A new procedure based on mutual information for fault diagnosis of industrial systems

Sylvain Verron, Teodor Tiplica, Abdessamad Kobi

A NEW PROCEDURE BASED ON MUTUAL INFORMATION FOR FAULT DIAGNOSIS OF INDUSTRIAL SYSTEMS

Sylvain Verron [*,1] **Teodor Tiplica** [*]
Abdessamad Kobi [*]

[*] *LASQUO/ISTIA, University of Angers, 49000 Angers, France*

Abstract: The purpose of this article is to present a new procedure for industrial process diagnosis. This method is based on bayesian classifiers. A feature selection is done before the classification between the different faults of a process. The feature selection is based on a new result about mutual information that we demonstrate. The performances of this method are evaluated on the data of a benchmark example: the Tennessee Eastman Process. Three kinds of fault are taken into account on this complex process. The challenging objective is to obtain the minimal recognition error rate for these 3 faults. Results are given and compared on the same data with those of other published methods.

Keywords: Fault diagnosis, bayesian network classifiers

1. INTRODUCTION

Nowadays, industrial systems have more and more sensors. So, from a process, we can obtain an important amount of data. An active research field is on the utilization of these data in order to control the process. We can view the process control as a four-step procedure. In the first step, the goal is to detect an abnormal situation, a disturbance in the process: this step is named fault detection. The goal of the second step, the fault identification, is to identify the most relevant variables for the diagnosis of the fault (disturbance). The third step is the fault diagnosis, which is the purpose of this article, it consists to determinate which fault has occurred in the process. Finally, the last step is the process recovery whose aim is to return the process in a normal state.

Many data-driven techniques for fault detection can be found in the literature: univariate statistical process control (Shewhart charts) (Shewhart, 1931), multivariate statistical process control (T^2 and Q charts) (Hotelling, 1947), and some PCA (Principal Component Analysis) based techniques like Multiway PCA or Moving PCA (Bakshi, 1998). In (Kano *et al.*, 2002), authors make comparisons between these different techniques. For the fault identification, one of the most relevant statistical techniques is the MYT decomposition (Mason *et al.*, 1995) in which authors decompose the T^2 statistic into orthogonal components.

The fault diagnosis procedure can be seen as a classification task. Indeed, process measurements are stored in a database when the process is in control, but also in case of identified out-of-control. Assuming that the number of types of fault (classes) and that the belonging class of each observation is in the database (learning sample), fault diagnosis can be viewed as a supervised classification task whose objective is to class new observations to one of the existing classes. Many classifiers have been developed. We can cite (Duda

[1] Supported by a PhD purpose grant from "Angers Loire Métropôle"

et al., 2001) FDA (Fisher Discriminant Analysis) , SVM (Support Vector Machine), kNN (k-nearest neighborhood), ANN (Artificial Neural Networks) and bayesian classifiers. But, performances of all these types of classifiers are reduced in the space described by all the variable of the process. So, before the classification, a feature selection is required in order to obtain better performances.

In this article, we present a new data-driven procedure to diagnosis the disturbances of an industrial system. This procedure includes a feature selection of the most informative group of variables of the system. Then, a bayesian classifier can easily discriminate between the types of fault of the process. The article is structured in the following manner. In the section 2, we present different bayesian classifiers. The section 3 is the demonstration of a new result about mutual information. So, a new fault diagnosis procedure based on this result is shown in the section 4. The section 5 is an application of this procedure on a benchmark problem: the Tennessee Eastman Process. Finally, the section 6 presents conclusion and outlooks of fault diagnosis with bayesian networks.

2. BAYESIAN NETWORK CLASSIFIERS

An interesting bayesian classifier is the Condensed Semi Naïve Bayesian Network (CSNBN) (Kononenko, 1991). The principle of this classifier is to represent some variables in a joint node. So, some normally distributed variables can be modeled with a node representing a multivariate normal distribution. In this way, all correlations of the system will be taken into account. A CSNBN will be composed of two nodes: the class node and a multivariate node. The CSNBN is equivalent to the discriminant rule of the quadratic discriminant analysis (Duda *et al.*, 2001). Although this classifier is well fitted for the classification task of faults in industrial systems, it still remains a problem. If we have non-informative descriptors, the performances (in term of correct classification rate) are poor. So, if we want to diagnosis a system which has many variables, even though only few are really important for classification, other (less important) must not be taken into account. We have to do a selection of the important variables for classification. We will use mutual information for representing the importance of a group of variables in the classification (discrimination) task.

3. A NEW RESULT ABOUT MUTUAL INFORMATION

In the information theory, the Mutual Information (I), or transinformation, of two random variables x and y is a quantity measuring the mutual dependence of the two variables (Cover and Thomas, 1991). It can be expressed by equation 1.

$$I(x;y) = \sum_{x,y} p(x,y) ln \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

So, if we compute mutual information between the class variable and each descriptor, we can know which descriptor is important for classification and quantify this importance (Perez *et al.*, 2006). But, main problem of this approach is the redundancy. Indeed, assuming two variables with high mutual information with the class variable, to have these 2 variables in the model is not optimal if they share the same information with C (redundancy of the information). So, more interesting is the computation of the mutual information between the class node of a CSNBN and his multivariate node. For multivariate nodes of same dimension, we can evaluate the most informative one. We demonstrate a novel result about the mutual information between a multivariate gaussian (multivariate normally distribution) variable and a multinomial (discrete) variable. This mutual information can be computed as indicated by equation 2. For this equation, it is assumed that: C is a multinomial random variable with r 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 $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$; $\boldsymbol{X}$ conditioned to $C = c$ follows a multivariate normal density function with parameters $\boldsymbol{\mu_c}$ et $\boldsymbol{\Sigma_c}$.

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

Proof of this result is demonstrated in the following equations. As demonstrated in chapter 9 of (Cover and Thomas, 1991), entropy of a multivariate normal distribution of dimension p can be written as:

$$h(\boldsymbol{X}) = -\int_x p(x) log p(x) dx = \frac{1}{2}\log(2\pi e)^p |\boldsymbol{\Sigma}| \tag{3}$$

And, definition of mutual information gives:

$$
\begin{aligned}
I(\boldsymbol{X};C) &= \sum_{c=1}^{r} \int_x p(c,x) \log \frac{p(c,x)}{P(c)p(x)} dx \\
&= \sum_{c=1}^{r} \int_x P(c)p(\,x|\,c) \log \frac{p(\,x|\,c)}{P(c)p(x)} dx \\
&= \sum_{c=1}^{r} P(c) \int_x p(\,x|\,c) \log p(\,x|\,c) dx \\
&\quad - \sum_{c=1}^{r} \int_x P(c)p(\,x|\,c) \log p(x) dx \qquad (4)
\end{aligned}
$$

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:

$$
\begin{aligned}
&\sum_{c=1}^{r} \int_x P(c)p(\,x|\,c) \log p(x) dx \\
&= \int_x \sum_{c=1}^{r} p(x,c) \log p(x) dx \\
&= \int_x p(x) \log p(x) dx \\
&= -\frac{1}{2} \log(2\pi e)^p |\boldsymbol{\Sigma}| \qquad (5)
\end{aligned}
$$

then,

$$
\begin{aligned}
I(\boldsymbol{X};C) &= \sum_{c=1}^{r} P(c) \left(-\frac{1}{2} \log(2\pi e)^n |\boldsymbol{\Sigma_c}| \right) \\
&\quad + \frac{1}{2} \log(2\pi e)^n |\boldsymbol{\Sigma}| \qquad (6) \\
&= \frac{1}{2} \left[\log(|\boldsymbol{\Sigma}|) - \sum_{c=1}^{r} P(c) \log(|\boldsymbol{\Sigma_c}|) \right] \qquad (7)
\end{aligned}
$$

So, I can be computed for all different groups of variables (descriptors) of an industrial system. The most important group of variables for the classification task will be those that have an important I value. Of course, comparison of I can only be done between variables groups of same dimension. Indeed, adding variables to the model increases information of the model. We can also give this result in the case of the univariate case for the normal variable. Indeed, for the case of a distribution $N(\mu, \sigma^2)$, $|\Sigma| = \sigma^2$ and the equation 3 becomes:

$$
I(X;C) = \frac{1}{2} \left[\log(\sigma^2) - \sum_{c=1}^{r} P(c) \log(\sigma_c^2) \right] \qquad (8)
$$

The result of this equation 8 has been demonstrated in (Perez *et al.*, 2006) and corresponding to a special case of the new demonstrated result of the equation 2.

4. PROCEDURE FOR FAULT DIAGNOSIS

The objective of this procedure is to select a group of variables S giving good performances of discrimination between the different faults (classes) of an industrial system. The optimal solution to obtain a good classifier would be to estimate the misclassification rate of all the possible groups S. But, assuming an industrial system with p variables, the number of possible groups that can be constructed is $\sum_{i=1}^{p} \binom{p}{i}$. So, for example, given a system of 20 variables, we must construct 1048575 different groups and evaluate them with a cross validation technique. Of course, this solution can be effective for system with few variables, but we consider that 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 fault diagnosis of an industrial system with p descriptors (variables). This procedure is composed of four main steps: firstly, search the best group S_k of k variables for $k = 1$ to p (obtaining p groups); secondly, for each group S_k selected at the first step, evaluate the misclassification rate (average and standard deviation) of a m-fold cross validation with the classifier; thirdly, select the group giving the lower average error with the lower dimension. Finally, learn the classifier on the total learning data and classify the new faulty observations. Of course, once that the diagnosis is correct, the new faulty observations and the class of the fault are added to the faults database. The schema of this procedure can be found on the figure 1. As the fourth step is basic, we will develop only the first three steps in the next sections. These three first steps represent a new heuristic search in the space of all possible groups of variables in order to select the group giving the best results for future classification.

4.1 Step 1: search in the space of possible groups variables

Always assuming an industrial system of p variables, the aim of this first step is to select the best (the more informative on the class) group of variables S_k for each possible dimension k (so, for $k = 1$ to p). For that, we will use the new result that we have demonstrated in the section 3: the mutual information between a multivariate gaussian variable and a multinomial variable. The method consists to search the most informative variable ($k = 1$) and select S_1 corresponding to the variable maximizing I computed with the

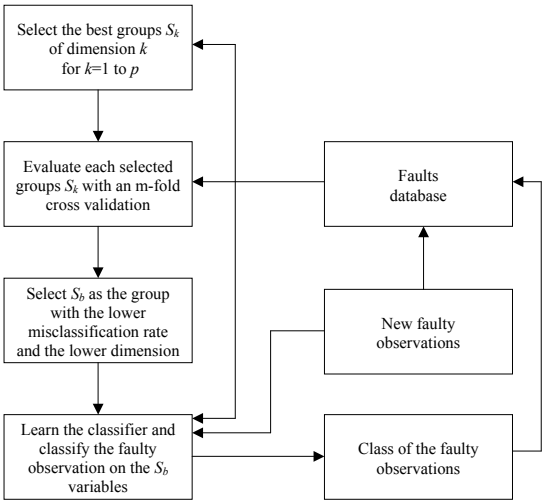

Fig. 1. The procedure for fault diagnosis

equation 8. In the next step the mutual information I of all possible groups of dimension k ($k = 2$) comprising S_1 are computed with equation 2, and the group having the maximum I is chosen as S_2, and iteratively so on. So, in this approach, we start with no variable and we add one iteratively. An example of the search for a 4 variables system is given on the figure 2.

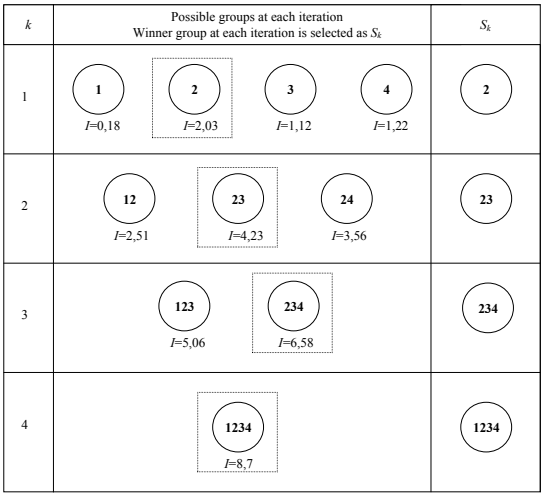

Fig. 2. Example of the search for a 4 variables system

At the end of one of these searches (forward or backward), $\sum_{i=1}^{p} \binom{i}{i-1}$ groups of variables have been evaluated (for $p = 20$ it is equal to 210) and p groups have been selected. But, it is still necessary to identify a performing group for classification (fault diagnosis).

4.2 Step 2: evaluation of the selected groups of variables

The objective of this second step is the evaluation, with the classifier, of the p different groups of variables that have been selected in the first step. Of course, the evaluation is done only on the training dataset (database). For this evaluation procedure, we apply a well known technique: the m-fold cross validation (Cover, 1969). In the m-fold cross validation, the training dataset is divided into m subsets. So, for each m, one of the m subsets is used as the testing set, and 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 (Duda *et al.*, 2001).

4.3 Step 3: selection of the best group of variables

The aim of this step is to select the best group of variables for the fault diagnosis. The first idea is the selection of the group giving the lower misclassification (average error) at the step 2 (m-fold cross validation). But, this is not a good way. Indeed, we can represent the error function of the number of feature (see figure 3).

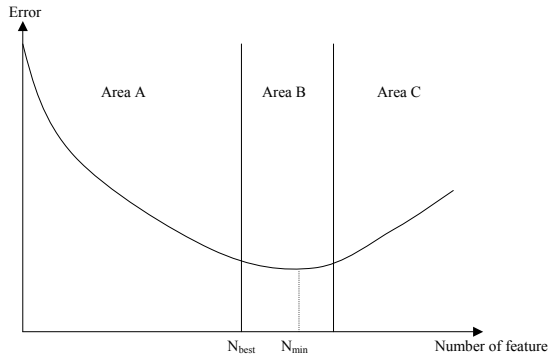

Fig. 3. Average error function of the number of features

We can distinctly see 3 areas: Area A represents an area where the error decreases when the number of feature increases, indicating that the number of feature is not quite important to obtain good discrimination; Area B is a quite constant error on some numbers of feature with N_{min} the number of feature having the lower error; and in area C, the error increases when the number of feature increases implying that too many features give noise for discriminating. So, if our goal is to select the group of variables giving the lower error but with the lower number of feature, N_{min} is not a good choice. Indeed, we can see that N_b is a better choice because the error is statistically equivalent to the error of N_{min} but with a lower number of feature. So, for this third step, the idea is to select the group S_{min} of dimension N_{min} giving the lower average misclassification rate. After that, hypothesis tests are made in order to compare the average error of S_{min} and the average error of all S_k with $k < N_{min}$. Then the group S_b (group where the average error is comparable to

the average error of S_{min} but with $N_b < N_{min}$) is selected.

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

5. APPLICATION TO THE TEP

5.1 Presentation of the TEP

The process simulator for 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 (Downs and Vogel, 1993). The Tennessee Eastman Process (TEP) is a chemical process. It is composed of five major operation units: a reactor, a condenser, a compressor, a stripper and a separator. Four gaseous reactant 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 (Downs and Vogel, 1993). 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 (Kano *et al.*, 2002; Lee *et al.*, 2004; Kruger *et al.*, 2004). Some fault diagnosis techniques have also been tested on the TEP (Chiang *et al.*, 2001; Chiang *et al.*, 2004; Kulkarni *et al.*, 2005). In (Chiang *et al.*, 2004; Kulkarni *et al.*, 2005), authors focus on only 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 others.

Consequently, we have taken into account 3 types of faults named: fault 4, 9 and 11 (see table 1). These three types of fault are good representations of overlapping data and so, are not easy to classify.

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 1. Description of fault datasets

For each type of faults, we have 2 datasets: a training sample and a testing sample, containing respectively 480 and 800 observations as indicated on the table 1. All computations implicating bayesian networks have been made on Matlab with the BNT (BayesNet Toolbox) developed by Murphy (Murphy, 2001). For an objective comparison to other methods, new faulty observations were not added to the faults database, like it was mentioned in the figure 1. The faults 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 is constant.

5.2 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 very important: 4.5036e+015 possible groups. So, we will apply the procedure that we have developed (1378 possible groups evaluated with mutual information). The search algorithm has selected the group 9,21,51. We have to classify 2400 new observations (800 of each type of fault). Result are given in the table 2. We are also giving the results of other methods on the same data. We notify that the results for the FDA (Fisher Discriminant Analysis), SVM (Support Vector Machines), PSVM (Proximal Support Vector Machines) and ISVM (Independent Support Vector Machines) methods are extracted from (Chiang *et al.*, 2004) and (Kulkarni *et al.*, 2005). In order to demonstrate the advantage of a reduced space for discrimination, we also added in the table 2 results of the different methods in the space of the 52 variables (All).

	Misclassification rate		
Method	All	9,51	9,21,51
FDA	38%	18%	
SVM	44%	6.5%	
PSVM	35%	6.0%	
ISVM	29.86%	6.0%	
CSNBN	18.83%	5.87%	5.67%

Table 2. Misclassification rate of the different methods in the different spaces

Firstly, we can see that all the methods gives better results in the reduced space than in the space of all the variables. So, as expected, a feature selection is necessary to well diagnosis the disturbances of an industrial system. Second interesting remark is the fact that on a same space, the best classifier is the CSNBN. Indeed, CSNBN outperforms all other classifiers (FDA, and SVM based classifiers) in the space of the 52 variables, but also in the reduced space 9,51. The fact that CSNBN outperforms FDA is not

surprising because FDA is a linear technique while CSNBN is a quadratic one. Although, CSNBN requires quite simple computation, its results are quite similar to SVM based techniques (SVM, PSVM, ISVM), which are techniques requiring more computational potential. Finally, we can view that the group selected with the new search algorithm, group 9,21,51, gives the best result of misclassification for this example of the TEP. So, we can say that this approach is relevant for feature selection of industrial systems.

6. CONCLUSION AND OUTLOOKS

The main interest of this article is the presentation of a new procedure for fault diagnosis of industrial processes. This procedure is a four steps procedure based on historical data process. Firstly, we have demonstrated a new result about mutual information (analytical form of the mutual information between a multivariate gaussian variable and a multinomial variable). This new result is exploited in two different algorithms whose aim is to search the best group of variables of each dimension of the process. Then, evaluation of these groups, with cross validation and hypothesis tests on the results, gives the group able to well discriminate between the types of fault of the process. This procedure has one constraint: the classifier must make the assumption that the data are normally distributed. So, we have used bayesian classifiers (NBN and CSNBN) making this assumption. The evaluation on the Tennessee Eastman benchmark problem demonstrates that the procedure is relevant for a good feature selection and for a good classification. Moreover, we remarked that on this TEP problem, the forward algorithm combined to the CSNBN classifier attains better results that other methods on the same data.

Outlooks of this procedures will be application of similar techniques, based on bayesian networks, in order to improve the fault diagnosis (cases of a non identified fault, case of non normally distributed data), but also in order to study in which way the fault detection step can be made with bayesian networks. Final goal is to have on the same procedure the fault detection and the fault diagnosis.

REFERENCES

Bakshi, Bhavik R. (1998). Multiscale pca with application to multivariate statistical process monitoring. *AIChE Journal* **44**(7), 1596–1610.

Chiang, Leo H., Evan L. Russell and Richard D. Braatz (2001). *Fault detection and diagnosis in industrial systems*. New York: Springer-Verlag.

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

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

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. John Wiley and Sons.

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

Duda, R. O., P. E. Hart and D. G. Stork (2001). *Pattern Classification 2nd edition*. Wiley.

Hotelling, Harold (1947). Multivariate quality control. *Techniques of Statistical Analysis* pp. 111–184.

Kano, M., K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss and B.R. Bakshi (2002). Comparison of multivariate statistical process monitoring methods with applications to the eastman challenge problem. *Computers and Chemical Engineering* **26**(2), 161–174.

Kononenko, Igor (1991). Semi-naive bayesian classifier. In: *EWSL-91: Proceedings of the European working session on learning on Machine learning*. pp. 206–219.

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

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

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

Mason, Robert L., Nola D. Tracy and John C. Young (1995). Decomposition of t2 for multivariate control chart interpretation. *Journal of Quality Technology* **27**(2), 99–108.

Murphy, Kevin Patrick (2001). The bayes net toolbox for matlab. In: *In Computing Science and Statistics : Proceedings of Interface*.

Perez, Aritz, Pedro Larranaga and Inaki Inza (2006). Supervised classification with conditional gaussian networks: Increasing the structure complexity from naive bayes. *International Journal of Approximate Reasoning* **In Press, Corrected Proof**, –.

Shewhart, Walter A. (1931). *Economic control of quality of manufactured product*. New York : D. Van Nostrand Co.