Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 52-1 (2019) 673–678

# Variable Selection for Fault Detection and Identification based on Mutual Information of Alarm Series ⋆

**Matthieu Lucke** *,** **Xueyu Mei** * **Anna Stief** ***
**Moncef Chioua** * **Nina F. Thornhill** **

* *ABB Corporate Research Germany, Wallstadter Strasse 59, 68526 Ladenburg, Germany (e-mail: matthieu.lucke@de.abb.com)*
** *Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK*
*** *ABB Corporate Research Center, ul. Starowislna 13a, 31-038, Krakow, Poland*

**Abstract:** Reducing the dimensionality of a fault detection and identification problem is often a necessity, and variable selection is a practical way to do it. Methods based on mutual information have been successful in that regard, but their applicability to industrial processes is limited by characteristics of the process variables such as their variability across fault occurrences. The paper introduces a new estimation strategy of mutual information criteria using alarm series to improve the robustness of the variable selection. The minimal-redundancy-maximal-relevance criterion on alarm series is suggested as new reference criterion, and the results are validated on a multiphase flow facility.



*Keywords:* Fault detection and diagnosis, Variable selection, Mutual information.

## 1. INTRODUCTION

Monitoring of industrial processes involves a large number of measured process variables often described as data rich but information poor (Ming and Zhao, 2017). Variable selection and feature extraction are the two main approaches to reduce the dimensionality of the fault detection and identification problem. Variable selection consists in identifying and selecting informative and discriminative variables. Feature extraction consists in applying a transformation to the original variables to highlight characteristics or reduce the dimension. Historically, variable selection methods have not received the same attention as feature extraction methods in the process fault detection and diagnosis literature (Ming and Zhao, 2017) although Ghosh et al. (2014) demonstrated that both approaches are complementary. Nevertheless, the recent review of Peres and Fogliatto (2018) highlights a growing interest in variable selection in the community.

A popular strategy for variable selection consists in combining filter criteria to preselect sets of variables with an optimization problem, where the set of variables leading to the best performance of the fault detection and identification algorithm is retained. Verron et al. (2008) combine a multivariate extension of the mutual information criterion with discriminant analysis for fault identification. Ardakani et al. (2016) extend the analysis and benchmark multiple combinations of relevance criterion (e.g. maximum-relevance) and redundancy criterion (e.g. minimum-redundancy) with different classifiers on the Tennessee Eastman process. Other criteria aiming at highlighting variables with abnormal variations have been suggested: Zhao and Gao (2017) identify the nonsteady faulty variables that are disturbed significantly using a stability factor, and Tong and Palazoglu (2016) use an index describing the degree of abnormal variation for each variable. Alternative approaches include genetic algorithms (Ghosh et al., 2014) or selection of the variables that correspond to the root causes of the faults (Shu et al., 2016).

The method designed by Verron et al. (2008) has become a benchmark in the literature. However, the analytical formulations of the univariate and multivariate mutual information of Verron et al. (2008) assume that the faults are stationary, that the process variables follow a Gaussian distribution and that the process variables conditioned to the faults follow Gaussian distributions. This is generally not the case in industrial systems. For this reason, data-driven estimation of mutual information is more appropriate.

Another challenge of industrial fault detection and identification is the variability of process variables from one fault occurrence to another as highlighted by Lucke et al. (2018), due for example to noise, external disturbances or different operating points. The number of fault occurrences available for training is limited, so the algorithms must be robust to cope with distorted patterns. This phenomenon also has an impact on variable selection, because variables

⋆ Financial support is gratefully acknowledged from the Marie Skodowska Curie Horizon 2020 EID-ITN project PROcess NeTwork Optimization for efficient and sustainable operation of Europe's process industries taking machinery condition and process performance into account PRONTO, Grant agreement No 675215.


10.1016/j.ifacol.2019.06.140

with variations directly related to the faults must be separated from variables with less relevant variations.

This paper proposes a new estimation strategy for the criteria based on mutual information using a discretized version of the process variables (the alarm series) instead of the original process variables. The discretization cuts the impact of variations within a certain range around the normal operating point (defined based on the noise level) on the variable selection. The faults are then detected and identified using variables with large variations compared to the noise level, and the effect of noise and external disturbances in the classification is reduced. In addition, the estimation of the joint probability distribution functions in the mutual information is easier on discretized quantities which makes the method more scalable, as highlighted by Yu et al. (2015) and Su et al. (2017) for the estimation of transfer entropy for root cause analysis. An extension of the mutual information criterion, the minimal-redundancy-maximal-relevance (mRMR) criterion (Peng et al., 2005), is proposed to take into account redundancy in the variable selection.

Section 2 defines the alarm series, the mutual information criterion and the mRMR criterion as well as the estimation strategies. Section 3 compares the performance of the two variable selection criteria when applied to process variables and when applied to alarm series, on an industrial case study where variables are not Gaussian and contain variability across fault occurrences. Section 4 explains why the variable selection performs better on the alarm series and Section 5 provides concluding remarks.

## 2. VARIABLE SELECTION BASED ON MUTUAL INFORMATION OF ALARM SERIES

### *2.1 Alarm series*

In this paper, an *alarm series* is a discretized version of the process variable based on statistical process control rules. The alarm series are generated using the standard deviation $\sigma$ of the process variables during normal operation, which gives an indication of the noise level in normal operation. An alarm series $\tilde{X}(t)$ is generated from a process variable $X(t)$ according to Eq. 1, where $\tilde{k}$ is a positive integer tuned according to the expected variability in the noise level. In addition, alarm chattering is removed using delay timers.

$$\tilde{X}(t) = \begin{cases} -1, & \text{if} \quad X(t) \leq -\tilde{k}\sigma \\ 0, & \text{if} \quad -\tilde{k}\sigma < X(t) < \tilde{k}\sigma \\ 1, & \text{if} \quad X(t) \geq \tilde{k}\sigma \end{cases} \tag{1}$$

### *2.2 Fault detection and identification*

Fault detection consists in determining whether a fault happened and fault identification consists in identifying the type of fault that occurred. Fault detection and identification is usually done using one (or several) model(s) on the process variables $X_i$. As summarized by Russell et al. (2000), it is common practice to consider the plant profiles at different times $t$ as the inputs of the model. The plant profile $Z(t)$ at sampling time $t$ is defined as the vector $[X_1(t)X_2(t)\ ..\ X_N(t)]$ where the $X_i(t)$ are the values taken by the process variables $X_i$ at time $t$.

One strategy to build the model is supervised learning, where a statistical model is trained considering the values of the plant-profiles $Z(t)$ at different sampling times $t$ and their corresponding class label $C(t)$ which relates to a specific type of fault or to normal operation. In this paper, a single classification model addressing both the fault detection and the fault identification is used. The model is trained on a set of plant profiles covering the trajectory of one occurrence of each type of fault, as well as on a set of plant profiles corresponding to normal operation data.

Variable selection has an impact on the performance of the fault detection and identification model as it consists in determining the best variables $X_i$ to be included in the plant profile $Z$. Variable selection based on mutual information is a popular strategy in the fault detection and diagnosis literature since it provides simple and systematic criteria, independent from the choice of the classification model.

### *2.3 Variable selection based on mutual information*

The mutual information $I(x;y)$ between two random variables $x$ and $y$ is a quantity measuring the mutual dependence of the two variables (Shannon and Weaver, 1949) that can be computed as:

$$I(x;y) = \iint P(x,y)\, log \frac{P(x,y)}{P(x)P(y)}\, dxdy \tag{2}$$

where $P(x,y)$ is the joint probability distribution function of $x$ and $y$, and $P(x)$ and $P(y)$ are the marginal probability distribution functions of $x$ and $y$ respectively.

When at least one of the random variables is continuous, the joint probabilities can be estimated using estimators based on binning, kernel density or nearest neighbours. The nearest neighbours estimator compares well to the two other approaches as it provides a data efficient and adaptive estimator (Kraskov et al., 2011). For this reason, the nearest neighbours estimator is used in the present work to estimate mutual information on process variables. When both random variables are discrete, the estimation can be done as a discrete sum.

In the context of classification, the purpose of variable selection based on mutual information is to find a set $S$ of $m$ variables $X_i$ that have the largest dependency on the class $C$. The max-dependency criterion is defined as:

$$\max\ d(S,C),\ d = I(\{X_i, i = 1, ..., m\}; C) \tag{3}$$

Since the joint probability distribution functions are difficult to estimate in practice for lack of samples, the max-dependency criterion is approximated using simplified criteria such as the mutual information criterion or the mRMR criterion.

### 2.4 Mutual information criterion

The simplest criterion is the univariate mutual information $I(X_i;C)$ between a variable $X_i$ and the class $C$. The higher the value of $I(X_i;C)$ is, the more relevant $X_i$ is considered for the classification. It can be expressed as:

$$I(X_i;C) = \iint P(x_i,c)\, log\frac{P(x_i,c)}{P(x_i)P(c)}\, dx_i dc \qquad (4)$$

where $x_i$ and $c$ represent all the possible values that $X_i$ and $C$ can take. The probability distribution functions are computed using the extension of the nearest neighbours estimator between a continuous and a discrete variable derived by Ross (2014).

The mutual information criterion based on alarm series $I(\tilde{X}_i;C)$ is proposed as an alternative in this paper and can be computed as a discrete sum:

$$I(\tilde{X}_i;C) = \sum_{\tilde{x}_i,c} P(\tilde{x}_i,c)\, log\frac{P(\tilde{x}_i,c)}{P(\tilde{x}_i)P(c)} \qquad (5)$$

### 2.5 Minimal-redundancy-maximal-relevance criterion

The mRMR criterion (Peng et al., 2005) takes into account both the relevance of the variables and their redundancy, although still in a pairwise manner. The max-relevance criterion is an approximation of the dependency criterion in Eq. 3 formulated as:

$$\max\ D(S,C),\ D = \frac{1}{|S|}\sum_{X_i\in S} I(X_i;C) \qquad (6)$$

The min-redundancy criterion is defined as:

$$\min\ R(S),\ R = \frac{1}{|S|^2}\sum_{X_i,X_j\in S} I(X_i;X_j) \qquad (7)$$

Both are combined as the mRMR criterion:

$$\max\ \Phi(D,R),\ \Phi = D - R \qquad (8)$$

As in Section 2.4, $I(X_i;C)$ is computed using the extension of the nearest neighbours estimator between a continuous and a discrete variable. $I(X_i;X_j)$ is computed using the nearest neighbours estimator.

In practice, an incremental search is done using Eq. 9. Assuming we know the set of $m-1$ variables $S_{m-1}$, the $m$th variable is selected from the remaining variables $\Omega_X - S_{m-1}$ as:

$$\max_{X_j\in\Omega_X - S_{m-1}} \left[ I(X_j;C) - \tfrac{1}{m-1}\sum_{X_i\in S_{m-1}} I(X_j;X_i) \right] \qquad (9)$$

The mRMR criterion based on alarm series is proposed as an alternative:

$$\max\ \tilde{D}(\tilde{S},C),\ \tilde{D} = \frac{1}{|\tilde{S}|}\sum_{\tilde{X}_i\in\tilde{S}} I(\tilde{X}_i;C) \qquad (10)$$

$$\min\ \tilde{R}(\tilde{S}),\ \tilde{R} = \frac{1}{|\tilde{S}|^2}\sum_{\tilde{X}_i,\tilde{X}_j\in\tilde{S}} I(\tilde{X}_i;\tilde{X}_j) \qquad (11)$$

$$\max\ \tilde{\Phi}(\tilde{D},\tilde{R}),\ \tilde{\Phi} = \tilde{D} - \tilde{R} \qquad (12)$$

In this case, both $I(\tilde{X}_i;C)$ and $I(\tilde{X}_i;\tilde{X}_j)$ are computed as discrete sums. The incremental search becomes:

$$\max_{\tilde{X}_j\in\Omega_{\tilde{X}} - \tilde{S}_{m-1}} \left[ I(\tilde{X}_j;C) - \tfrac{1}{m-1}\sum_{\tilde{X}_i\in\tilde{S}_{m-1}} I(\tilde{X}_j;\tilde{X}_i) \right] \qquad (13)$$

### 2.6 Design of experiment

The objective of the paper is to demonstrate that the variable ranking obtained using the mutual information criterion (respectively its extension the mRMR criterion) on the alarm series is more appropriate than the ranking obtained using the same criterion on the original process variables.

The variable ranking is done on a training dataset containing the plant profiles of one occurrence of each fault and plant profiles corresponding to normal operation. Four variable rankings are computed:

- Variable ranking using mutual information criterion on process variables.
- Variable ranking using mRMR criterion on process variables.
- Variable ranking using mutual information criterion on alarm series.
- Variable ranking using mRMR criterion on alarm series.

For each variable ranking, each set of variable $S_n$ ($n = 1\ ..\ N$ with $N$ the total number of variables) where $S_n = \{X_{N_1},\ ..\ ,\ X_{N_n}\}$ corresponds to the $n$ best ranked variables is assessed. The accuracy of the classification model based on the plant profiles $Z(t) = [X_{N_1}(t)\ ..\ X_{N_n}(t)]$ is computed on the same training dataset using a five-fold cross-validation strategy.

The classification model chosen for the experiment is a $k$ Nearest Neighbours classifier (Fix and Hodges, 1951). In contrast to discriminant analysis suggested by Verron et al. (2008), kNN does not assume Gaussianity of the variables, which does not hold in this paper.

Finally, the robustness of each variable ranking is evaluated on a test dataset containing another occurrence of each fault and new normal operation data. For each variable ranking obtained on the training dataset, the classification accuracy for each variable set $S_n$ is computed, training the model on the training dataset and testing it on the test dataset.

## 3. INDUSTRIAL CASE STUDY

### 3.1 Description of the industrial case study

The case study is a multiphase flow facility located at the Process System Engineering laboratory of Cranfield University described by Stief et al. (2018). Water and air are mixed at the entrance of the horizontal section and then separated. A process diagram is shown in Figure 1.

The system is operated at two different points and normal operation data is gathered for each operating point. Operating point A corresponds to an air flow rate of 120 $sm^3$

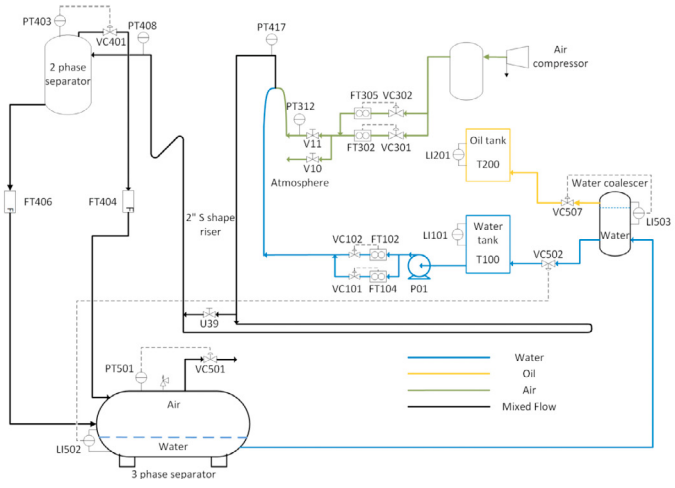


Fig. 1. Process diagram of the multiphase flow facility from Stief et al. (2018) [1]

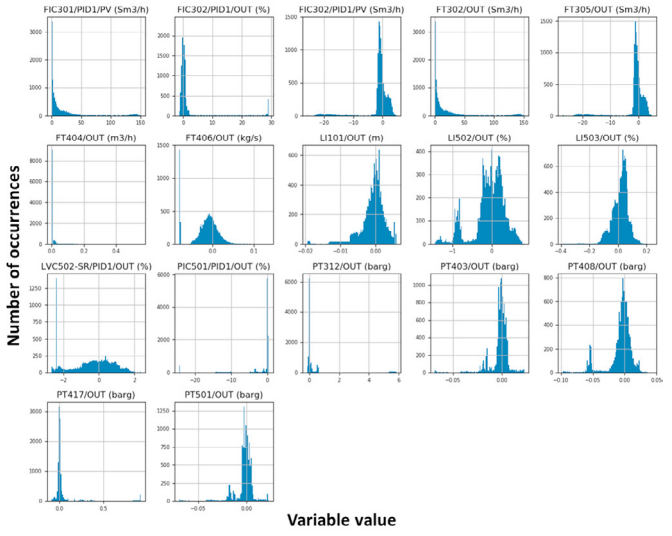


Fig. 2. Distribution plots of the mean-centered variables in operation point A.

$h^{-1}$ and a water flow rate of 0.1 kg $s^{-1}$. Operating point B corresponds to an air flow rate of 150 $sm^3$ $h^{-1}$ and a water flow rate of 0.5 kg $s^{-1}$. Three types of fault are introduced successively at each operating point:

- Air leakage: valve V10 is gradually opened so that the air is partially leaked out to the atmosphere.
- Air blockage: valve V11 is gradually closed to simulate a developing blockage in the input airline.
- Diverted flow: bypass valve U39 is gradually opened so that the mixed flow is partially led straight to the riser and partially led into the horizontal pipeline before joining the riser.

The data are separated into four classes: normal operation data and the three faulty episodes (blockage, leakage and diverted flow).

[1] ©2018 International Federation of Automatic Control. Reproduced with permission from the original publication in IFAC-PapersOnline, 51/18.

Table 1. Process variables.

| Sensor tag | Process variable |
|---|---|
| FT305/OUT | Inlet air flow rate (AFR) 1 |
| FT302/OUT | Inlet air flow rate (AFR) 2 |
| PT312/OUT | Air delivery pressure |
| PT417/OUT | Pressure in the mixing zone |
| PT408/OUT | Pressure at the riser top |
| PT403/OUT | Pressure in the top separator |
| FT404/OUT | Top separator output air flow rate |
| FT406/OUT | Top separator output water flow rate |
| PT501/OUT | Pressure in the 3-phase separator |
| PIC501/PID1/OUT | Air outlet valve 3-phase separator |
| LI502/OUT | Water-oil 3- level phase separator |
| LI503/OUT | Water coalescer level |
| LVC502-SR/PID1/OUT | Water coalescer outlet valve |
| LI101/OUT | Water tank Level |
| FIC302/PID1/OUT | Inlet AFR controller 1 valve opening |
| FIC302/PID1/PV | Inlet AFR controller 1 process value |
| FIC301/PID1/PV | Inlet AFR controller 2 process value |

The analysis focusses on the 17 process variables (list available in Table 1) with a standard deviation greater than a given threshold to eliminate variables that do not move at all in the available dataset. The variables are mean-centered to cope with the two operating points, the alarm series for operating point A are generated using the standard deviation of the normal operation data in point A, and the alarm series for operating point B are generated using the standard deviation of the normal operation data in point B. The noise level in the system has a high variability, therefore $\tilde{k}$ is set to 10.

### *3.2 Variable selection*

Variable selection and training of the classification algorithm are done on the data of one operating point, and tested on the data of the other operating point. Two scenarios are presented:

- Scenario AB: train on point A and test on point B.
- Scenario BA: train on point B and test on point A.

Figure 2 represents the distribution plots of the variables for the operating point A. Most variables have a skewed distribution which reflects the incipient development of the faults. The analytical computation of mutual information proposed by Verron et al. (2008) assumes that the variables $X$ follow a Gaussian distribution and that $X$ conditioned to $C = c$ for all the classes $c$ follow Gaussian distributions. These assumptions are not valid here, since some $X$ do not follow Gaussian distributions and neither do the $X$ that are conditioned to faults.

## 4. DISCUSSION

Figure 3 and Figure 4 summarize the accuracy scores of the classification with mutual information and mRMR variable ranking respectively for scenario AB and BA. The figures highlight the importance of variable selection for the fault detection and identification. The accuracy scores on the training data and on the test data increase as the number of variables selected gets higher, until a certain number of variables. The least informative variables bring noise that is actually deteriorating the performance of the model.

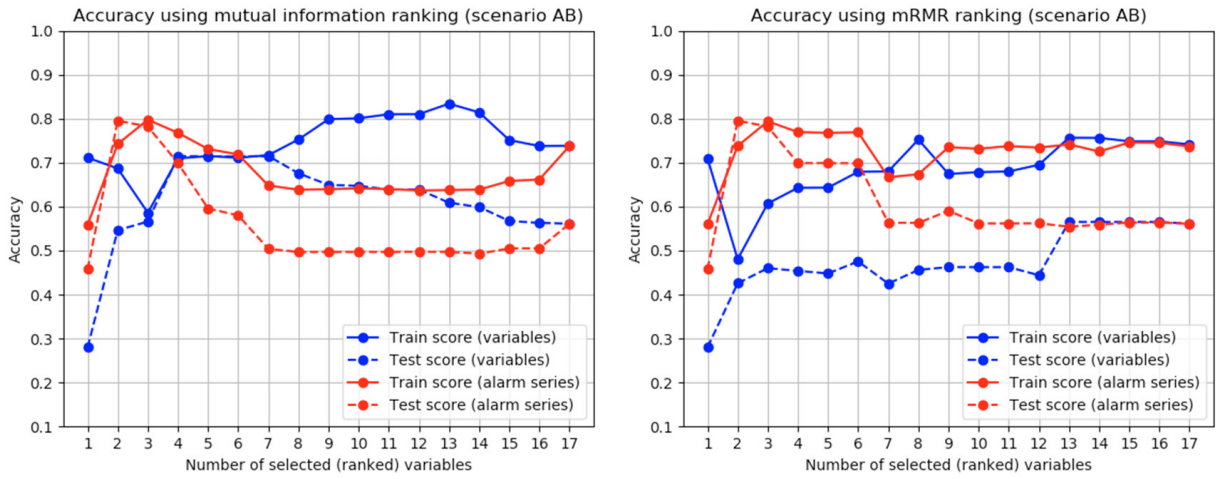


Fig. 3. Accuracy scores on training and test data for scenario AB using mutual information variable ranking (left) and mRMR variable ranking (right) applied to process variables and alarm series.

The robustness of the variable ranking can be assessed by comparing the training and test accuracy curves in Figures 3 and 4 for each case, in particular for the most informative variables. While the training and test curves follow similar shapes with the variable rankings on alarm series (both with mutual information and mRMR), the training and test curves with the variable rankings on process variables show different behaviours. Variables that improve the accuracy in the training case do not improve it in the test case and vice versa. The variables selected as the most informative in the training case do not help discriminating the faults during another occurrence. The variations in those variables are not the most representative of the faults.

A specific example is the most informative variable according to mutual information and mRMR on process variables in scenario AB (cf. Figure 3). Classification based on this variable (PT408/OUT) presents an accuracy of 71.2% on the training data but only of 28.0% on the test data. The good accuracy score on the training data corresponds to overfitting, as the variations of this variable are not representative of the faults. Figure 5 shows the most informative variable obtained when using respectively mutual information on process variables (i.e. PT408/OUT) and mutual information on alarm series (i.e. PIC501/PID1/OUT) in scenario AB. It can be noticed that PT408/OUT displays small variations that can be described as secondary variations compared to e.g. PIC501/PID1/OUT. Those variations are actually triggered by the fault, but only after it has propagated to the whole system: PT408/OUT is the pressure at the riser top, so when the pressure drops in the system (e.g. due to the leakage fault), PT312/OUT and PT417/OUT also drops, which eventually affects PT408/OUT. In addition, the alarm series corresponding to PT408/OUT does not activate in any of the faulty scenarios during the training occurrence (cf. Figure 5). For this reason, PT408/OUT is not considered informative when applying mutual information on the alarm series, which ranks the air outlet valve opening setpoint of the 3-phase separator PIC501/PID1/OUT as the most informative variable to discriminate the three faults.

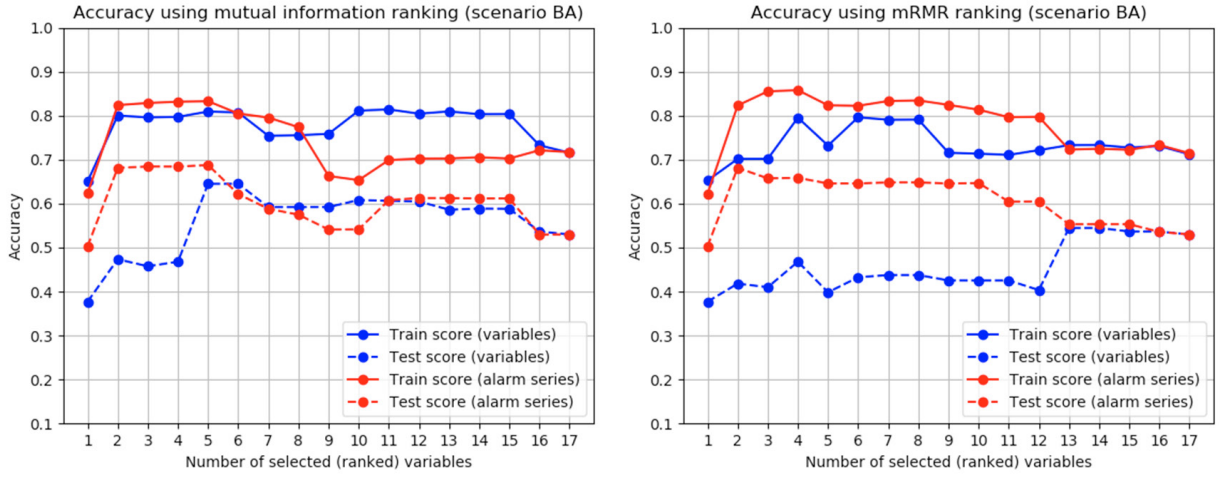


Fig. 4. Accuracy scores on training and test data for scenario BA using mutual information variable ranking (left) and mRMR variable ranking (right) applied to process variables and alarm series.

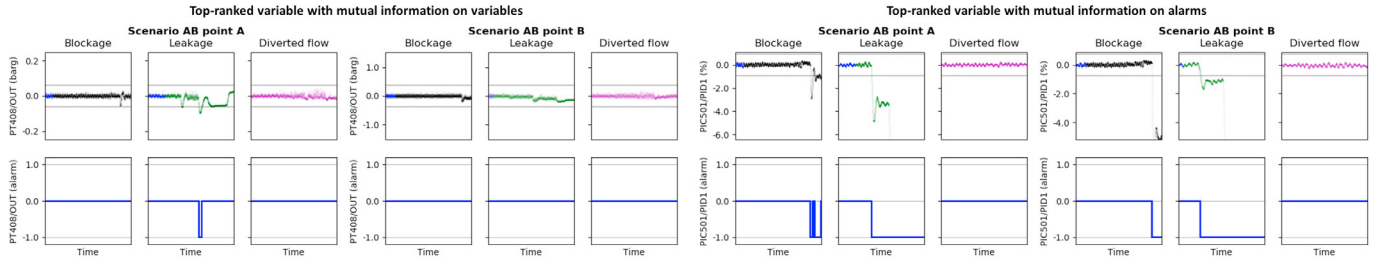


Fig. 5. Highest ranked variable using mutual information on process variables (left) and on alarm series (right) for scenario AB. Both the process variable and the corresponding alarm series are displayed for each variable and for each faulty episode: blockage (black), leakage (green) and diverted flow (magenta). The dotted lines indicate the alarm thresholds as defined in Eq. 1.

Therefore, the estimation of the relevance of the variables with regard to the class is more robust on alarm series than on process variables, and so is the redundancy estimation. The gap in classification accuracy between curves with variable selection on variables (in blue) and variable selection on alarm series (in red) gets larger when taking into account the redundancy criterion (mRMR), as pictured in the right plots of Figure 3 and Figure 4.

## 5. CONCLUSION

Variable selection techniques based on mutual information provide a scalable means to identify the relevant variables for fault detection and identification. This paper argues that variables with small variations are affected by noise and external disturbances, and thus do not represent the best variables for classifying faults. A discretization procedure is proposed to cut off small variations in the variables, so that the mutual information criteria focus on large variations. The estimation strategy based on alarm series improves the robustness of the variable selection in addition to facilitating the computation of the joint probability distribution functions.

## REFERENCES

Ardakani, M., Askarian, M., Shokry, A., Escudero, G., Graells, M., and Espuña, A. (2016). Optimal features selection for designing a fault diagnosis system. *Computer Aided Chemical Engineering*, 38, 1111–1116.

Fix, E. and Hodges, J.L. (1951). Discriminatory analysis - nonparametric discrimination: consistency properties. *USAF School of Aviation Medicine, Randolph Field, Texas*.

Ghosh, K., Ramteke, M., and Srinivasan, R. (2014). Optimal variable selection for effective statistical process monitoring. *Computers and Chemical Engineering*, 60, 260–276.

Kraskov, A., Stoegbauer, H., and Grassberger, P. (2011). Estimating mutual information. *Phys. Rev. E*, 16(1), 51–54.

Lucke, M., Chioua, M., Grimholt, C., Hollender, M., and Thornhill, N.F. (2018). On improving fault detection and diagnosis using alarm-range normalisation. In *Proceedings of 10th SAFEPROCESS Symposium, Warsaw, Poland, August 29-31, 2018*.

Ming, L. and Zhao, J. (2017). Review on chemical process fault detection and diagnosis. In *Proceedings of 6th AdCONIP Symposium, Taipei, Taiwan, May 28-31, 2017*, 457–462.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238.

Peres, F.A.P. and Fogliatto, F.S. (2018). Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers and Industrial Engineering*, 115, 603–619.

Ross, B.C. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2), e87357.

Russell, E.L., Chiang, L.H., and Braatz, R.D. (2000). *Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes*. Advances in Industrial Control. Springer, London.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois.

Shu, Y., Ming, L., Cheng, F., Zhang, Z., and Zhao, J. (2016). Abnormal situation management: Challenges and opportunities in the big data era. *Computers and Chemical Engineering*, 91, 104–113.

Stief, A., Tan, R., Cao, Y., and Ottewill, J.R. (2018). Analytics of heterogeneous process data : Multiphase flow facility case study. In *Proceedings of 10th ADCHEM Symposium, Shenyang, China, July 25 - 27, 2018*.

Su, J., Wang, D., Zhang, Y., Yang, F., Zhao, Y., and Pang, X. (2017). Capturing causality for fault diagnosis based on multi-valued alarm series using transfer entropy. *Entropy*, 19(12), 5–8.

Tong, C. and Palazoglu, A. (2016). Dissimilarity-based fault diagnosis through ensemble filtering of informative variables. *Industrial and Engineering Chemistry Research*, 55(32), 8774–8783.

Verron, S., Tiplica, T., and Kobi, A. (2008). Fault detection and identification with a new feature selection based on mutual information. *Journal of Process Control*, 18(5), 479–490.

Yu, W., Yang, F., and Knuth, K.H. (2015). Detection of causality between process variables based on industrial alarm data using transfer entropy. *Entropy*, 17, 5868–5887.

Zhao, C. and Gao, F. (2017). Critical-to-fault-degradation variable analysis and direction extraction for online fault prognostic. *IEEE Transactions on Control Systems Technology*, 25(3), 842–854.