Ciencia e Ingeniería Neogranadina
ISSN: 0124-8170
ISSN: 1909-7735
Universidad Militar Nueva Granada

Cote-Ballesteros, Jorge E.; Grisales Palacios,
Victor Hugo; Rodriguez-Castellanos, Jhon Edisson

A Hybrid Approach Variable Selection Algorithm Based on Mutual Information for Data-Driven Industrial Soft-Sensor Applications*

Ciencia e Ingeniería Neogranadina, vol. 32, no. 1, 2022, January-June, pp. 59-70
Universidad Militar Nueva Granada

DOI: https://doi.org/10.14482/INDES.30.1.303.661

Available in: https://www.redalyc.org/articulo.oa?id=91172083005

Ciencia e Ingenieria Neogranadina 2022 Vol. 32(1) enero-junio 2022 ■ ISSN: 0124-8170 · ISSN-e: 1909-7735 ■ pp. 59-70

Editorial Neogranadina

DOI: https://doi.org/10.18359/rcin.5644

A Hybrid Approach Variable Selection Algorithm Based on Mutual Information for Data-Driven Industrial Soft-Sensor Applications*

Jorge E. Cote-Ballesteros[a] ■ Victor Hugo Grisales Palacios[b] ■ Jhon Edisson Rodriguez-Castellanos[c]

Abstract: The development of virtual sensors predicting the desired output requires a careful selection of input variables for model construction. In an industrial environment, datasets contain many instrumentation system measures; however, these variables are often non-relevant or excessive information. This paper proposes a variable selection algorithm based on mutual information examination, redundancy analysis, and variable reduction for soft-sensor modeling. A relevance calculation is performed in the first stage to select important variables using the mutual information criterion. Then, the detection and exclusion of redundant variables are carried out, penalizing undesired variables. Finally, the most relevant variables subset is determined through a wrapper method using Mallowssans' Cp metric to assess the fitting prediction performance. The approach was successfully applied to estimate the ethanol concentration for a distillation column process using an adaptive network-based fuzzy inference system architecture as a non-linear dynamic regression model. A comparative study was performed considering the application of correlation analysis and the method proposed in this study. Simulation results show the effectiveness of the proposed approach in the variable selection providing a reduction in search of suitable models that achieve faster results for developing soft sensors oriented to industrial applications.

Keywords: Data-driven, distillation column feature selection, industrial processes, mutual information, soft sensor.

* Research paper.

a MS Industrial Automation. Universidad ECCI, Bogotá, Colombia. E-mail: jcoteb@ecci.edu.co ORCID: http://orcid.org/0000-0002-9830-7726

b PhD. Automated Systems. Universidad Nacional de Colombia, Bogotá, Colombia. E-mail: vhgrisalesp@ unal.edu.co ORCID: http://orcid.org/0000-0002-8625-7821

c MS Industrial Automation. Universidad ECCI, Bogotá, Colombia. E-mail: jhrodriguezc@ecci.edu.co ORCID: http://orcid.org/0000-0002-7790-605X

Recibido: 01/03/2021 **Aceptado:** 29/11/2021

Disponible en línea: 03/06/2022

Cómo citar: J. E. Cote-Ballesteros, V. H. Grisales Palacios, and J. E. Rodriguez-Castellanos, "A Hybrid Approach Variable Selection Algorithm Based on Mutual Information for Data-Driven Industrial Soft-Sensor Applications", *Cien.Ing.Neogranadina*, vol. 32, no. 1, pp. 59-71, Jun 2022.

Un algoritmo de selección de variables de enfoque híbrido basado en información mutua para aplicaciones de sensores blandos industriales basados en datos

Resumen: El desarrollo de sensores virtuales que predicen el resultado o producto deseado requiere una cuidadosa selección de variables de entrada para la construcción del modelo. En un entorno industrial, los conjuntos de datos contienen muchas medidas del sistema de instrumentación; sin embargo, estas variables suelen ser información no relevante o excesiva. Este artículo propone un algoritmo de selección de variables basado en el examen de información mutua, el análisis de redundancia y la reducción de variables para el modelado de sensores blandos. En la primera etapa se realiza un cálculo de relevancia para seleccionar variables importantes utilizando el criterio de información mutua. Luego, se realiza la detección y exclusión de variables redundantes, penalizando las variables no deseadas. Finalmente, el subconjunto de variables más relevante se determina a través de un método de envoltura utilizando la métrica Cp de Mallows para evaluar el rendimiento de la predicción de ajuste. El enfoque se aplicó con éxito para estimar la concentración de etanol para un proceso de columna de destilación utilizando una arquitectura de sistema de inferencia difusa basada en red adaptativa como un modelo de regresión dinámica no lineal. Se realizó un estudio comparativo considerando la aplicación del análisis de correlación y el método propuesto en este estudio. Los resultados de la simulación muestran la efectividad del enfoque propuesto en la selección de variables proporcionando una reducción en la búsqueda de modelos adecuados que logren resultados más rápidos para el desarrollo de sensores blandos orientados a aplicaciones industriales.

Términos del índice: basado en datos; selección de características de la columna de destilación; procesos industriales; información mutua; Sensor blando

Introduction

In industrial processes, online quality measurements are critical to ensure suitable process parameters, automatic control of key-variables, and timely decision-making. The online calculation of these variables is, in some cases, difficult or money and time expensive. When hardware sensors are unavailable or unsuitable, data-based inferential estimators called soft sensors or virtual sensors developed in recent years can be used. The soft sensor indirectly estimates primary variables through inference from process observations. In the industrial sector, the estimation of the desired process variable is usually based on secondary easy-to-measure variables, such as temperature, flow rate, level, among others. Some works show a complete methodology for soft-sensor design, for instance in [1], [2].

Thus, each methodology reported considers the problem of datasets contain many variables, which are challenging to handle since they often include redundant and non-relevant information. Therefore, prediction performance is reduced due to overfitting and data dimensionality problems. This situation has been studied, and several works have been reported to reduce the number of variables by removing irrelevant information variables. Three main approaches are used for variable selection in machine learning. Filtering techniques use statistical metrics to score variables regarding the target variable [3]. Wrapper methods use a learning model to select relevant variables, making them more accurate to fit the particular model through an error metric for selection. In contrast, they are more computationally expensive than filtering methods [4]. A third approach is a hybrid, in which filtering and wrapper elements are combined to improve the quality of selection and computational cost reduction [5].

Several successful soft sensors have been reported dealing with feature dimensionality problems through filtering approaches. For instance, [6] implemented a soft sensor based on artificial neural networks (ANN), [7] designed a support vector machines (SVM) to predict the melt index for a polymer in, and [8] employed fuzzy systems to obtain methanol concentration from a simulated distillation column. These cases provide a filtering approach based on correlation analysis through Pearson's correlation coefficient. Although some other metrics such as information theory, non-linear correlation or even expert knowledge have been reported in soft-sensor design, Pearson's correlation analysis is the most used. Likewise, several soft sensors have been reported using wrapper methods for variable selection. The wrapper approach uses the performance of the learning technique itself to provide the most representative subset from several options. For example, [9] implements a soft sensor for monitoring gases emission, for which variable selection was carried out through ANN training using different subsets of variables. In [10], a soft sensor based on ANN is implemented to estimate kerosene's endpoint. The work reports a wrapper selection of secondary variables using Mallow's Cp as the model performance metric. Also, [10] uses several ANN architectures and mean square error (MSE) to select the most relevant variables for predicting the target variable.

Recently, hybrid methods have drawn more attention from soft-sensor designers to achieve a suitable prediction performance and reduce computational costs employing combined filtering and wrapper methods. Thus, filtering scores such as correlation coefficient, Spearman coefficient, among others, show linear dependence between two variables. However, industrial processes present strongly non-linear relationships between their variables. Therefore, more powerful techniques using information theory-based metrics such as mutual information (MI) and gain information (GI) have been used. These are entropy-based metrics; thus, the dependency is measured through the quantity of information computing. In [11]–[13], the authors show soft-sensor selecting variables through MI as filtering metric and regression techniques such as principal component analysis (PCA) or partial least squares (PLS), to obtain a suitable subset of variables. Although many hybrid algorithms have been proposed to detect the relevance of variables of a dataset, scarce attention has been paid to non-linear dependencies between variables

of industrial datasets considering the redundancy and model performance measurement.

In [14], [15] the authors present a complete review of feature selection techniques for industrial processes soft sensors. They have considered application cases of industrial soft sensors where filtering, wrapper and hybrid approaches were applied. Although many of these approaches have been used on successful soft-sensor applications, the more recent works explore hybrid method due to its balance between accuracy and computational cost. Also, MI has been drawing attention from researchers, since it deals with nonlinear datasets and low computational cost. Lastly, soft-sensor researchers have focused on performance metrics regarding machine learning technique; however, the prediction performance may be strongly affected by redundant variables in the dataset. Therefore, redundancy should be measured and involved in the feature selection algorithm.

This work proposes a hybrid selection variable algorithm based on the MI coefficient, followed by redundancy analysis and reduction, for industrial processes soft sensors. The approach's effectiveness lies in the variable selection, decreasing the search for suitable models that achieve consistent and accurate results for developing soft sensors oriented to industrial applications. The proposed method starts by computing the relevance for each process variable using the MI coefficient. The variables are then scored by a redundancy coefficient, and finally, the subset of the suitable variables is reduced using Mallow's Cp coefficient as the performance metric. Moreover, a study case demonstrates the algorithm's use and results applied to a distillation column process for water-ethanol mix separation. The soft sensor has been designed using an adaptive neuro-fuzzy inference system (ANFIS) to estimate ethanol concentration at the top of the column. A comparative study was performed to show the results of applying correlation analysis and the proposed method in developing a soft sensor for the distillation column dataset.

This paper is organized as follows: The second section introduces the mutual information coefficient and the model performance metric's mathematical basis and presents the proposed variable selection algorithm. The third section describes the study case's simulation results for applying a soft sensor based on ANFIS. Finally, the fourth section gives concluding remarks.

Materials and methods

Mutual information coefficient

Industrial processes often require finding the best subsets of easy-to-measure variables to estimate the hard-to-measure variable regarding the available dataset. Several soft sensors have been proposed, including feature selection techniques based on correlation and collinearity, for instance [16]. These methods were designed for linear behavior datasets; however, industrial systems are non-linear processes. Hence, correlation metrics are not a suitable choice for feature selection. Recently, information theory metrics have been used to solve non-linear feature selection problems such as entropy (E), MI, and GI, among others.

Mutual information is a non-linear dependency metric between two variables of a system. It can be calculated through information entropy as follows:

$$H(x) = -\int f_x(x)\log\big(f_x(x)\big)\,dx \tag{1}$$

$$H(y|x) = \int\int f_{x,y}(x,y)\log\left(\frac{f_x(x)}{f_{x,y}(x,y)}\right)dx\,dy \tag{2}$$

Entropy is the measure of uncertainty of a random variable [17]. Equ. (1) shows the entropy calculation for a continuous random variable, where f_x is the probability density function for a random variable x. Equ. (2) shows the conditional entropy calculation for two continuous random variables, where $f_{x,y}$ is the conditional probability density function for variable y given variable x. Although Equs. (1) and (2) allow entropy calculation for continuous variables, these density probability functions are hard to obtain.

Because of this, assuming x and y as discrete random variables, it is necessary to obtain the entropy as follows:

$$H(x) = -\sum_i p(x_i)\log_2\left(p(x_i)\right) \quad (3)$$

$$H(y|x) = -\sum_j \sum_i p(y_j, x_i)\log_2 p(y_j|x_i) \quad (4)$$

In this manner, MI can be calculated as

$$MI(x,y) = H(y) + H(x) - H(y|x) \quad (5)$$

$$MI(x,y) = -\sum_{i,j} p(y_j, x_i)\log_2 \frac{p(y_j,x_i)}{p(y_j)p(x_i)} \quad (6)$$

where i and j represent the input and output variables to analyze respectively.

Greater MI means greater dependency between x and y. MI is more relevant for describing the relation between variables, valid for both linear and non-linear cases.

Measure for predicting performance

Filtering methods on industrial datasets may not be accurate by themselves, since variables are not being considered as part of a whole model. Wrapper selection methods deal with input variable selection by model performance measurement through several metrics such as MSE, RMSE, R2, Mallow's Cp, among others. The soft-sensor model's predictive performance is frequently measured through error metrics, such as MSE or root mean squared error (RMSE), among others, expressed by

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2 \quad (7)$$

where y_j are the observations, $\hat{y}_j$ predicted values of a variable, and n the number of observations available for analysis. However, the minimal error model is not necessarily optimal since high dimensional models might result in biased error metrics by overfitting [2] . Therefore, to deal with this problem, it is common to use a measurement that penalizes overfitting. For example, Mallow's Cp [18] allows determining the optimal tradeoff between model size and model performance by penalizing those overfitted models. Mallow's Cp can be calculated as

$$C_p = \frac{\sum_{j=1}^{n}\left(y_j - \widehat{y_j}(k)\right)^2}{\sigma_d^2} - n + 2p \quad (8)$$

where y_j (k) are outputs obtained by using a subset of p variables, y_j are the predicted values, n is the number of samples and σ_d^2 are the residuals for the full trained dataset. Thus, Cp-value measures the relative bias and variance of a model with p variables. Therefore, the unbiased model's value will be p so that the optimal model will have the Mallow's Cp number closest to p [1].

Proposed variables selection algorithm

This subsection presents a new variable selection algorithm based on the MI coefficient and a wrapper technique to obtain a suitable subset of variables for soft sensors with industrial orientation. The MI-based variable selection method may be considered as filtering, where the MI coefficient is used to score relevance for each process variable [19], [20]. Industrial datasets contain a large number of variables, thus some of these measurements might provide redundant information and prediction performance degradation. Therefore, redundancy between two random variables may be calculated as [21]

$$R\left(x_i, x_j\right) = \frac{MI\left(x_i,x_j\right)}{H(x_i)+H(x_j)} \quad (9)$$

where R takes values between 0 and maximum valor. Therefore, normalized R can be written as [21]

$$\hat{R}\left(x_i, x_j\right) = \frac{MI\left(x_i,x_j\right)}{min\{H(x_i),H(x_j)\}} \quad (10)$$

The stages of the proposed algorithm are as follows. First, the selection of essential variables is performed using the mutual information criterion. Then detection and exclusion of redundant variables are carried out, penalizing relevant variables selected previously. Finally, the subset of the most suitable variables is determined using the wrapper method to assess the prediction performance with

Mallow's Cp metric as the selection criterion. To calculate MI, Equ. (6) is employed.

The proposed algorithm's detailed structure is presented below; at the final iteration, S is the selected set of variables for developing the soft sensor.

Algorithm 1 Variable Selection Algorithm

1: procedure

//filtering stage

2: Given F={f_i} the total variable set and S an empty set; i=1,2,…,k

3: Compute MI for each feature regarding target values; MI(x_i,y_j)

4: Adding the first variable to S; s_1=max{MI(x_i,y_j)} have the answer if r is 0

5: while Mallow's Cp decrease:

//Wrapper stage and redundancy computing

6: Compute redundancy for s_1 regarding all remaining s∈S

7: Selection of next feature/variable by max $\{MI(x_i, y_i) - \frac{1}{k}\sum_{i=1}^{k} R_k\}$

8: Training and validation by using a machine learning technique with the obtained subset

9: Compute Mallow's Cp performance prediction

10: end while

11: end procedure

High redundancy variables are strongly penalized by using average redundancy and possibly removing them, otherwise, the variables are selected for the S subset.

The Hill climbing method is used as the wrapper and consists of an iterative trial and error technique, starting with an empty variables subset and progressively incrementing one variable by one until the best performance subset is complete. Lastly, Cp-value is chosen as the prediction performance index. The Mallow's Cp will be degraded if high dimensionality affects soft-sensor performance prediction.

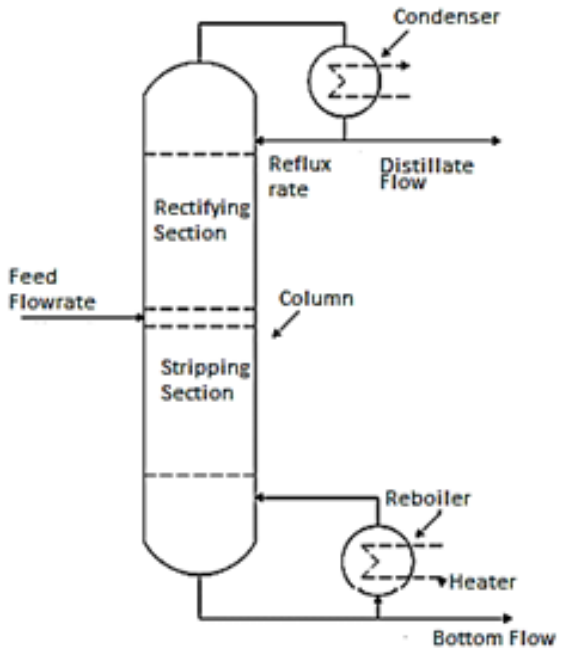
Result and discussions

The proposed study case is a distillation column process that is broadly used for petrochemical and food factory industries. Much work has been reported with successful applications of distillate product soft sensors [10], [11], [22].

Simulation dataset

A distillation process consists of separating two or more compounds by applying energy to lead vapor toward the column's top. The remaining liquid is transported to the bottom of the column. This process is repeated until the separation is completed. Fig. 1 shows a distillation column scheme.

Fig. 1. Distillation Column Scheme.



Source: The authors

This work considers a non-linear simulated mathematical model of a binary distillation with 12 trays to separate the water-ethanol mix. A dataset containing 60 input variables with 4000 observations (data points) was collected at a sample rate of 10 samples per hour with no-shutdown phases. The dataset was contaminated with 10% amplitude noise, and random 5% of total observations are outliers, to represent common environmental conditions of industrial conditions. Ethanol distillate concentration is the dataset output, and the model inputs are shown in Table 1.

Table 1. Labels for distillation column variables

Variables description	Label
Pressure (at each stage)	$u_{\_1}$-$u_{\_14}$
Liquid flowrate (at each stage)	u_{15}-u_{28}
Vapor flowrate	u_{29}
Temperature (at each stage)	u_{30}-u_{43}
Hold on mass (at each stage)	u_{44}-u_{57}
Distillate flowrate	u_{58}
Bottom flowrate	u_{59}
Feed flowrate	u_{60}

Source: The authors

Outliers detection

The presence of outliers in industrial datasets is expected due to environmental issues; therefore, these observations should be detected and replaced, aiming for a suitable training and prediction performance. A Hampel filter is applied to the dataset to avoid high deviated outliers sensitivity [1], [14], [23]. Thus, the Hampel filter uses median absolute deviation (MAD), an outlier-resistant metric, and applies the filter through a moving window with two tuning parameters, threshold, and width. MAD can be implemented with [19]

$$MAD = 1.4286 * median\{|x_i - x^*|\} \quad (11)$$

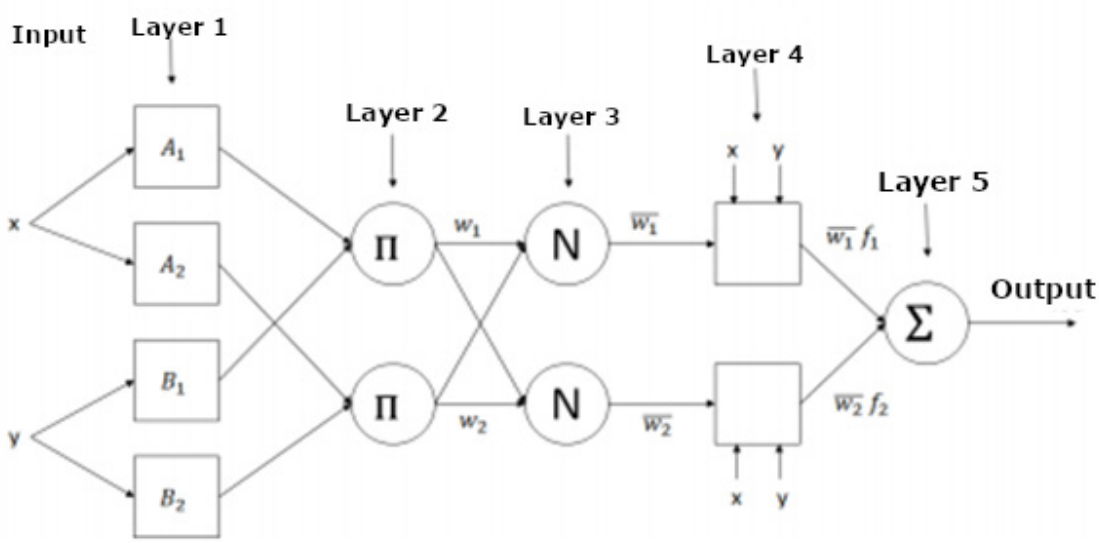
where x_i are the values of the data sequence and x^* is the median.

Training and validation

The soft sensor was developed using a five-layer ANFIS model with gaussian membership functions, N input variables, and Takagi-Sugeno consequent. Fig. 2 shows the considered architecture. The training was carried out with a hybrid approach using a back-propagation algorithm for tuning the membership function parameters and least squares for adjusting Takagi-Sugeno function parameters [24], [25].

A first-order, non-linear autoregressive with exogenous input (NARX) model was considered for ethanol distillate concentration prediction. MSE and RMSE were used as training performance metrics.

Fig. 2. Five-layer ANFIS architecture for the industrial soft sensor.



Source: The authors

The proposed algorithm was applied to the initial 60 variables dataset from Table 1 to obtain the most suitable subset for prediction. First, the ranking stage was performed to determine ethanol concentration's dependence regarding distillation column secondary variables; the results are presented below. Table 2 shows the top ten variables obtained through mutual information coefficient, for which ethanol concentration depends on several flow-rates and trays temperature. A greater MI coefficient means greater dependence regarding ethanol concentration.

Table 2. Top-ten relevant variables regarding ethanol concentration based on mutual information

Variable	Label	MI coefficient
Temperature plate	u_{30}	3.42
Temperature plate	u_{31}	2.47
Temperature plate	u_{33}	2
Temperature plate	u_{35}	1.52
Temperature plate	u_{36}	1.38
Temperature plate	u_{41}	0.98
Liquid flowrate	u_{15}- u_{28}	0.69
Bottom flowrate	u_{59}	0.57
Feed flowrate	u_{60}	0.23
Distillate flowrate	u_{58}	0.17

Source: The authors

Table 3 shows the ten highest Pearson correlation coefficients in magnitude, between the distillation column inputs and the ethanol concentration. It shows that the estimated variable has a high dependence only on the temperatures in the column plates. However, Pearson's coefficient does not consider a non-linear correlation between the variables, while the MI coefficient does.

The results of this first stage of the algorithm, after an extensive graphical study, indicate high dependence between the temperatures in the trays and ethanol concentration, however, these results show redundancy between the variables. Therefore, it is likely that the soft sensor will have a performance degradation if all these temperatures are considered for the ANFIS model. Additionally, the correlation method only considers linear relationships between the variables, therefore it is likely that at the end of the list there are variables that represent the ethanol concentration in a non-linear way. On the other hand, the MI coefficient presents very different results regarding the variables with the highest weight. In this case, the first 10 variables correspond to some temperatures but also some flows, which indicates that in these variables there is information that represents the ethanol concentration, regardless of whether these relationships are linear or non-linear.

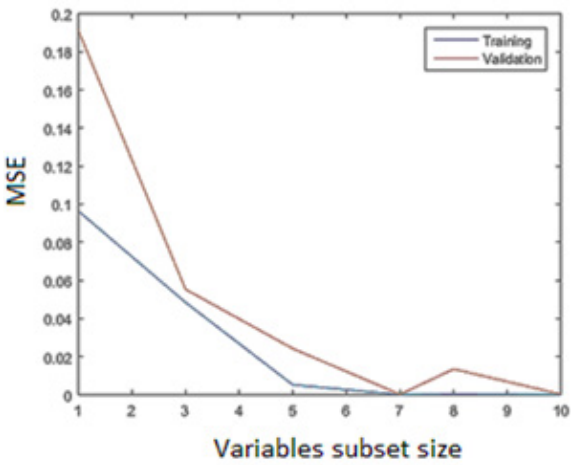
Table 3. Top-ten relevant variables regarding ethanol concentration based on correlation analysis

Variable	Label	Pearson's coefficient
Temperature plate 1	u_{30}	-0.9903
Temperature plate 2	u_{31}	-0.9658
Temperature plate 3	u_{32}	-0.9645
Temperature plate 4	u_{33}	-0.9486
Temperature plate 5	u_{34}	-0.9180
Temperature plate 6	u_{35}	-0.8873
Temperature plate 7	u_{36}	-0.8528
Temperature plate 8	u_{37}	-0.8134
Temperature plate 9	u_{38}	-0.7714
Temperature plate 10	u_{39}	-0.7249

Source: The authors

Next, the algorithm proposed was applied, aiming for the best prediction subset to estimate ethanol concentration. Fig. 3 presents the behavior of the algorithm with several groups of variables. Training and validation processes were performed for each variable's subset to assess the performance prediction in these cases. Fig. 3 shows a decrease of MSE with subset size increment, although soft-sensor performance is degraded with subset size greater than seven variables.

Fig. 3. Training and validation MSE for several subsets size.



Source: The authors

Table 4 shows several candidate groups of variables and their validation metrics. From Fig. 3 and Table 4 it is possible to conclude that subset 5 achieves the best balance between subset size and prediction accuracy.

Table 4. Hill climbing metrics results for MI coefficient

Subset	Labels	Subset Size	MSE Train	MSE Validation	Mallow's Cp
1	u_{30}	1	0.0964	0.1911	$5.019x10^6$
2	u_{30},u_{31},u_{58}	3	0.0484	0.0553	$1.446x10^6$
3	$u_{30},u_{31},u_{35},$ u_{58},u_{60}	5	0.0052	0.0242	$6.370x10^5$
4	$u_{30},u_{31},u_{35},$ u_{41}, u_{58}, u_{59},u_{60}	7	0.0001	0.0001856	4933
5	$u_{15},u_{30},u_{31},$ $u_{33},u_{35},$ u_{41},u_{58},u_{60}	8	0.0003	0.0135	$3.637x10^5$
6	$u_{15},u_{30},u_{31},$ $u_{33},u_{35},$ $u_{36},u_{41},u_{58},$ u_{59},u_{60}	10	0.0001	0.0002491	6609

Source: The authors

From Table 5, an 8-variable subset is selected from the algorithm; trays 1, 2, 3, 5 and 18 temperatures are highly relevant for prediction. Likewise, distillate and feed flowrate affect ethanol concentration. Expert knowledge confirms the results obtained from the proposed algorithm.

Table 5 presents variable selection considering correlation analysis. Results show subset 6 as the most suitable regarding Mallow's Cp value. After applying the correlation analysis and wrapper methodology, subset 6 presents a minimal validation metric. However, this MSE is higher than the proposed method because correlation only considers the temperature of trays, and probably information redundancy exists. In this case, at least six variables are necessary to obtain an acceptable performance of the proposed soft sensor

Table 5. Hill climbing metrics results for correlation analysis

Subset	Labels	Subset Size	MSE Train	MSE Validation	Mallow's Cp
1	u_{30}	1	0.0834	0.8312	$8.021x10^6$
2	u_{30},u_{31},u_{32}	3	0.0571	0.0761	$3.381x10^6$
3	$u_{30},u_{31},u_{32},$ u_{33},u_{34}	5	0.0135	0.0517	$7.631x10^5$
4	$u_{30},u_{31},u_{32},$ $u_{33},u_{34},$ u_{35},u_{36}	7	0.0021	0.0128	$4.131x10^5$
5	$u_{30},u_{31},u_{32},$ $u_{33},u_{34},u_{35},$ u_{36},u_{37}	8	0.0015	0.00461	$1.637x10^4$
6	$u_{30},u_{31},u_{32},$ $u_{33},u_{34},u_{35},$ $u_{36},u_{37},$ u_{38},u_{39}	10	0.0004	0.00137	$0.983x10^4$

Source: The authors

After variable selection and NARX model architecture were applied, an ANFIS-based soft sensor was obtained. Thus, it predicts ethanol concentration at the output of a distillation column for water-ethanol mix separation. Several operation points at steady state are considered during 120 hours, without shutdown phases. The data were collected at six minutes time-sampling, the noise of 10% amplitude has been added and 5% of total samples were contaminated with outliers. These are typical conditions of industrial instrumentation systems.

Fig. 4.a. shows the ANFIS well-trained system's validation results with selected subset by the MI-based proposed method. The vertical axis represents ethanol concentration at the top of the distillation column, and the horizontal axis is time. Fig. 4.b. presents residual error from soft-sensor validation. Results show an error near 0 for each time instant.

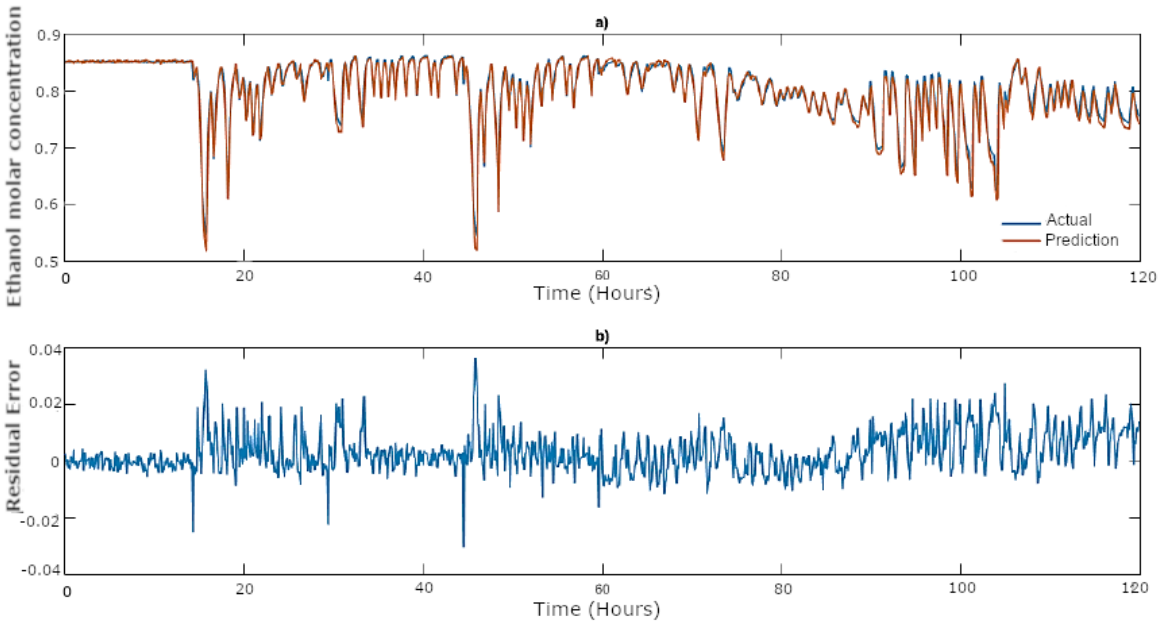
Table 6 shows a comparison between soft sensor based on proposed algorithm and other algorithms reported in literature. Minimal redundancy maximal relevance technique (MRMR) [20] and Mutual information feature selection algorithm (MIFS) [19] have comparable results with the proposed algorithm, however the proposed algorithm has a lower computational cost.

Table 6. MSE for ANFIS soft sensor based on several feature selection algorithms

Feature Selection Algorithm	ANFIS MSE
Proposed	0.000207
Correlation Analysis	0.004989
LASSO	0.002485
MRMR (Nonlinear)	0.000358
MIFS (Nonlinear)	0.000317

Source: The authors

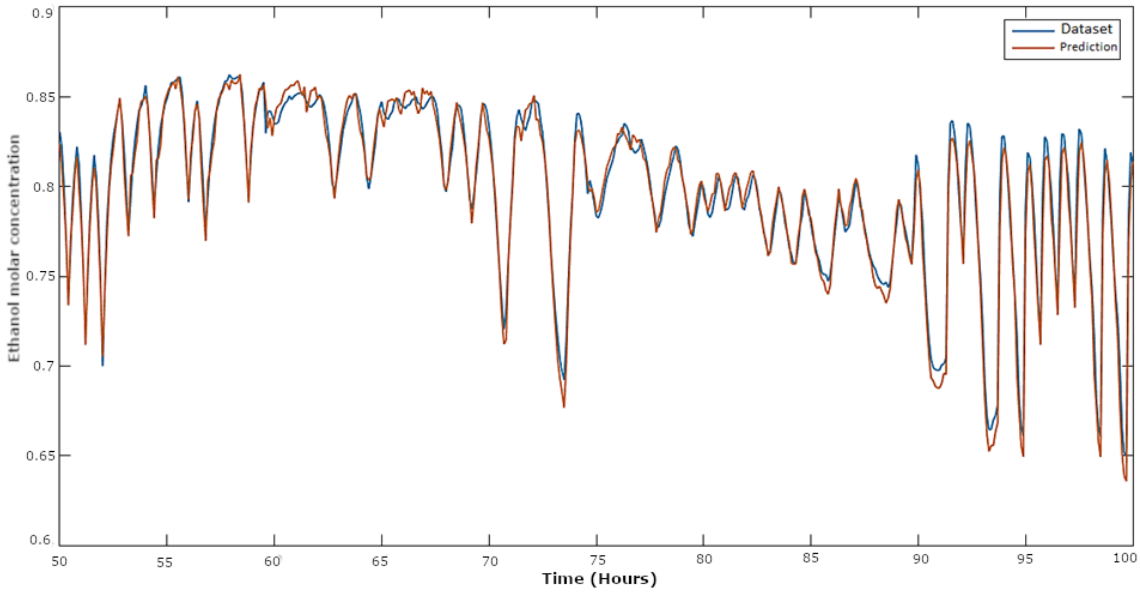
Fig. 4. a) Ethanol molar concentration actual and prediction. b) Residual error.



Source: The authors

Fig. 5 shows 50 hours of operation where the red line represents predicted data, and the blue line is actual data. After the validation stage, MSE was obtained, which has suitable accuracy for this type of measurements.

Fig. 5. Detailed 50-hour operation of the distillation column and the developed soft sensor.



Source: The authors

Conclusions

This paper has proposed a mutual-information (MI) based algorithm to select soft-sensor design variables in an industrial context. This algorithm's main advantage over previously reported methods is the ranking of variables by relevance and redundancy before applying the wrapper method. Thus, the classification of variables reduced the search in the subsets space and provided relevant results faster.

The proposed algorithm was applied to a distillation column to separate a water-ethanol mix for a data-driven soft-sensor application based on ANFIS. A comparative analysis was carried out to explore the best performance between correlation analysis and the proposed MI-based method, which improves prediction accuracy, showing suitable performance for industrial environments.

The study case concludes that ethanol concentration depends on eight variables such as 4 tray temperatures, liquid flowrate, distillation flowrate, among others. These results were compared to expert knowledge; most of the control systems act over these variables, which shows that it was a suitable variable selection result.

References

[1] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, "A systematic approach for soft-sensor development," Comput. Chem. Eng., vol. 31, no. 5-6, pp. 419-425, 2007. DOI: https://doi.org/10.1016/j.compchemeng.2006.05.030

[2] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven Soft-sensors in the process industry," Computers and Chemical Engineering, vol. 33, no. 4. pp. 795-814, 2009. DOI: https://doi.org/10.1016/j.compchemeng.2008.12.012

[3] I. Guyon, A. Elisseeff, and A. M. De, "An Introduction to Variable and Feature Selection," J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[4] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1-2, pp. 273-324, 1997. DOI: https://doi.org/10.1016/S0004-3702(97)00043-X

[5] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," 2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014, no. 1997, 2015. DOI: https://doi.org/10.1109/ICCIC.2014.7238499

[6] L. Fortuna, S. Graziani, and M. G. Xibilia, "Soft-sensors for product quality monitoring in debutanizer distilla-

tion columns," Control Eng. Pract., vol. 13, no. 4, pp. 499-508, 2005. DOI: https://doi.org/10.1016/j.conengprac.2004.04.013

[7] S. B. Chitralekha and S. L. Shah, "Application of support vector regression for developing soft-sensors for nonlinear processes," Can. J. Chem. Eng., vol. 88, no. 5, pp. 696-709, 2010. DOI: https://doi.org/10.1002/cjce.20363

[8] E. Y. Nagai, L. Valeria, and R. De Arruda, "Soft-sensor based on Fuzzy Model indentification."

[9] M. Liukkonen, E. Hälikkä, T. Hiltunen, and Y. Hiltunen, "Dynamic soft-sensors for NOx emissions in a circulating fluidized bed boiler," Appl. Energy, vol. 97, no. x, pp. 483-490, 2012. DOI: https://doi.org/10.1016/j.apenergy.2012.01.074

[10] A. Rogina, I. Šiško, I. Mohler, Z. ˇUjević, and N. Bolf, "Soft-sensor for continuous product quality estimation (in crude distillation unit)," Chem. Eng. Res. Des., vol. 89, no. January, pp. 2070-2077, 2011. DOI: https://doi.org/10.1016/j.cherd.2011.01.003

[11] X. Yuan, H. Yang, and N. S. Wang, "A method of variables selection for soft-sensor based on distributed mutual information," vol. 7, no. 3, pp. 1164-1169, 2015.

[12] F. Souza, R. Araújo, S. Soares, and J. Mendes, "VARIABLE SELECTION BASED ON MUTUAL INFORMATION FOR SOFT-SENSORS APPLICATIONS," in Proceedings of the 9th Portuguese Conference on Automatic Control (Controlo 2010), 2009.

[13] Q. Li, X. Du, W. Liu, and W. Ba, "Soft-sensor modelling based on mutual information variable selection and partial least squares," Proc. - 2017 Chinese Autom. Congr. CAC 2017, vol. 2017-Janua, pp. 3649-3654, 2017. DOI: https://doi.org/10.1109/CAC.2017.8243414

[14] F. Curreri, S. Graziani, and M. G. Xibilia, "Input selection methods for data-driven Soft-sensors design: Application to an industrial process," Inf. Sci. (Ny)., 2020. DOI: https://doi.org/10.1016/j.ins.2020.05.028

[15] F. Curreri, G. Fiumara, and M. G. Xibilia, "Input selection methods for soft-sensor design: A survey," Futur. Internet, vol. 12, no. 6, pp. 1-24, 2020. DOI: https://doi.org/10.3390/fi12060097

[16] V. H. Alves Ribeiro and G. Reynoso-Meza, "Feature selection and regularization of interpretable soft-sensors using evolutionary multi-objective optimization design procedures," Chemom. Intell. Lab. Syst., vol. 212, no. February, p. 104278, 2021. DOI: https://doi.org/10.1016/j.chemolab.2021.104278

[17] T. M. Cover and J. A. Thomas, Elements of Information Theory, Second Ed. Wiley Jhon & sons, 2006.

[18] C. L. Mallows, "Some Comments on Cp," Technometrics, vol. 15, no. November, pp. 87-94, 1973. DOI: https://doi.org/10.2307/1267380

[19] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural-Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, 1994. DOI: https://doi.org/10.1109/72.298224

[20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226-1238, 2005. DOI: https://doi.org/10.1109/TPAMI.2005.159

[21] L. Gao and W. Wu, "Relevance assignation feature selection method based on mutual information for machine learning," Knowledge-Based Syst., vol. 209, p. 106439, Dec. 2020. DOI: https://doi.org/10.1016/j.knosys.2020.106439

[22] M. Mittal, S. C. Satapathy, V. Pal, B. Agarwal, L. M. Goyal, and P. Parwekar, "Prediction of coefficient of consolidation in soil using machine learning techniques," Microprocess. Microsyst., vol. 82, p. 103830, Apr. 2021. DOI: https://doi.org/10.1016/j.micpro.2021.103830

[23] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data," Ecol. Inform., vol. 61, no. January, p. 101224, 2021. DOI: https://doi.org/10.1016/j.ecoinf.2021.101224

[24] J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," IEEE Trans. Syst. Man Cybern., vol. 23, no. 3, pp. 665-685, 1993. DOI: https://doi.org/10.1109/21.256541

[25] J.-S. R. Jang, "Input selection for ANFIS learning," Proc. IEEE 5th Int. Fuzzy Syst., vol. 2, pp. 1493-1499, 1996.