlinear interactions between input variables in the variable and delay selection problem. The performance results have shown that the application of Algorithm 1 that is based on the mutual information criterion leads to a higher prediction performance when compared with the application of the method based on the traditional correlation coefficient performance index in the variable selection technique. The proposed MI-based selection procedure selects a lower number of (variable, delay) pairs while attaining a

Table 4: Description of the variables of the simulated WWTP.

Variables Description	
Dissolved Oxygen in effluent	u_1
Dissolved Oxygen in reactor 1	u_2
Settler sludge height	u_3
Q_{in} (Influent flow rate)	u_4
Q_e (Effluent flow rate)	u_5
Q_{pu} (Primary settler underflow flow rate)	u_6
Influent temperature	u_7
Effluent temperature	u_8
Instantaneous sludge wastage rate	u_9
Instantaneous methane production	u_{10}
Instantaneous total gas flow normalized to P-atm	u_{11}
Instantaneous pumping energy	u_{12}
COD in effluent	y

better performance.

6.2 Simulated WWTP

This section presents experimental results of a case study concerning the estimation of the total chemical oxygen demand (COD) at the effluent of a wastewater treatment plant (WWTP). The WWTP study was conducted using the Benchmark Simulation Model No. 2 (BSM2) [5]. BSM2 is a platform-independent WWTP simulation environment defining a plant layout, a process model, influent data, test procedures and evaluation criteria. The benchmark is evaluated for two years with acquisition data for the variables being available with a 2 hour sampling interval. There are 12 input variables in the dataset which are candidates for the variables and time-lag selection problem. We define the set of input variables as: $U^{(k)} = \{u_1(k), \ldots, u_{12}(k)\}$ and the output, COD, as y. The description of each variable, in $U^{(k)}$ is given in Table 4.

For each input variable, five possible input delays were chosen $(0, 1, 2, 3, 4)$. Thus, the ensembled candidates input set is given by $X^{(k)} = \{u_1(k), u_1(k-1), u_1(k-2), u_1(k-3), u_1(k-4), u_2(k), \ldots, u_{12}(k-4)\}$, the original number of samples in the dataset is 4368, but due to the consideration of delays this number is reduced for 4364. To apply the input variable and time-lag selection algorithm, the whole data set of was divided into a training data set of 3819 points, and a test data set of 545 points.

Similarly to the debutanizer experiment (Sec. 6.1), two algorithms are compared for variable and delay selection: the first is based on the correlation coefficient, and the second method consists of Algorithm 1. In the first method, the criterion used to remove the least significant variables was the following: input variables with an absolute value of CC below the threshold of 0.4 were rejected in the selection procedure. Table 5 shows the selected variables and respective time lags for correlation coefficient procedure (marked with '$*$'). Eleven input (variable, delay) pairs, of six input variables, were selected, where variables u_1, u_2, u_7, u_8 were not selected for any delay. The final model was trained using the least squares SVM with the selected inputs. Table 6 (line "CC+LS-SVM") shows the prediction results of the prediction setting designed with the correlation coefficient based variable selection procedure.

Table 6: Performance results on the COD estimation in the WWTP experiment.

Method+Model	Performance Results				
	$\|S\|$	RMSE	NMSE	MAE	Correlation
MI + LS-SVM	6	2.11	0.236	9.27	0.87
CC + LS-SVM	11	2.80	0.413	13.23	0.77
No Selection	12	2.19	0.254	17.03	0.86

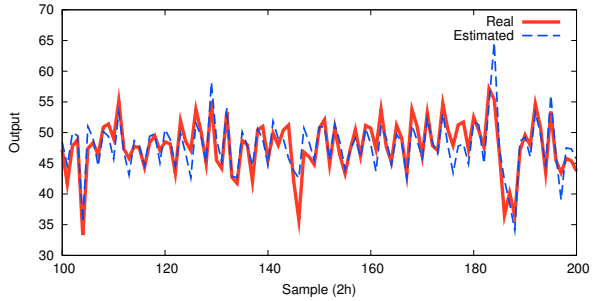

Figure 2: Comparison between the real and predicted COD in the WWTP experiment.

When the mutual information based forward selection procedure proposed in Algorithm 1 is employed, five variables and six (variable, delay) pairs were selected. The selected best input variables are indicated in Table 5 (marked with '$\circ$'). Variables u_1, u_2, u_4, u_5, u_6, u_7 and u_8 were not selected as inputs, for any delay. Table 3 (line "MI+LS-SVM") shows the prediction results of the prediction setting designed with the variable selection procedure of Algorithm 1. The final prediction, for the test set, is depicted in Figure 2. Table 6, in the line "No Selection", presents the results of prediction performance obtained when variables and delays selection is not performed.

Again in this second experiment, it is possible to conclude that the mutual information criterion can adequately handle non-linear interactions between input variables in the variable and delay selection problem. The performance results have shown that the application of Algorithm 1 that is based on the mutual information criterion leads to a higher prediction performance when compared with the application of the method based on the traditional correlation coefficient performance index in the variable selection technique. The proposed MI-based selection procedure selects a lower number of (variable, delay) pairs while attaining a better performance. This is a positive point for soft sensor settings where it is desirable to have lower requirements in terms of the number of sensors required.

Table 5: Selected variables and respective time lags in the WWTP experiment.

Time Lag/Variable	MI ($\circ$), CC ($*$)											
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}
(k)			$*$	$*$	$*$	$*$			$\circ *$	$\circ$	$\circ$	$*$
$(k-1)$			$\circ$	$*$	$*$	$*$			$*$			$*$
$(k-2)$									$\circ$			
$(k-3)$												
$(k-4)$												$\circ$

7 Discussion

Correlation coefficient is a classical criterion used for linear variable selection. However, the main shortcoming of individual variable selection is that it does not take into consideration the interaction among variables, and can not remove redundant variables, because redundant variables have similar values. Moreover, it is still necessary to provide a threshold to remove less relevant variables.

It is known that when irrelevant variables are deleted from the model, prediction results can be improved. An advantage of the use of the mutual information is that it can be used as a tool to deal with the problem of redundancy: it can be used as a basis for selecting relevant variables for complex and highly nonlinear processes. Often the use of mutual information permits to attain a better choice of variables and delays on nonlinear processes, thus ultimately permitting to get a better understanding of the process. The method proposed in this paper can replace the use of the correlation coefficient in the initial data analysis steps in soft sensor development.

8 Conclusions

This work presented a method to perform variable and time-lag selection for industrial applications. The method was successfully applied in two datasets. The case studies indicate that the proposed method can give better results when compared with correlation coefficient variable selection, and when compared to the situation when variables and delays selection is not performed. Moreover, unlike the case of the procedure based on the correlation coefficient, the proposed method has the advantage that it is not necessary to define a prior threshold value for the stopping condition. The results indicate that the proposed method can replace with advantage the method based on the correlation coefficient.

Acknowledgment

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 has been supported by Fundação para a Ciência e a Tecnologia (FCT) under grant SFRH/BD/63454/2009.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 1991.

[2] M. R. Delgado, E. Y. Nagai, and L. V. R. de Arruda. A neuro-coevolutionary genetic fuzzy system to design soft sensors. *Soft Computing*, 13(5):481–495, December 2008.

[3] L. Fortuna, S. Graziani, and M. G. Xibilia. *Soft Sensors for Monitoring and Control of Industrial Processes*. Springer, 2007.

[4] I. Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

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

[6] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence Archive*, 97(1-2):273–324, December 1997.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, June 2004.

[8] Ls-svmlab, 2011. [Online]. Available: http://www.esat.kuleuven.be/sista/lssvmlab/ .

[9] O. Ludwig, U. Nunes, R. Araújo, L. Schnitman, and H. A. Lepikson. Applications of information theory, genetic algorithms, and neural models to predict oil flow. *Communications in Nonlinear Science and Numerical Simulation*, 14(7):2870–2885, 2009.

[10] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005.

[11] M. Shakil, M. Elshafei, M. Habib, and F. Maleki. Soft sensor for nox and o2 using dynamic neural networks. *Computers & Electrical Engineering*, 35(4):578–586, 2009.

[12] F. Souza, P. Santos, and R. Araújo. Variable and delay selection using neural networks and mutual information for data-driven soft sensors. In *Proc. 2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA 2010)*, pages 1–8, September 2010.

[13] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle. *Least Squares Support Vector Machiness*. World Scientific, 2002.

[14] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. *JMLR: Workshop and Conference Proceedings*, 4:5–20, 2008.

[15] K. Warne, G. Prasad, S. Rezvani, and L. Maguire. 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, 2004.