Fault detection and identification with a new feature selection based on mutual information

Sylvain Verron, Teodor Tiplica, Abdessamad Kobi

Fault detection and identification with a new feature selection based on mutual information

Sylvain Verron * , Teodor Tiplica , Abdessamad Kobi

LASQUO/ISTIA, University of Angers, 62, Avenue Notre Dame du Lac, 49000 Angers, France

Abstract

This paper presents a fault diagnosis procedure based on discriminant analysis and mutual information. In order to obtain good classification performances, a selection of important features is done with a new developed algorithm based on the mutual information between variables. The application of the new fault diagnosis procedure on a benchmark problem, the Tennessee Eastman Process, shows better results than other well known published methods.

Key words: FDI, Discriminant Analysis, Mutual Information

1 Introduction

Nowadays, the control of complex manufacturing systems is becoming an essential task in order to reduce the variability of products, or to insure a safety

* Corresponding author. Tel.: 00 33 2 41 22 65 80; fax: 00 33 2 41 22 65 21
Email address: sylvain.verron@univ-angers.fr (Sylvain Verron).

production (for humans and materials). In order to achieve this activity of supervisory control, some authors call this AEM (Abnormal Event Management) [1]. This is composed of three principal steps: firstly, a timely detection of an abnormal event; secondly, diagnosing its causal origins (or root causes), which is the purpose of this article; and finally, taking appropriate decisions and actions to return the process in a normal working state. We will call "fault" an abnormal event, it is classically defined as a departure from an acceptable range of an observed variable or a calculated parameter of the process [1]. So, a fault can be viewed as a process abnormality or symptom, like an excessive pressure in a reactor, or a low quality of a part of a product, and so on.

Three major categories of methods can be identified to achieve process control: data-driven, analytical and knowledge-based [2,1]. The knowledge-based category represents methods based on qualitative models (FMECA - Failures Modes Effects an Critically Analysis; Fault Trees; Decision Trees; Risk Analysis) [3,4]. The analytical methods compare real process data to those obtained by mathematical models of the system [5]. But, for large systems, obtaining reliable detailed models is difficult and can often conduct to false conclusions on the state of the system. As a consequence, for systems with no (or not enough reliable) models, one could prefers the application of data-driven techniques which are quantitative models based on rigorous statistical development of the process data.

In the literature, many data-driven techniques for fault detection and diagnosis can be found. Thus, Statistical Process Control (SPC) [6] techniques are used in the case where, for different reasons, one is not able to control continuously the process (or some process variables) and as a consequence samples are taken (at constant or variable time steps). In the frame of the SPC, the

control charts are very simple and useful fault detection tools. The $\overline{X}$ [7], CUSUM (CUmulative SUM) [8] or EWMA (Exponentially Weighted Moving Average) [9] control charts are widely used in manufacturing plants to monitor the mean of a process variable. In order to monitor and control the dispersion of a process parameter, control charts like: R, S and S^2 are frequently employed [6]. In the case of qualitative variables, control charts like p, np, c, u have been proposed to control the non-conformities or the non-conforming products [6]. Unfortunately, due to the complexity of the manufacturing processes, monitoring process parameters on different control charts like those presented above is not sufficient to control the whole manufacturing process. Indeed, this strategy has some major drawbacks : the false alarm rate are inflated, the correlation between the variables is not taken into account, the monitoring is complicated especially when the number of charts is becoming important. For these reasons, in order to control several process parameters in the same time, multivariate control charts like the T^2 of Hotelling [10], the MEWMA (Multivariate EWMA) [11] and MCUSUM (Multivariate CUSUM) [12] have been developed. The multivariate control charts give an easy to understand representation of a multivariate process evolution but have also a major disadvantage: they do not give any information concerning the process parameter responsible for an out of control situation detected on the chart. Otherwise, we know that the process is out of control but we are unable to do a reliable diagnosis by saying which process parameter changed or what was wrong with the process. Many researches have been done in the last few years and a great number of methods were proposed to overcome this inconvenience ([13–23]). A non-exhaustive comparative study on the recent diagnosis techniques in the field of the SPC can be found in [24].

Other FDI approaches are based on data analysis methods: like the Principal Component Analysis (PCA) [25], Discriminant Analysis, Projection to Latent Structures (PLS), supervised or not-supervised classification techniques, etc. The PCA method captures the variability of a process in a lower dimensional space than the process space. One can monitor the T^2 metric on the new PCA axes or monitor the residuals (Q chart) of the PCA model [26]. Some extensions of the PCA method like the Moving PCA [27], the Multiway PCA [28], have been proposed in order to deal with serial correlations in the process data or to monitor some batch processes. Numerous applications of PCA for the fault detection and diagnosis can be found [29–31]. For example, we can cite Harkat et al. [29] in which the authors propose a new detection index D_i on the PCA scores, in order to monitor an air quality monitoring network. The PLS (Projection to Latent Structures) based approaches are used to establish a relationship between the space of the final product quality characteristics and the space of the process parameters. These techniques maximize the covariance between a predicted matrix (generally product quality data) and a predictor matrix (all other variables of the system) [32]. Some extensions of the PLS techniques, like the Multiway PLS [33], can be found in the literature for batch processes. A good comparative study concerning some of these fault detection techniques can be found in [34].

Although all these techniques are well designed for the fault detection, one of the most relevant technique for the diagnosis is the supervised classification. Indeed, it is usual to see the fault diagnosis as a classification task whose objective is to class new observations to one of the existing classes. Many methods have been developed for supervised classification. The Fisher Discriminant Analysis (FDA) [35], the Support Vector Machine (SVM) [36], the k-Nearest

Neighborhood (kNN) [37], the Digital Filtering and Discriminant Analysis (DFDA) [38], etc. Other recent and new emerging classification approaches are based on the use of some artificial intelligence techniques. Of course, we can cite here the Artificial Neural Networks (ANN) [35] and particularly the MultiLayer Perceptron (MLP) which is a very efficient non linear classifier; the Bayesian network classifiers [39] like the Naïve Bayesian Network (NBN) [40], the Tree-Augmented bayesian Network (TAN) [39], the k-dependence Bayesian classifier [41], the Condensed Semi Naïve Bayesian Network (CSNBN) [42], etc.

Nevertheless, the classification (diagnosis) is an hard task in the detection space like the PCA space or the original space (like used for the different control chart). Indeed, the space used for the detection is not frequently a space where the different potential fault of a process can be well discriminated : there are too many variables giving some noise for the discrimination, or the transformation made in order to increase the detection performance leads to decrease the classification performance. Moreover, just some variables of the space are significant for the discrimination of the different faults. So, it seems important to highlight here that in the case of non-informative (insignificant) variables, the performances (in term of classification error rate) of the classifiers presented above are decreasing as the number of non-informative variables increases. Therefore a selection of the informative variables for the classification task should be done in order to increase the accuracy of the classification [43].

In this article, we present a new data-driven technique to diagnose the faults of a system in steady state conditions. This procedure includes a feature selection of the most informative variables of the system. Then, fault diagnosis is made on these important variables with a discriminant analysis. Assumptions are

made that a fault detection technique has detected an observation of the system as faulty, and that two or more faults cannot occur simultaneously.

The article is structured in the following manner: in the section 2, we present briefly some basic theoretical aspects concerning the discriminant analysis; in the section 3 we propose a new fault diagnosis procedure based on the use of the discriminant analysis technique on the variables selected according to theirs informative potential for the classification task; the section 4 is an application of this procedure on a benchmark problem - the Tennessee Eastman Process; finally, in the section 5 some concluding remarks and outlooks of the proposed fault diagnosis method are mentioned.

2 Discriminant Analysis

Concerning the Discriminant Analysis (DA), one can distinguish two different aspects: the descriptive and the predictive aspect. Given a system with p variables (descriptors) and with k identified classes, the descriptive discriminant analysis (or Fisher Discriminant Analysis) is generally used to find $k-1$ new descriptors of the system. These new $k-1$ descriptors (which are a linear combination of the original descriptors) are supposed to maximally discriminate between the k identified classes of the system. The other aspect of the discriminant analysis is the predictive one. The purpose of the predictive aspect is principally to allocate a new observation to one of the k identified classes of the system. In the remaining of this article, we will focus mainly on the predictive potential of the Discriminant Analysis and we will consider Discriminant Analysis as a supervised classification method [35].

Principally, two decision rules can be applied with a Discriminant Analysis: the geometric one and the probabilistic one. The geometric one attributes to the observation the class with the nearest mean to the observation. This rule can conduct to false attribution if the variability of the classes are not identical. We will prefer the probabilistic DA that is based on the Bayes decision rule. Giving k classes C_i ($i \in \{1, \ldots, k\}$) a priori known, this rule allocates a new observation $\boldsymbol{x}$ to the class C_i with the maximum a posteriori probability $P(C_i|\boldsymbol{x})$ giving the value of each descriptor, as defined in the equation 1.

$$\boldsymbol{x} \in C_i,\ if\ i = \underset{i=1,\ldots,k}{\operatorname{argmax}}\{P(C_i|\boldsymbol{x})\} \tag{1}$$

This decision rule is named "Bayes decision rule" because it is based on the Bayes rule which gives the value of $P(C_i|x)$ as stated in equation 2.

$$P(C_i|\boldsymbol{x}) = \frac{P(C_i)P(\boldsymbol{x}|C_i)}{P(\boldsymbol{x})} \tag{2}$$

where $P(C_i)$ is the a priori probability of $\boldsymbol{x}$ to belong to the class C_i. This probability can be fixed differently : uniformly on all the classes ($P(C_i) = P(C_j)$ for each couple i, j); based on the historical data, with $p(C_i) = \frac{n_i}{n}$, where n_i is the number of observations of the class C_i, and n is the total number of observations ($n_1 + \ldots + n_k = n$); or under assumption that some classes are most probable than other. In certain cases (noisy data, overlapping classes), the discrimination is not easy with the decision rule of equation 2. thus, it can be applied some techniques allowing the rejection of ambiguous observations: ambiguity rejection, distance rejection (see [44,45]). In the equation 2, we can see that for each class, the denominator is the same, so, it is not implicated

in the discriminant function. Then, equation 1 can be rewritten as:

$$\boldsymbol{x} \in C_i,\ if\ i = \underset{i=1,\ldots,k}{\operatorname{argmax}}\{P(C_i)P(\boldsymbol{x}|C_i)\} \tag{3}$$

More, like in numerous articles ([46] for example), in order to simplify following equations in the case of a gaussian distribution, we can rewrite equation 3 by using the cost function K:

$$K_i(\boldsymbol{x}) = -2\log(P(C_i)P(\boldsymbol{x}|C_i)) \tag{4}$$

where *log* represents the natural logarithm. So, for each observation $\boldsymbol{x}$, the allocating rule becomes:

$$\boldsymbol{x} \in C_i,\ if\ i = \underset{i=1,\ldots,k}{\operatorname{argmin}}\{K_i(\boldsymbol{x})\} \tag{5}$$

In this article, we consider the classical assumption that data follow a multivariate normal distribution. The density function f of a normal variable conditionally to a class C_i can be written as in equation 6, where p is the dimension of the observation $\boldsymbol{x}$, $\boldsymbol{\mu}_i$ is the mean vector of the class C_i, $\boldsymbol{\Sigma}_i$ is the covariance matrix of the class C_i, the symbol ^{t} represents the transpose of a vector or a matrix, and $|\boldsymbol{\Sigma}_i|$ represents the determinant of the matrix $\boldsymbol{\Sigma}_i$.

$$f(\boldsymbol{x}|C_i) = \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i))}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \tag{6}$$

We are reminding here that, for n_i samples of the class C_i, the Maximum Likelihood Estimation (MLE) gives [35]:

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_j \tag{7}$$

and:

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)^t \tag{8}$$

We notify that in certain cases (presence of outliers in the data, limited number of samples), the MLE can be quite inaccurate. In these cases, other estimators, called robust estimators, can be used. There are several types of robust estimators, we can cite for example M-estimator (maximum likelihood type estimator) or R-estimator (estimator based on rank transformation) [47,48].

In the case of the multivariate normal distribution, the cost function K_i (see equation 4) becomes:

$$\begin{aligned} K_i(x) = & (\boldsymbol{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - 2log(P(C_i)) \\ & + log(|\boldsymbol{\Sigma}_i|) + plog(2\pi) \end{aligned} \tag{9}$$

and we can see that the term $plog(2\pi)$ is the same for each class C_i, so it does not contribute to the discrimination between classes. This cost function is named Quadratic Discriminant Analysis (QDA). This classification rule makes quadratic boundaries between each class.

If not enough data are available for a good estimation of each covariance matrix $\boldsymbol{\Sigma}_i$, it can be assumed that each $\boldsymbol{\Sigma}_i$ is equal to $\boldsymbol{\Sigma}_{pool}$ which is a pooled covariance matrix given in equation 10 (where n_i is the number of observed samples for the class C_i).

$$\boldsymbol{\Sigma}_{pool} = \frac{(n_1 - 1)\boldsymbol{\Sigma}_1 + \ldots + (n_k - 1)\boldsymbol{\Sigma}_k}{n - k} \tag{10}$$

So, in the case of pooled covariance matrix, the cost function given in the equation 9 can be reduced to equation 11:

$$K_i(x) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_{pool}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - 2log(P(C_i)) + Cst \tag{11}$$

where Cst is a constant equal to $log(|\boldsymbol{\Sigma}|) + plog(2\pi)$. The cost function given in the equation 11 is named Linear Discriminant Analysis (LDA). This classification rule makes linear boundaries between each class. If we consider that each observation has the same weight, under the assumption that we have the same number of observations in each class, $P(C_i)$ is equal for each class and so the decision rule becomes to attribute to a new observation the class with the least Mahalanobis distance (term $(\boldsymbol{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_{pool}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$) to the class center.

We have shortly presented some basic concepts of the QDA and the LDA, but as indicated in the first section the performances of these classifiers are not optimal in the presence of non-informative variables. It means that each variable give some information for the classification task, but give also a certain noise. A non-informative variable will be a variable giving little or none information for the classification task, but bringing some noise which will increase the classification error. So, a selection of informative variables is necessary in order to obtain low classification error. In the next section we will propose a new fault diagnosis method including a variable selection algorithm based on the concept of the mutual information.

3 Fault diagnosis based on discriminant analysis and mutual information

In this section, we propose a fault diagnosis method including a variable selection algorithm based on the mutual information. The algorithm is based on two major steps: firstly, the variables are sorted according to theirs shared mutual information with the class variable and secondly the more informative variables are chosen based on the classification error rate.

3.1 Sorting the variables of the system with Mutual Information

The Mutual Information (I), or transinformation, of two random variables x and y can be viewed as a quantity measuring the mutual dependence of the two variables [49,50]. The mutual information is widely used in applications area like the training of hidden Markov models, the prediction of the ribonucleic acid, the registration of medical images and in feature selection for machine learning.

The mutual information between two random variables x and y can be computed as indicated in the equation 12, 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.

$$I(x; y) = \sum_{x,y} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \tag{12}$$

In supervised classification, one can view the classes as a multinomial random variable (that we will name C) with k possible values (where k is the number of

classes of the system). So, one will be able to compute the mutual information and identify the informative variables.

3.1.1 First approach: an univariate approach

In [42], authors demonstrate that the mutual information between a gaussian (normaly distributed) variable and a multinomial (discrete) variable can be computed as indicated by equation 13. In this equation, it is assumed that: C is a multinomial random variable with k possible values and a probability distribution given by $P(C = c) = P(c)$; X is a random variable with a normal density function of parameters μ and σ^2; X conditioned to $C = c$ follows a normal density function with parameters μ_c and ${\sigma_c}^2$.

$$I(X;C) = \frac{1}{2}\left[\log(\sigma^2) - \sum_{c=1}^{k} P(c)\log(\sigma_c^2)\right] \tag{13}$$

In this way, the mutual information (I) can be computed for all the variables (descriptors) of the system. The most important variables for the classification task will be those having an high I value comparing to other variables. So, we can sort the variables in decreasing order of the mutual information that they share with the class variable, and thus we obtain the variables sorted from the most informative one to the least informative one for the classification task.

This approach is very fast but has a major drawback: the redundancy. Indeed, assuming two variables with high mutual information with the class variable, having these 2 variables in the model is not optimal if they share the same information with the class variable (redundancy of the information), because they bring the same information for the classification, but each one add his

own noise to the classification, giving more misclassification errors. So, the goal is to select a group of variable giving maximum information, but adding a minimum of noise for the classification.

3.1.2 Second approach: a multivariate approach

We demonstrate (see appendix A) a new result about the mutual information between a multivariate gaussian variable and a multinomial (discrete) variable. This mutual information can be computed as indicated by equation 14. For this equation, it is assumed that: C is a multinomial random variable with k possible values and a probability distribution given by $P(C = c) = P(c)$; $\boldsymbol{X}$ is a random variable with a multivariate normal density function of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$; $\boldsymbol{X}$ conditioned to $C = c$ follows a multivariate normal density function with parameters $\boldsymbol{\mu_c}$ and $\boldsymbol{\Sigma_c}$.

$$I(\boldsymbol{X}; C) = \frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) - \sum_{c=1}^{k} P(c)\log(|\boldsymbol{\Sigma_c}|)\right] \tag{14}$$

It can be observed easily that if we assume that $\boldsymbol{X}$ is univariate, then $\boldsymbol{X} \sim N(\mu, \sigma^2)$, $|\Sigma| = \sigma^2$ and consequently the mutual information given in equation 14 is computed in the same way as in the equation 13. So, the result demonstrated by Perez (equation 13) can be view as a particular case of our demonstrated result (equation 14).

The mutual information (I) can be computed for all different groups of variables (descriptors) of a system. The most important group of variables for the classification task will be the one that has a large I value. Indeed, adding more variables to the model increases the amount of information of the model. In

order to exploit this new result about mutual information, we have developed a new sorting algorithm of the variables taking into account both the information and the redundancy. This sorting algorithm is given below where V_i represents the variable selected at step i, and p is the number of variables of the system.

(1) (Initialization) $\boldsymbol{V} \leftarrow \emptyset$

(2) (Selection loop) for $i = 1$ *to* p

  a) (Computation of the mutual information) compute I for all possible groups of dimension i including $\{V_1, \ldots, V_{i-1}\}$

  b) (Selection of the variable) select as V_i the variable allowing the maximization of the mutual information I

(3) Output the set $\boldsymbol{V}$ containing the ordered variables.

The figure A.1 illustrates the above algorithm for a system described by 4 variables. This system has two different classes, and we have simulated 100 observations for each. Parameters of the two classes are given by :

$$\boldsymbol{\mu}_1 = [1\ 2\ 2\ 1] \qquad \boldsymbol{\mu}_2 = [2\ 1\ 1\ 2]$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.3 & 0.6 & 0.4 \\ 0.3 & 1 & 0.4 & 0.2 \\ 0.6 & 0.4 & 1 & 0.3 \\ 0.4 & 0.2 & 0.3 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.8 & 0.3 & 0.6 \\ 0.8 & 1 & 0.2 & 0.5 \\ 0.3 & 0.2 & 1 & 0.5 \\ 0.6 & 0.5 & 0.5 & 1 \end{pmatrix}$$

In the first step, the mutual information of each variable is computed, and the

variable 3 is retained. At the second step, all possible groups of dimension 2, but containing the first retained variable (variable 3), are formed and evaluated with mutual information. We can see that the group {3,1} maximize the I. So, the variable retained for this step is the variable 1. In the same way, at step three, we formed all possible groups of dimension 3, but containing the two first selected variables (variable 1 and 3), and evaluated them with the I, and so on.

[Fig. 1 about here.]

This multivariate approach is more computation-consuming than the univariate one. The complexity of the univariate approach is equivalent only to the first step of the multivariate one. But, the multivariate approach that we propose is more exact than the univariate approach because it takes into account the redundancy of information between the variables. To illustrate that on the 4 variables example, the univariate approach gives $\{2, 4, 3, 1\}$, compared with the multivariate approach which selects the variable 3 before the variable 4 and gives $\{2, 3, 4, 1\}$. It signifies that information given by the variable 4 is more redundant than the information given by the variable 3. This can be due to the fact that the part of the mutual information given by the correlation between variable 2 and 3 is more important than the part of the mutual information given by the correlation between variable 2 and 4

3.2 Fault diagnosis procedure

The method is based on a wrapper approach [51]. The objective of this procedure is to select a group of variables S giving good discrimination performances

between the different faults (classes) of a system. The optimal way to obtain a good classifier would be to estimate the misclassification rate of all the possible groups S. This solution can be effective for system with few variables, but in many cases the number of possible groups is too high for this exhaustive and time-consuming search. So, we propose a new procedure for the fault diagnosis of a system with p descriptors (variables) as illustrated on the figure A.2. We precise here that in order to take into account processes with multiple set-points, a normalization of the data is made before the application of the proposed procedure.

[Fig. 2 about here.]

In the first step, we sort the variables from the most informative to the least informative one as presented previously (see paragraph 3.1). We precise that if the QDA is used, the equation 14 is directly applied, but if the LDA is used the equation 14 becomes:

$$\begin{aligned} I(\boldsymbol{X};C) &= \frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) - \log(|\boldsymbol{\Sigma}_{pool}|)\right] \\ I(\boldsymbol{X};C) &= \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_{pool}|}\right) \end{aligned} \tag{15}$$

In the second step of the procedure, we iteratively compute the misclassification rate of each group S_i of dimension i, composed of $V_1, V_2, \ldots, V_i$ where V_1 is the first most informative variable, V_2 is the second most informative variable, and so on. The misclassification rate is obtained with a well known technique: the m-fold cross validation [52]. In the m-fold cross validation, the training dataset is divided into m subsets and one of this is used as the testing set while the m-1 other subsets are put together to form the training set. Then the average and the standard deviation of the error for all m trials is computed

[35].

Once the misclassification rate is computed for the p groups, the error function of the number of features (variables) can be drawn (see figure A.3).

[Fig. 3 about here.]

On the figure A.3, we can see 3 areas: Area I represents an area where the error decreases when the number of features increases, indicating that the number of features is not sufficient to obtain good discrimination; Area II is a quite constant error zone with N_{min} the number of features having the lower error; and in area III, the error increases when the number of features increases implying that some of the features are not informative for the classification. Our goal is to select the group of variables giving the lower error but with the lower number of features. Thus, we will consider that N_{best} is a better choice than N_{min} if its classification error is statistically equivalent to the classification error of N_{min}. So, the main idea is to select the group S_{min} of dimension N_{min} giving the lower average misclassification rate. After that, hypothesis tests [53] are made in order to compare the average error of S_{min} and the average error of all S_i with $i < N_{min}$. The test realized is a classical equality test of the mean of two distributions (assumed normal) with unknown (but estimated) variances. Then the group S_b (group where the average error is statistically equal to the average error of S_{min} but with $N_{best} < N_{min}$) is selected.

Using the same example as in the previous paragraph 3.1.2 (see the 4 variables system), one can see (on the table A.1) the average and the standard deviation of the error for the 4 groups of variables selected. We can see that the minimum average error is obtained for the group S_3, so we compare the average error of

S_3 with the average error of the group containing less variables than S_3 (so, S_1 and S_2). We give the result of the hypothesis tests (such as 1 signifies that errors are statistically equal and 0 indicates that errors cannot be considered equal) and select $S_2 = \{2, 3\}$ as the the best group.

[Table 1 about here.]

Once the best group S_b has been identified, we can easily use QDA or LDA with those variables, and classify new observations of the system more accurately. If the plant engineers confirm the diagnosis procedure, the new faulty observation and his belonging class are stored in the database. This iteration will improve the diagnosis performances of the procedure for future faulty observations.

Now, we will see an application of this approach on a benchmark problem: the Tennessee Eastman Process (figure A.4).

[Fig. 4 about here.]

4 Application to the TEP

4.1 Presentation of the TEP

The Tennessee Eastman Industrial Challenge Problem was created by the Eastman Chemical Company to provide a realistic industrial process in order to evaluate process control and monitoring methods [54]. The Tennessee Eastman Process (TEP) is a chemical process. It is composed of five major operation units (see figure A.4): a reactor, a condenser, a compressor, a stripper and a separator. Four gaseous reactants A, C, D, E and an inert B are fed

to the reactor where the liquid products F, G and H are formed. This process has 12 input variables and 41 output variables. It has 20 types of identified faults.

The TEP is entirely described in the article of Downs and Vogel [54]. This process was simulated on Matlab by Ricker [55]. This plant is open-loop unstable. So, it is also a benchmark problem for control techniques. Some fault detection approaches have been tested on the TEP [34,56,57]. Some fault diagnosis techniques have also been tested on the TEP [2,58–61] with the plant-wide control structure recommended in Lyman and Georgakis [62]. On the same control structure and for steady state conditions (base case chosen on the simulator), we have taken into account 3 types of faults named: fault 4, 9 and 11 (see table A.2) because they are good representations of overlapping data and so, are not easy to classify. In other articles [59,60], authors focus only on 3 types of faults and give the datasets they used. For this reason, we will take the same data that in these articles and we will compare our approach to theirs.

[Table 2 about here.]

For each type of fault, we have 2 datasets: a training sample and a testing sample, containing respectively 480 and 800 observations as indicated on the table A.2. Of course, having such big amount of data in the database (1440 samples) is not very realistic in the case of real processes. We are using these 1440 samples of training only in order to make correct comparison with the results of other authors. Nevertheless, we will also take into account the scenario where only a small number of observations is available for each class of fault.

All computations have been made on Matlab. For an unbiased comparison

with the previously cited methods, new faulty observations were not added to the fault database, as it was mentioned in the figure A.2. The fault database is composed of the 3 training samples. We also have to notify that only 52 variables are taking into account in this problem because an input variable (the reactor agitator speed) is constant.

4.2 The new procedure applied on the TEP

As this application has 52 variables, an exhaustive search of all possible groups that can be formed is impossible: 4.5036e+015 possible groups. So, we will apply the procedure that we have developed which is a free choice classifier (the only condition is that the classifier assumes that the data are class conditional normally distributed).

4.2.1 First step

We have compared, at this step, the univariate and the multivariate sorting approaches. We have also applied the multivariate approach with LDA and QDA. We remind that for sorting the variables, the choice of LDA or QDA has no effect with the univariate approach (because no covariance matrix is computed). A graphical result of the univariate approach is given in the figure A.5.

[Fig. 5 about here.]

On the figure A.5, we can see that 2 variables have very high values of mutual information (variables 51 and 9). We can also observed that 3 variables (variables 50, 19 and 18) obtain medium value of mutual information, and that

several variables have a quite little value (variables 20, 38, 21, 37 and 46).

The results of the different approaches (univariate and multivariate) for the eight first variables are given in the table A.3.

[Table 3 about here.]

We observe that the variables 9 and 51 seems to be the most informative one because they are selected in the two first positions by each algorithm. Others ordered variables are quite different between each algorithm. So, it seems to be difficult, for the moment, to state that a variable in particular (other than variables 9 and 51) will be important for the classification task.

4.2.2 Second step

In a second step, we have to apply a m-fold cross validation. Of course, the choice of m is dependent of the number of example of the training dataset. Here, as we have numerous samples of each classes, a value of 10 is chosen (this value is suggested in many articles applying a cross validation). So, we apply a 10-fold cross validation on the group of 52 variables. The results are given in the figure A.6 where the upper graphs are the application of the LDA and the lower graphs are the application of the QDA. In each case, the left graph represents the results given with the univariate approach and the right graphs are the results of the multivariate approach. For each graph, the N_{best} and the N_{min} (see section 3.2) are represented.

[Fig. 6 about here.]

On the figure A.6, we can firstly observe that the misclassification average error of the Linear Discriminant Analysis (LDA) is more important than the one

induced by the Quadratic Discriminant Analysis (QDA). More, we can also see that (like it was shown on the figure A.3) that to increase the number of features can, in a first time, decrease the average error, but than in a second time, the increasing of the number of feature leads to an increased average error. We can finally observed that just one variable (the variable 51) is not able to discriminate correctly between the different classes of fault.

4.2.3 Third step

At the third step, the group giving the lower misclassification rate is selected. Then, hypothesis tests (with $\alpha = 5\%$) are made in order to find the smaller group for an equivalent misclassification rate. The table A.4 gives the results of this step.

[Table 4 about here.]

We can see that in the four cases, variables 9 and 51, (respectively the reactor temperature and the reactor cooling water valve position) are selected. We can conclude that these variables are very important in order to discriminate the 3 types of faults. Of course, as the faults 4 and 11 are both a change in the reactor cooling water inlet temperature, it is logical that the variables representing the reactor temperature (variable 9) and the reactor cooling water valve position (variable 51) can help to discriminate between these two faults. But the advantage of the feature selection algorithm is that it concludes that these variables can also discriminate the fault 9 (D feed temperature), which has no evident link with the 2 variables. With different methods (contribution charts, discriminant partial least square, genetic algorithms combined with Fisher discriminant analysis), Chiang (in [59]) has obtained the same conclu-

sion.

But, we can view that the multivariate algorithm has also selected the variable 21 (reactor cooling water outlet temperature). It is surprising to see that this algorithm conclude on the importance of this variable even though the univariate approach ordered it at the 8th position (see table A.3). So, it will be interesting to compare classification performance between {51,9} and {51,9,21}, this study will be done in the next paragraph.

4.2.4 Fourth step

As mentioned in the procedure of fault diagnosis (see figure A.2), in the fourth step we learn the classifier and to classify a new faulty observation of the industrial system. In an objective evaluation purpose of our procedure and to compare it with the results of other published methods (like Support Vector Machines), in a first case we classified 2400 new observations (800 of each type of fault) of the TEP. The results are given in the table A.5. For the LDA and QDA methods, we compute the misclassification rate (percentage of observations which are not well classified). We are also giving the results of other methods on the same data. The results for the SVM (Support Vector Machines), PSVM (Proximal Support Vector Machines) and ISVM (Independent Support Vector Machines) methods are extracted from [59] and [60]. The results of the different methods in the space of the 52 variables {All} are also given in table A.5 in order to demonstrate the advantage of a well chosen reduced space for discrimination. But, as mentioned before, having so many faulty observations is not common in practice. So, we analyzed the application of the LDA and QDA techniques in the case of less data (60 samples of train-

ing for each classes). The results (noted LDA_{60} and QDA_{60}) are also given in table A.5.

[Table 5 about here.]

An evident remark is that all the methods give better results in the reduced space than in the space of all the variables. So, as expected, in order to well diagnosis the disturbances of an industrial system, a feature selection is necessary.

We can also view that on the same space ({All} or {51,9}) QDA outperforms all the other methods. The fact that QDA outperforms LDA is not surprising because LDA is a linear technique while QDA is a quadratic one. A graphical explanation for these good results can be obtained in figure A.7, where we can view that the 3 classes are overlapping in the multidimensional space with non-linear frontiers for separation. These shapes signify that data are normal, and this is an assumption of the QDA classifier. So, it is logical that QDA obtains good performances on this problem.

An interesting remark is the fact that the results of QDA on the reduced space ({51,9} or {51,9,21}) are quite similar to the results of the SVM based techniques (SVM, PSVM, ISVM). Indeed, SVM are techniques requiring more computational potential than QDA.

[Fig. 7 about here.]

An other important point (see table A.5), is that the best classification result is obtained by the QDA classifier in the space {51,9,21}. This reduced space has been found with our algorithm for feature selection. So, we can say that this approach is relevant for the feature selection in the industrial systems.

Finally, we can see that with a training data set of 60 samples by classes, results are quite good in the reduced space. But, as attended, in the space of all the variables poor performances are obtained. Indeed, obtaining a correct estimation of a 52×52 covariance matrix with only 60 observation is quite difficult. In this case (no many data), other estimation approaches should be applied [63,64,46] in order to increase the precision of the estimation.

The confusion matrix for the QDA in the space {51,9,21} is given on table A.6 and gives us the possibility to see how the discrimination of the different faults is done by the QDA technique. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. For example, for 800 tested observations of fault 4, the diagnosis procedure gives 6 observations as the fault 11, and 794 observations as the fault 4, so 0,75% (6/800) of misclassified observations for the fault 4.

[Table 6 about here.]

We can see that the fault 4 and 9 are well discriminated. But, the fault 11 is less discriminable than the 2 others because this fault overlaps on the two other.

We think that the proposed procedure can be applied online. Indeed, on a Pentium 2.4GHz with Matlab, the first step (ordering the variable, multivariate approach) takes near 15 seconds, the second step (the 10-fold cross validation) takes about 8 seconds, the third step (the hypothesis tests) is about 0.15 second, and the last step (classification tests) takes 0.03 second for 2400 observations. So, as the feature selection steps (the three first steps) can be made during normal working condition of the process, the real diagnosis step (the last step) takes near 1×10^{-5} seconds, and can be used online.

5 Conclusion and outlooks

We have presented a new supervised procedure for FDI. This method, after a feature selection, makes classification of faulty observation with a discriminant analysis (QDA or LDA). The feature selection algorithm exploits a new result that we have demonstrated on the mutual information between the class variable and a normal multivariate variable. The application on the TEP of this FDI procedure demonstrates that this method is relevant in order to do a good feature selection and a good classification, and we obtained better results than other published methods.

The feature selection algorithm and the classifier share a common assumption: the normality of the data. It will be interesting to study the effect of non-normality on this approach. But, it will be more motivating to improve this method in order to take into account not normally distributed data. An attractive way of research can be the use of gaussian mixture models [35]. But, for instance, an analytical form of the entropy of a gaussian mixture does not exist. So, mutual information between a gaussian mixture and a class variable cannot be computed directly. This can be an interesting research field. An other outlook of interest would be the extension of the method in order to apply it to a classical regression model.

Another field of interest is the diagnosis of a non-identified fault type (no historical data on this fault type). For the moment, the method we propose cannot detect if a new type of fault appeared and the procedure randomly attributes it to one already identified type of fault. So, extension of the method with some procedure like described in [44] would be interesting.

6 ACKNOWLEDGMENTS

Sylvain VERRON is Supported by a PhD purpose grant from "Angers Loire Métropôle". The authors gratefully acknowledge the contribution of the reviewers comments.

A About mutual information

This appendix presents the demonstration of the equation 14 which is the mutual information between a multinomial variable and a multivariate normal variable.

As demonstrated in chapter 9 of [50], the entropy h of a multivariate normal distribution of dimension p can be written as:

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{x}} P(\boldsymbol{x}) \log(P(\boldsymbol{x}))d\boldsymbol{x} = \frac{1}{2}\log\left((2\pi e)^p|\boldsymbol{\Sigma}|\right)$$

where $P(\boldsymbol{x})$ represents the density function of a p normal variable, as given in the equation 6. The definition of the mutual information gives:

$$\begin{aligned}
I(\boldsymbol{X};C) &= \sum_{c=1}^{k}\int_{\boldsymbol{x}} P(c,\boldsymbol{x})\log\left(\frac{P(c,\boldsymbol{x})}{P(c)P(\boldsymbol{x})}\right)d\boldsymbol{x} \\
&= \sum_{c=1}^{k}\int_{\boldsymbol{x}} P(c)P(\boldsymbol{x}|\,c)\log\left(\frac{P(c)P(\boldsymbol{x}|\,c)}{P(c)P(\boldsymbol{x})}\right)d\boldsymbol{x} \\
&= \sum_{c=1}^{k} P(c)\int_{\boldsymbol{x}} P(\boldsymbol{x}|\,c)\log\left(P(\boldsymbol{x}|\,c)\right)d\boldsymbol{x} \\
&\quad -\sum_{c=1}^{k}\int_{\boldsymbol{x}} P(c)P(\boldsymbol{x}|\,c)\log\left(P(\boldsymbol{x})\right)d\boldsymbol{x}
\end{aligned}$$

We can see that the integral of the first term is the definition of the entropy of a multivariate normal distribution with mean $\boldsymbol{\mu_c}$ and covariance matrix $\boldsymbol{\Sigma_c}$. The second term can be developed as follow:

$$\begin{aligned}
&\sum_{c=1}^{k} \int_{\boldsymbol{x}} P(c)P(\boldsymbol{x}|\, c) \log\left(P(\boldsymbol{x})\right) d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \sum_{c=1}^{k} P(\boldsymbol{x}, c) \log\left(P(\boldsymbol{x})\right) d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} P(\boldsymbol{x}) \log\left(P(\boldsymbol{x})\right) d\boldsymbol{x} \\
&= -\frac{1}{2} \log\left((2\pi e)^p |\boldsymbol{\Sigma}|\right)
\end{aligned}$$

then,

$$\begin{aligned}
I(\boldsymbol{X}; C) = &\sum_{c=1}^{k} P(c) \left(-\frac{1}{2} \log\left((2\pi e)^p \, |\boldsymbol{\Sigma_c}|\right)\right) \\
&+\frac{1}{2} \log\left((2\pi e)^p \, |\boldsymbol{\Sigma}|\right) \\
= &-\frac{1}{2} \log\left((2\pi e)^p\right) - \frac{1}{2} \sum_{c=1}^{k} P(c) \log\left(|\boldsymbol{\Sigma_c}|\right) \\
&+\frac{1}{2} \log((2\pi e)^p) + \frac{1}{2} \log(|\boldsymbol{\Sigma}|) \\
= &\frac{1}{2} \left[\log\left(|\boldsymbol{\Sigma}|\right) - \sum_{c=1}^{k} P(c) \log(|\boldsymbol{\Sigma_c}|)\right]
\end{aligned}$$

References

[1] V. Venkatasubramanian, R. Rengaswamy, K. Yin, S. Kavuri, A review of process fault detection and diagnosis part i: Quantitative model-based methods, Computers and Chemical Engineering 27 (3) (2003) 293–311.

[2] L. H. Chiang, E. L. Russell, R. D. Braatz, Fault detection and diagnosis in industrial systems, New York: Springer-Verlag, 2001.

[3] D. H. Stamatis, Failure Mode and Effect Analysis: FMEA from Theory to Execution, ASQ Quality Press, 2003.

[4] B. Dhillon, Reliability, Quality, and Safety for Engineers, CRC Press, 2005.

[5] R. J. Patton, P. M. Frank, R. N. Clark, Issues of Fault Diagnosis for Dynamic Systems, Springer, 2000.

[6] D. C. Montgomery, Introduction to Statistical Quality Control, Third Edition, John Wiley and Sons, 1997.

[7] W. A. Shewhart, Economic control of quality of manufactured product, New York : D. Van Nostrand Co., 1931.

[8] E. S. Page, Continuous inspection schemes, Biometrika 41 (1954) 100–115.

[9] S. W. Roberts, Control chart tests based on geometric moving averages, Technometrics 1 (3) (1959) 239–250.

[10] H. Hotelling, Multivariate quality control, Techniques of Statistical Analysis (1947) 111–184.

[11] C. A. Lowry, W. H. Woodall, C. W. Champ, S. E. Rigdon, A multivariate exponentially weighted moving average control chart, Technometrics 34 (1) (1992) 46–53.

[12] J. Pignatiello, G. Runger, Comparisons of multivariate cusum charts, Journal of Quality Technology 22 (3) (1990) 173–186.

[13] D. M. Hawkins, Regression adjustment for variables in multivariate quality control, Journal of Quality Technology 25 (3) (1993) 170–182.

[14] N. Doganaksoy, F. W. Faltin, W. T. Tucker, Identification of out of control quality characteristics in a multivariate manufacturing environment, Communications in Statistics - Theory and Methods 20 (9) (1991) 2775–2790.

[15] M. K. Chua, D. C. Montgomery, Investigation and characterization of a control scheme for multivariate quality control, Quality and Reliability Engineering International 8 (1992) 37–44.

[16] R. L. Mason, N. D. Tracy, J. C. Young, Decomposition of t^2 for multivariate control chart interpretation, Journal of Quality Technology 27 (2) (1995) 99–108.

[17] D. Montgomery, Contributors to a multivariate statistical process control chart signal, Communications in Statistics - Theory and Methods 25 (10) (1996) 2203–2213.

[18] G. Zhang, A new type of control charts and a theroy of diagnosis with control charts, in: World Quality Congress Transactions, Beijing Inst of Post &, Telecommunication, Beijing, China, Beijing Inst of Post & Telecommunication, Beijing, China, 1984, pp. 175–185.

[19] N. D. Tracy, J. C. Young, R. L. Mason, Bivariate control chart for paired measurements, Journal of Quality Technology 27 (4) (1995) 370–376.

[20] C. Fuchs, Y. Benjamini, Multivariate profile charts for statistical process control, Technometrics 36 (2) (1994) 182–195.

[21] T. Kourti, J. F. MacGregor, Multivariate spc methods for process and product monitoring, Journal of Quality Technology 28 (4) (1996) 409–428.

[22] D. L. D. Micheaux, Indicateurs multivariés pour une maîtrise globale de processus, Qualita (2001) 143–151.

[23] S. Yoon, J. MacGregor, Fault diagnosis with multivariate statistical models part i: Using steady state fault signatures, Journal of Process Control 11 (4) (2001) 387–400.

[24] T. Tiplica, A. Kobi, A. Barreau, Synthèse et comparaison des méthodes pour la

maîtrise statistique des processus multivariés, in: Actes du congrès QUALITA, Annecy, France, 2001, pp. 134–142.

[25] E. J. Jackson, Multivariate quality control, Communication Statistics - Theory and Methods 14 (1985) 2657 – 2688.

[26] J. Westerhuis, S. Gurden, A. Smilde, Standardized q-statistic for improved sensitivity in the monitoring of residuals in mspc, Journal of Chemometrics 14 (4) (2000) 335–349.

[27] B. R. Bakshi, Multiscale PCA with application to multivariate statistical process monitoring, AIChE Journal 44 (7) (1998) 1596–1610.

[28] P. Nomikos, J. F. MacGregor, Monitoring batch processes using multiway principal component analysis, AIChE Journal 40 (8) (1994) 1361–1373.

[29] M.-F. Harkat, G. Mourot, J. Ragot, An improved pca scheme for sensor fdi: Application to an air quality monitoring network, Journal of Process Control 16 (6) (2006) 625–634.

[30] R. Dunia, S. Qin, T. Edgar, T. McAvoy, Identification of faulty sensors using principal component analysis, AIChE Journal 42 (10) (1996) 2797–2811.

[31] T. Kourti, J. MacGregor, Multivariate spc methods for process and product monitoring, Journal of Quality Technology 28 (4) (1996) 409–428.

[32] J. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Engineering Practice 3 (3) (1995) 403–414.

[33] B. Wise, N. Gallagher, The process chemometrics approach to process monitoring and fault detection, Journal of Process Control 6 (6) (1996) 329–348.

[34] M. Kano, K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss, B. Bakshi, Comparison of multivariate statistical process monitoring methods with

applications to the eastman challenge problem, Computers and Chemical Engineering 26 (2) (2002) 161–174.

[35] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification 2nd edition, Wiley, 2001.

[36] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[37] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1967) 21–27.

[38] T. Tiplica, A. Kobi, A. Barreau, Optimisation et maîtrise des processus multivariés. la méthode FNAD, Journal Européen des Systèmes Automatisés 37 (4) (2003) 477–500.

[39] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (2-3) (1997) 131–163.

[40] P. Langley, W. Iba, K. Thompson, An analysis of bayesian classifiers, in: National Conference on Artificial Intelligence, 1992.

[41] M. Sahami, Learning limited dependence bayesian classifiers, in: Second International Conference on Knowledge Discovery in Databases, 1996.

[42] A. Perez, P. Larranaga, I. Inza, Supervised classification with conditional gaussian networks: Increasing the structure complexity from naive bayes, International Journal of Approximate Reasoning 43 (2006) 1–25.

[43] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1-2) (1997) 273–324.
URL `citeseer.ist.psu.edu/kohavi96wrappers.html`

[44] T. Denoeux, M. Masson, B. Dubuisson, Advanced pattern recognition techniques for system monitoring and diagnosis : A survey, Journal Européen des Systèmes Automatisés 31 (9-10) (1997) 1509–1539.

[45] B. Dubuisson, Diagnostic, intelligence artificielle et reconnaissance des formes, Traité IC2 information. Série productique, Hermès sciences publications, 2001.

[46] C. Thomaz, D. Gillies, R. Feitosa, A new covariance estimate for bayesian classifiers in biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology 14 (2) (2004) 214–223.

[47] P. Huber, Robust Statistics, Wiley, 1981.

[48] A. Kobi, Diagnostic de procesus continus : application à la détection de valeurs aberrantes dans les signaux d'entrée et de sortie de système, Ph.D. thesis, Institut national polytechnique de Lorraine (1994).

[49] C. E. Shannon, A mathematical theory of communication, Bell Sys. Tech. J. 27 (1948) 379–423, 623–656.

[50] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley and Sons, 1991.

[51] G. H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: International Conference on Machine Learning, 1994, pp. 121–129.

[52] T. M. Cover, Learning in pattern recognition, s. watanabe (ed.) Edition, Methodologies of Pattern Recognition, NY, 1969.

[53] P. I. Good, Permutation, Parametric, and Bootstrap Tests of Hypotheses, Springer, 2004.

[54] J. Downs, E. Vogel, Plant-wide industrial process control problem, Computers and Chemical Engineering 17 (3) (1993) 245–255.

[55] N. Ricker, Decentralized control of the tennessee eastman challenge process, Journal of Process Control 6 (4) (1996) 205–221.

[56] J.-M. Lee, C. Yoo, I.-B. Lee, Statistical monitoring of dynamic processes based on dynamic independent component analysis, Chemical Engineering Science 59 (14) (2004) 2995–3006.

[57] U. Kruger, Y. Zhou, G. Irwin, Improved principal component monitoring of large-scale processes, Journal of Process Control 14 (8) (2004) 879–888.

[58] L. H. Chiang, R. J. Pell, Genetic algorithms combined with discriminant analysis for key variable identification, Journal of Process Control 14 (2) (2004) 143–155.

[59] L. Chiang, M. Kotanchek, A. Kordon, Fault diagnosis based on fisher discriminant analysis and support vector machines, Computers and Chemical Engineering 28 (8) (2004) 1389–1401.

[60] A. Kulkarni, V. Jayaraman, B. Kulkarni, Knowledge incorporated support vector machines to detect faults in tennessee eastman process, Computers and Chemical Engineering 29 (10) (2005) 2128–2133.

[61] A. Singhal, D. Seborg, Evaluation of a pattern matching method for the tennessee eastman challenge process, Journal of Process Control 16 (6) (2006) 601–613.

[62] P. Lyman, C. Georgakis, Plant-wide control of the tennessee eastman problem, Computers and Chemical Engineering 19 (3) (1995) 321–331.

[63] J. H. Friedman, Regularized discriminant analysis, J. Amer. Statist. Assoc 84 (405) (1989) 165–175.

[64] J. P. Hoffbeck, D. A. Landgrebe, Covariance matrix estimation and classification with limited training data, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (7) (1996) 763–767.

List of Figures

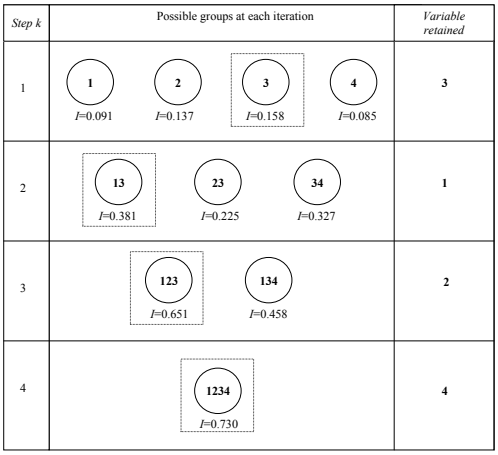

Fig. A.1. Example of the searching algorithm for a system described by 4 variables

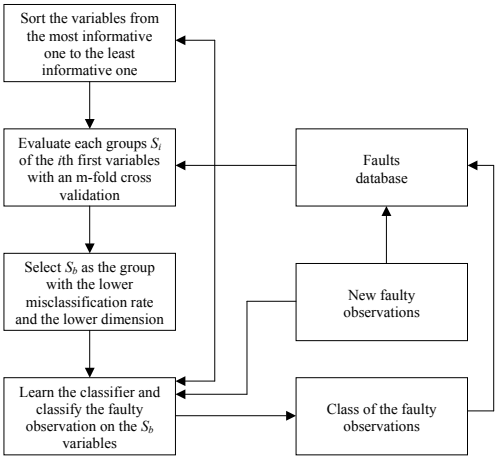

Fig. A.2. The procedure for fault diagnosis

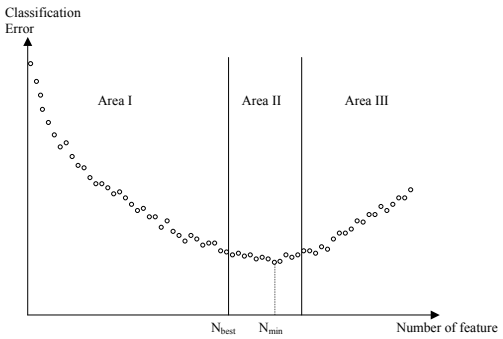

Fig. A.3. Average error function of the number of features

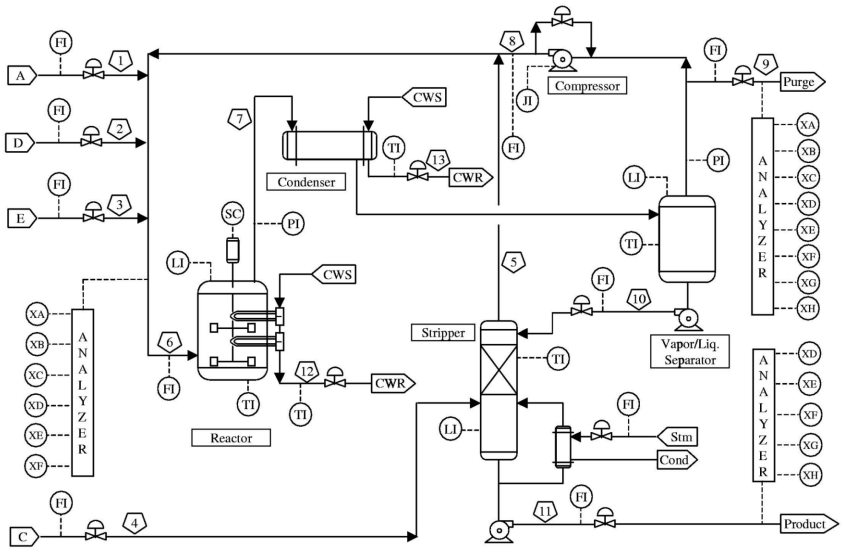

Fig. A.4. Process flowsheet of the TEP

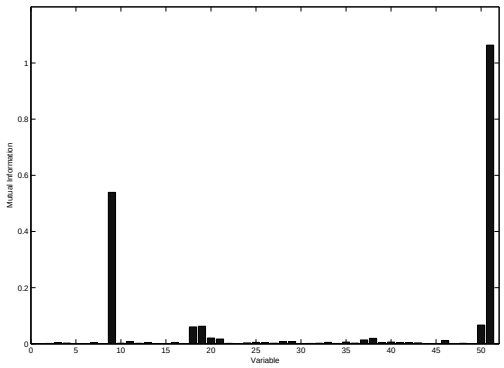

Fig. A.5. Mutual information of each variable

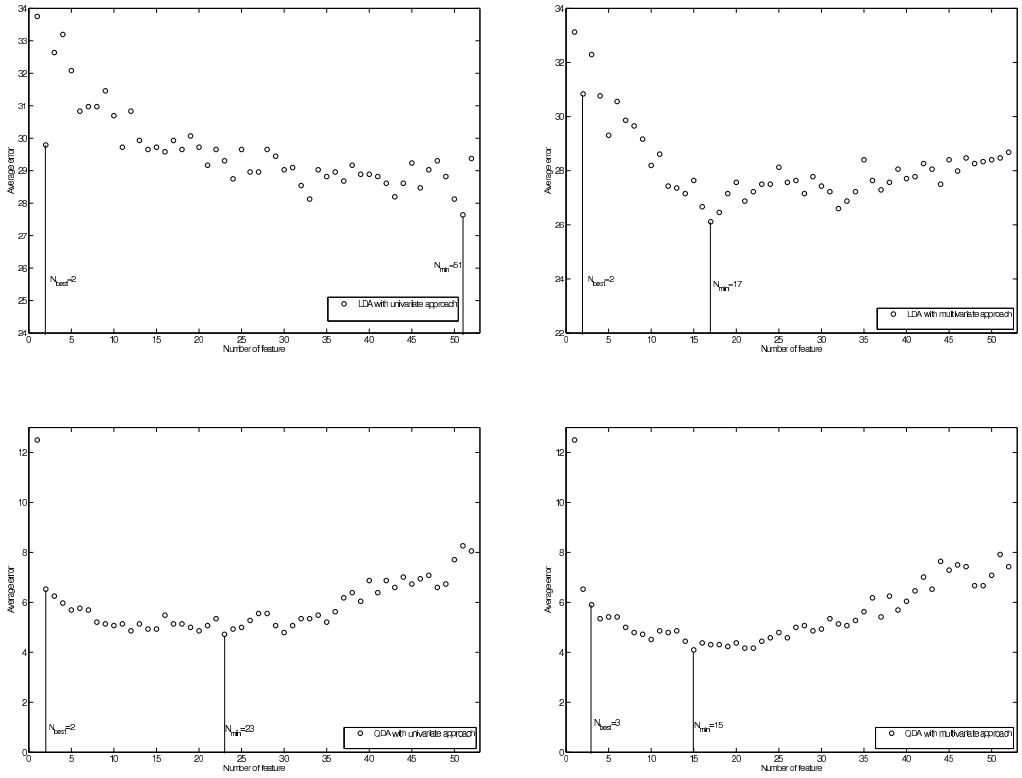

Fig. A.6. Results of the 10-fold cross validations

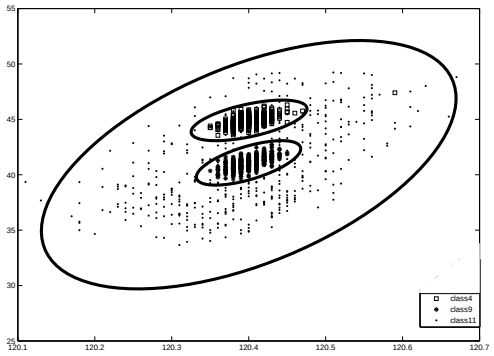

Fig. A.7. Learning data on variables 51 and 9

List of Tables

Group	S_1	S_2	S_3	S_4
Variables	{3}	{3,1}	{3,1,2}	{3,1,2,4}
Average error	26	15.5	10.5	11
Standard deviation	12.2	8.9	7.2	7.2
Hypothesis test	$S_3 - S_1$	$S_3 - S_2$	–	–
Result of the test	0	1	–	–
Best group		{2,3}		

Table A.1
Selection of the best group for the example of the 4 variables system

Class	Fault type	Train data	Test data
1	Fault 4: step change in the reactor cooling water inlet temperature	480	800
2	Fault 9: random variation in D feed temperature	480	800
3	Fault 11: random variation in the reactor cooling water inlet temperature	480	800

Table A.2
Description of fault datasets

Approach	Ordered variables
Univariate	51, 9, 50, 19, 18, 20, 38, 21
Multivariate (LDA)	51, 9, 19, 29, 16, 20, 18, 7
Multivariate (QDA)	51, 9, 21, 50, 20, 38, 7, 18

Table A.3
Order of the 8 first variables for the example of the TEP

Approach	Best group
Univariate with LDA	{51, 9}
Univariate with QDA	{51, 9}
Multivariate with LDA	{51, 9}
Multivariate with QDA	{51, 9, 21}

Table A.4
Selection of the best group for the example of the TEP

	Misclassification rate		
Method	{All}	{51,9}	{51,9,21}
480 obs/classes			
SVM	44%	6.5%	
PSVM	35%	6.0%	
ISVM	29.86%	6.0%	
LDA	42.04%	31.58%	32.87%
QDA	18.83%	5.87%	5.67%
60 obs/classes			
LDA_{60}	58.21%	30.75%	32.29%
QDA_{60}	60.75%	7.38%	7.33%

Table A.5
Misclassification rate of the different methods in the different spaces

Class	Fault 4	Fault 9	Fault 11	Total
Fault 4	794	0	34	828
Fault 9	0	777	73	850
Fault 11	6	23	693	716
Total	800	800	800	2400

Table A.6
Confusion matrix of QDA on the TEP data {9,21,51}