Identification and classification protocol for complex systems

N-H. Tran[1], M-F. Bouaziz[2] and E. Zamaï[1]

[1]*G-SCOP, CNRS UMR 5272, Grenoble-INP, France*
Ngoc-Hoang.Tran@grenoble-inp.fr
eric.zamai@grenoble-inp.fr

[2]*CRAN, CNRS UMR 7039, Lorraine University, France*
Mohammed-farouk.bouaziz@univ-lorraine.fr

ABSTRACT

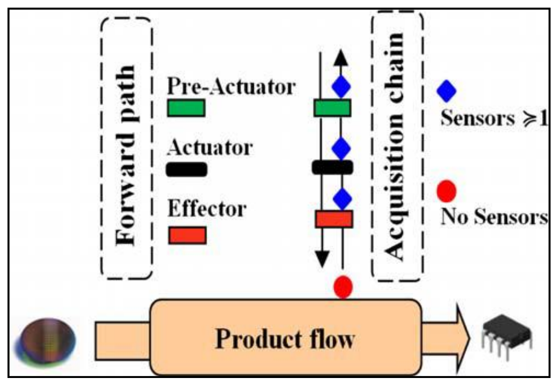
This paper proposes a test protocol for drift identification and classification in a complex production system. The key objective here is to develop a classifier for failure causes where variables depend on a set of measured parameters. In the context of our work, we assume that the drift problem of a production system is generally observed in control products phase. The model proposed in this paper for failure causes classification is structured in the form of a causes-effects graph based on Hierarchical Naïve Bayes formalism (HNB). Our key contribution in this is the methodology that allows developing failure causes classification test model in the complex and uncertain manufacturing context.

1. INTRODUCTION

Nowadays, the industrial market is characterized by capital investment and growing international competition. In this scenario, success depends on the competitiveness of products. In order to achieve this, manufacturers aim to maximize the performance and quality of services through three criteria: cycle time, costs and productivity (Kunio et al, 1995). These can only be achieved by improving manufacturing equipment availability. The manufacturing processes have become very complex and automated (Zio, 2009), and requires accuracy while executing production steps in the context of automated manufacturing systems (AMS), especially for the production equipment.

The equipment act directly on the product and they can be represented according to three parts: *(i)* the product flow that includes processed product, assembled product, finished product, etc., *(ii)* the controlled system including actuators, sensors and effectors, and *(iii)* the supervision, monitoring and control system (detection, diagnosis, prognosis, etc.), as shown in the Figure 1.

However, sensors are not directly positioned on the product for technical reasons. Therefore, the manufacturing process has the risk of not observing perturbation that affects the product quality. Also, the production equipment do not have internal mechanism to confirm that recipe applied to the product has been carried out correctly (Bouaziz et al., 2013). Therefore, many drifts are unavoidable in the production process.



Figue 1. Internal structure of the production equipment. (Bouaziz et al., 2013).

This article is structured as follows: in section 2, we present the approaches details of the identification and classification processes. Section 3 is devoted to present state of the art in the field of classification (main techniques). In section 4, we propose an introduction to Hierarchical Naïve Bayes technique. Then in section 5, we present an application of our approach on the Tennessee Eastman Process example.

2. IDENTIFICATION AND CLASSIFICATION PROCESS

In this section, we present the four steps of our methodology. The process of identification and classification is performed according to Figure 2.

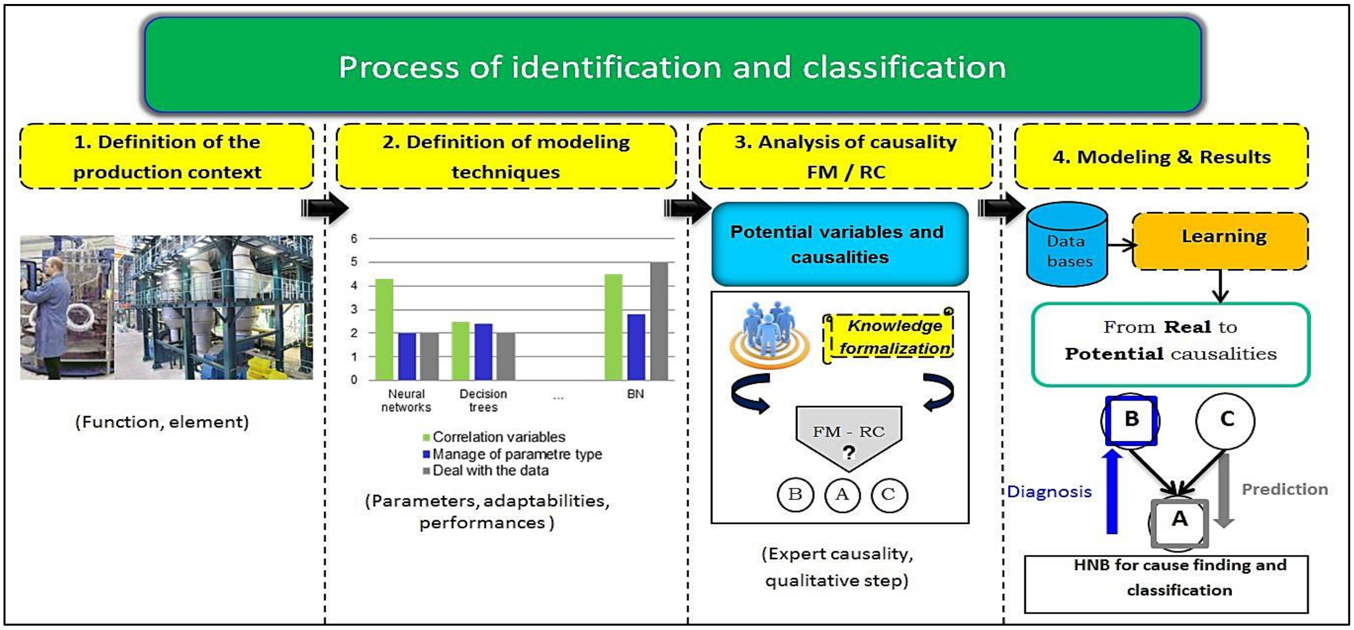


Figure 2. Identification and classification process.

2.1. Definition of the production context

This phase presents the production system context that is characterized by high complexity and uncertainty. Industrial production system is even more complex with multiple manufacturing processes running on the same production line and competing for available production resources. It means that there are a large number of elementary operations to manufacture a finished product (especially in the semiconductor and the pharmaceutical industries) and long production periods (8 to 10 weeks in semiconductor production). Also, the industrial production environment is naturally uncertain (equipment drifts, human errors...) that can impact the process control and maintenance contexts.

2.2. Definition of modeling techniques

In this second step, we analyze several methods based on the criteria defined within the production context. We analyse in particular if the model can:

- Manage diversity of the parameter types (discrete, continuous, qualitative and quantitative). Examples: time, digital measurements, samples …
- Manage multiple hierarchical classes of equipment parameters (sensors, motors...) and products.
- Manage diversity of variables: (observed variables and unobserved variables).
- Take into account correlation between variables or causal events.
- Deal with uncertain data and/or missing data (complete data and incomplete data).
- Be suitable: It is defined as the flexibility of the model for different purposes and problems (diagnosis, prognosis…).
- Be efficient: it is defined as the computation time of variables distributions (performance).

After making a synthetic comparison between the different methods (Neural networks, Decision trees, BN...), we found that modeling technique must be suitable to the context of production; and, this study is oriented towards probabilistic method: Bayes Network.

2.3. Analysis of causality (FM/RC)

The FMECA (Failure Modes, Effects and Criticality Analysis) approach is used to identify a list of failure modes (FM) and root causes (RC) by the expert. It is based on the priorities which are identified for the qualitative classification of failure modes by experts based on their knowledge. It results in the list of causalities (correlation between variables) (Bouaziz et al., 2013).

2.4. Modeling

In this last phase, we propose a mechanism to verify the causalities proposed by experts and/or find new causalities (Zaarour et al., 2004). An automated tool is proposed for this purpose that searches correlations by classification (Bouaziz et al., 2013) by learning them from a historical database.

The classifiers inputs (parameters and graphical structure) are calculated from the measured data and experts' knowledge. The output tool helps to make decisions to either verify and/or find existing or new causalities by calculating various probability distributions of the graphical model. In our case, as we propose to work in both diagnosis and prognosis; hence, we present a generic methodology for developing a simulation tool to assist this decision making.

3. THE TECHNIQUES IN THE FIELD OF CLASSIFICATION

Thereafter this section is designed to introduce techniques in the field of classification. It is necessary to know that in our case the classification phase is used for diagnosis/prognosis aspects. That is to say, the objective of this phase is to present a study of different types of classifiers with their advantages and disadvantages in the context that there is no single classifier that is better in all applications. We distinguish the classification algorithms in two categories as supervised and unsupervised classification. This section is dedicated to introduce some techniques often used in the each of these categories.

3.1. The supervised classification

In the process, when a failure causes are diagnosed, we classify the collected data according to different causes associated with degradation. The key purpose of supervised classification is to find, from the examples already classified (training sets), a model to predict the classes for new data. Following is the list of supervised classification methods used more often:

- K-nearest neighbors (k Nearest Neighborhood or kNN): The idea of this method is to observe the k nearest neighbors of a new observation to determine the class membership of this new observation (Belur, 1991). To predict the class of a new variable, the algorithm finds the K nearest neighbors of the new cases and predicts the most common response of them. This method is used on continuous data. It is possible to take into account binary data (discrete variable with 2 modalities), but not multinomial (discrete variable with n modalities) (Cover & Hart, 1967). It is difficult to find the class in case of insufficient data because it also needs a lot of examples for learning.
- Decision trees data set: It is a recognized discrimination between different classes tools. The main advantage of decision trees is that they can be easily used with the understandable rules. If the attribute is binary, we have two possible decisions, whereas if the attribute has k modalities, we have k possible decisions. Indeed, although the execution is fast, but the construction of the tree uses much more time. Also, it do not actually support the continuous values. In addition, it is always possible to discretize but the problem here is how to optimize discretization (lose the least amount of information compared to the original variable). So the decision trees work well with criteria to manage diversity parameters and variables whereas with others, they are not accurate (Verron et al, 2010).
- SVM Support Vector Machines: These are binary classifiers. The purpose of this technique is to find wide margin classifier to separate the data and maximize the distance between two classes. This linear classifier is called “hyperplane”. The closest points are called Support Vector (Verron et al, 2010). That “hyperplane” must be optimal which passes through the middle among the “hyperplanes” valid. This method has shown its effectiveness in many fields of applications such as image processing and medical diagnosis with large dimension datasets. However, the SVM application is not effective with the incomplete data.

3.2. The unsupervised classification

As we have discussed, when classes exist and that we have a large number of data already classified, we can classify new data (supervised classification). Unlike this technique, unsupervised classifications do not have a training set. There are two main families of unsupervised classification methods.

Hierarchical classification: Its purpose is to create a hierarchy in groups of variables. It means that identified classes of variables are assigned different levels.

Non-hierarchical classification: The hierarchy is not presented in this type of classification. The algorithms of this type produce classes but without forming a hierarchy (all classes are created in the same level).

- Agglomerative Hierarchical Clustering (AHC): It is a method of classification based on simple principle. We begin by calculating dissimilar objects among N. Then we combine the two objects according to criterion aggregation, thus creating a class for these two objects. We then calculate the dissimilarity between this class with other N-2 objects using this criterion to create another class. Then the two classes of objects or grouping minimizing the aggregation criterion objects are grouped. And we continue until all objects are grouped.
- Divisive Hierarchical Clustering (DHC): It is the inverse of the previous method where classes are created step by step. We initially assume that all individuals belong to the same class, and in turn we cut into two. This step is repeated until you get as many classes as individuals.
- Bayesian Networks (BN) (Pearl, 1988): This method can be used on both discrete and continuous variables. Indeed, we can build a BN model with a graph that reflects the discrete or continuous data, modeled in the probability tables. The extracted data are used for learning and the level of complexity for the

computation depends on the amount of data. A BN may represent variables by nodes and prioritization of classes with a Hierarchical Naïve Bayes networks HNB. The probabilities calculations can be provided by Maximum Likelihood Estimation / Expectation–Maximization algorithm (MLE/EM) and are used to represent correlations between nodes. Moreover, the advantage of Bayesian Networks is its adaptability. A Bayesian Network allows the consideration of the temporal dimension using Dynamic Bayesian Networks DBN (Verron et al., 2010).

In this paper, we want to remind that our study is directed towards the probabilistic methods, so it is really a method that can fulfill all of these criteria. Moreover, in our study, data is not supervised with the need for Hierarchical priorities, we would present the following details of this method in the next section.

4. INTRODUCTION OF THE HNB TECHNIQUE

4.1. Background and principle

A Bayes Network is a system representing knowledge and to calculate conditional probabilities providing solutions to different kinds of problems. The structure of this type of network is simple: a graph in which nodes represent random variables and arcs are connected by conditional probabilities (uncertainty knowledge) (Jensen, 1996). These variables may be discrete or continuous, observable or unobservable, detected or not detected.

In the general case, $X = \{X_1, X_2, \ldots, X_n\}$, the joint probability distribution P(X) is written as follows:

$$P(X)=\prod_{i=1}^{n} P(X_i / Parents(X_i)) \tag{1}$$

The calculation of BN is based on the Bayes theory (Bayes, 1763):

$$P(X_2 / X_1)=\frac{P(X_1 / X_2)\cdot P(X_2)}{P(X_1)} \tag{2}$$

- P(X_1) is the a priori probability (or Marginal) of X_1.
- P(X_2/X_1) is the posterior probability of X_2 (knowing X_1).
- P(X_1/X_2) is the likelihood function of X_1 (knowing X_2).

The marginal distribution $P(X_2)$ is calculated by the formula:

$$P(X_2)=P(X_2 / X_1).P(X_1)+P(X_2 / \overline{X}_1).P(\overline{X}_1) \tag{3}$$

The Naïve Bayes Network also called Bayes classifier is the Bayes classifier with the simplest structure. This classifier is very famous because of its performance, especially in the case where all variables are discrete (Verron et al., 2010). Naïve Bayes networks have a simple and unique structure that includes two levels. The first level contains a single parent node and the second is several children with high hypothesis of conditional independence of children *(X)* to the parent. Nodes $X_1 \ldots X_n$ are independent conditional on X_c class. They are widely used to solve classification problems expressed by Eq. (4) and Figure 3:

$$P(X_c, X_1, X_2 \ldots X_n)=P(X_c)\prod_{i=1}^{n} P(X_i / X_c) \tag{4}$$

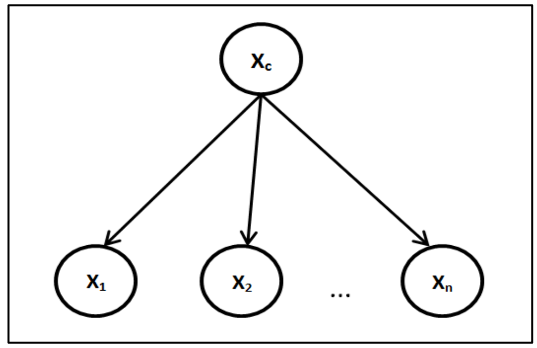


Figure 3. Naïve Bayes models.

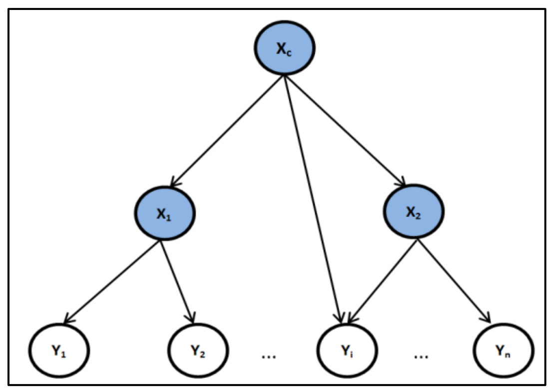


Figure 4. Hierarchical Naïve Bayes models.

In fact, the knowledge provided by an expert can also result in the creation of latent variables between two or more nodes. This is the case for example unsupervised problems where the class is never measured. Therefore, it is possible to provide the equivalent of a Naïve Bayesian network, the latent model, where classes (shown in blue in the following figure) are not part of the measured variables. A latent class (LC) model includes X_c, X_1 and X_2 latent and manifest variables Y_1, Y_2... Yn. Latent Hierarchical models illustrated in Figure 4 have been proposed by (Bishop & Tipping, 1998) for data visualization and unsupervised classification.

4.2. Learning and inference

Different families of learning and inference algorithms are proposed in the literature (Naïm et al., 2007) with three criteria of classification:

- Objective: learning or inference.
- Data: Complete or incomplete.
- Judgments of the expert: with or without the expert knowledge.

In this paper, we are working with two criteria (objective and data availability). (Bouaziz et al., 2013) presents a synthesis for probabilistic algorithms mostly used for Bayes networks. For a more detailed description, we recommend reading (Heckerman, 1998), (Neal & Hinton, 1998), (Pearl, 1988), (Jensen, 1996) and (Kappen, 2002).

These learning methods find the structure of Bayes network (structure learning) and estimate conditional probabilities (parameter learning) or acquire knowledge (experts' judgments). The inference algorithms are used for very large networks. There are many libraries for Bayes networks (BNT Matlab, BNJava, Java- Bayes, PNL…) and quality software (ProBT, BayesiaLab, Netica, Elvira…) that are useful (Naim et al., 2007).

The conditional probabilities of variables are computed based on the Bayes theory for TEP model. These results can help to make decision support components for metrology and maintenance (Bouillaut et al., 2008).

5. APPLICATION TO TENNESSEE EASTMAN PROCESS

5.1. Description

Tennessee Eastman Process (TEP) is a complex process developed by Eastman Company to provide a simulation of a real industrial process to test process monitoring methods. There are reactive gases A, C, D, E and inert gas B in the reactor. G and H are two products (liquid). The chemical reactions of the method are given by the equation system in Eq. (5).

$$\begin{aligned} &A\ (g) + C\ (g) + D\ (g) && \rightarrow G\ (liq)\\ &A\ (g) + C\ (g) + E(g) && \rightarrow H\ (liq)\\ &A(g) + E(g) && \rightarrow F\ (liq)\\ &3D(g) && \rightarrow 2F(liq) \end{aligned} \quad (5)$$

TEP has five elements: Reactor, Condenser, Compressor, Separator and Stripper. At first, the products leave the reactor while catalyst remains in there. Then the product gas is cooled through a condenser that moved to the vapor liquid separator. The uncondensed vapors in the separator return to the reactor via compressor. The inert gas B and derivative F are purged from the separator in this process. At last, the condensed stream into the separator is sent to the stripper to remove the last traces of reagents (Figure 5).

The TEP includes 53 variables: 41 measurements and 12 manipulated variables. Among these 41 variables, there are 22 continuous variables (these are the values of the sensors of the process), while other measures are compositions such as concentrations, which are not readily available but continuously sampled. TEP is subjected to 20 different faults. These faults are of different natures: step, random variation (the increasing level variability of certain variables) or other actuators such as a blocked valve. The description of these 20 mistakes and 53 variables is presented in detail in (Li & Xiao, 2011). Furthermore, we propose to work on the faults that cannot be observed (F_{16} to F_{20}).

5.2. Modeling

In our work, we propose to determine a set of variables representing the case study TEP according to steps 3 and 4 (see Figure 2). Therefore, the variables used in the illustrative models, we describe in this section, are inherently based on the experience and inference (Verron et al., 2010). Through this model, our objective is to describe the evolution and identify one or more failures in the system. We identified four distinct categories of variables:

- Failure modes of the process *FM*: We assume that the states of the variable *(FM)* takes two possible values (detected, not detected).
- Primary failure causes (level 1) RC_i (i= 1→6): these are quantitative variables defined by expert opinion. They correspond to six elements of process TEP (see table 1). All variables have a binary mode (observed or unobserved).

Node	**Variables**
RC_1	Reactor feed flow
RC_2	Reactor temperature
RC_3	Reactor pressure
RC_4	Condenser cooling water
RC_5	Separator temperature
RC_6	Stripper valve

Table 1. Primary failure causes.

- Intermediate failure causes (level 2) F_j (j=1→20). In our work, the failure causes are defined by the experts; however, for detailed description, we recommend to read (Verron et al, 2010). All faults have a binary mode (observed or unobserved).
- Parameter descriptions X_m (m=1→53): they are determined by the real process. We have 53 variables that correspond to the measurement and manipulated variables in TEP. Each variable has either a binary mode true or false.

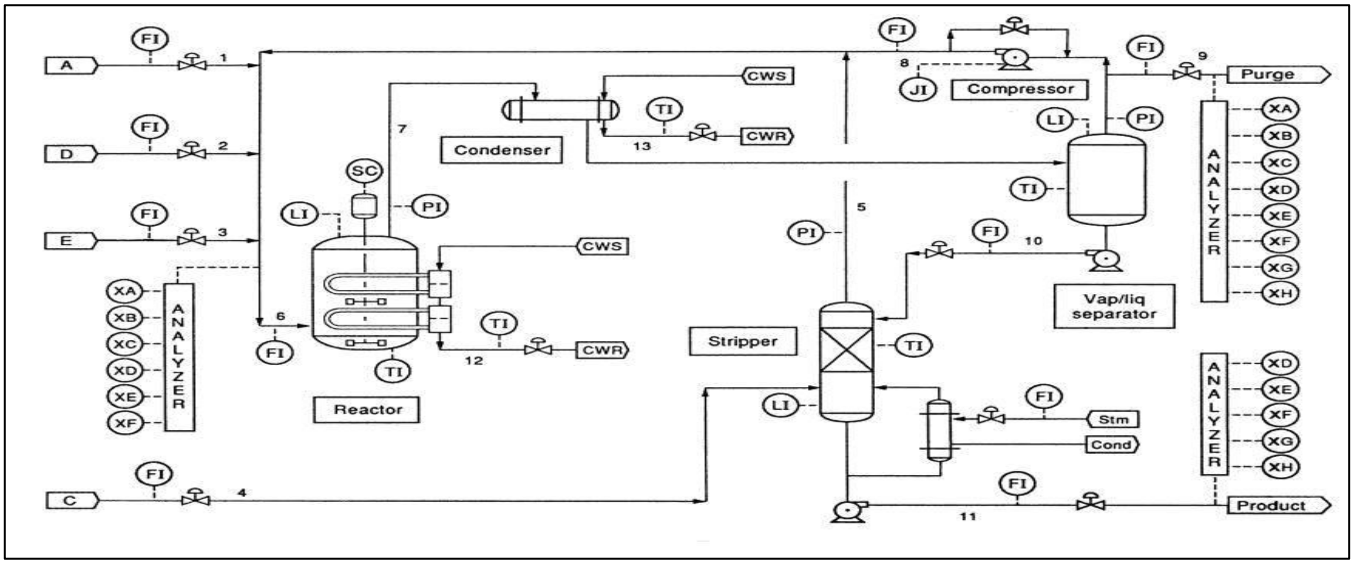


Figure 5. TEP flow sheet adopting control structure proposed by (Lyman & Georgakis, 1995).

In follows, we propose a graph structure of the model and calculate the probability distributions associated with each of variable in the graph. A classification structure from RC_i with the known observation parameters X_m and structure prognosis/diagnosis of FM based on the observations on RC_i is shown in Figure 6 below.

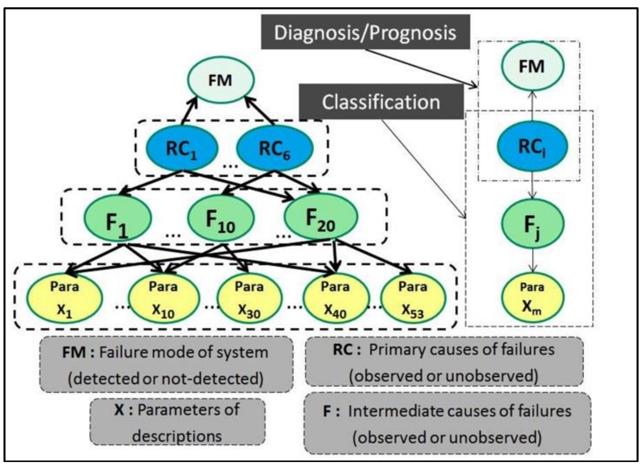


Figure 6. Identification and classification model by Hierarchical naïve Bayes network.

This model offers to classify failures causes in 2 hierarchicals RC_i and F_j. At the same time, we specify which are failures causes of the FM and predict the future state of the system or a component. To continue, our result would be presented in the next section.

6. RESULTS

In this section, first we present the preliminary results of learning with simulation in two cases complete and incomplete data.

At first, a square matrix (80 x 80) corresponding to 80 variables (53 parameters X_m + 20 variables F_j + 6 variables RC_i + 1 variable FM) and 80 samples for learning the probabilities are created by BNT Matlab © library (Murphy, 2001). The calculation of probabilities is done by MLE (Maximum Likelihood Estimation) algorithm that is a statistical estimate of the probability based on its occurrence (frequency of occurrence) in the dataset. Similarly, we have created incomplete data by adding many hidden variables in complete data.

Columns represent probabilities of variables. With FM (failure mode) variable we have 2 largest columns that represent probabilities of detected and undetected failure. We found that there are few different probability variables (Figure 7). This is unavoidable with incomplete data. However, we saw probabilities FM (failure mode) in two cases (0.77 and 0.74) is similar which is an acceptable result.

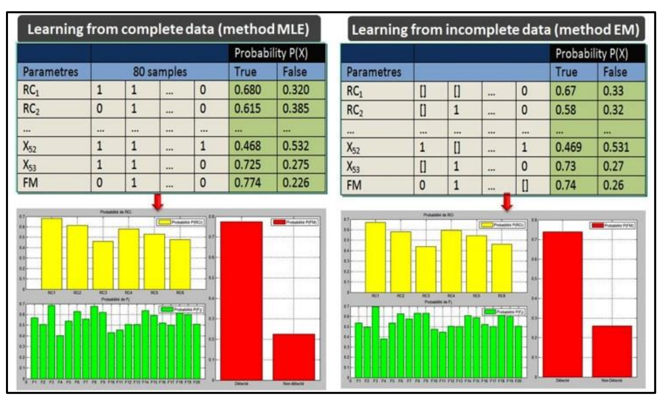


Figure 7. The results of learning algorithme.

Thereafter, we present the simulation results for the failure causes classification and prognosis after the appearance of failure. In the framework of this paper, we present un simple exemple model in Figure 8 to calculate probability distributions with the failure causes cooling water in condenser process (TEP) and related parametres.

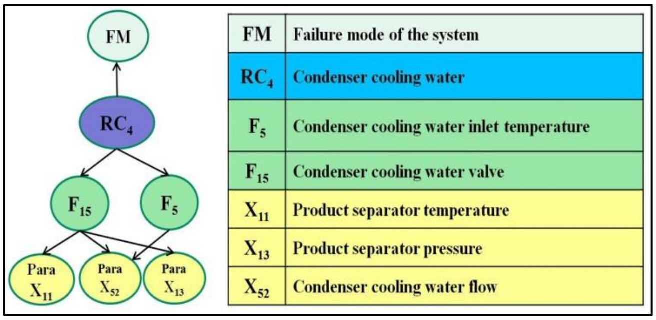


Figure 8. Exemple model to calculate probability distributions.

Figures 9 and 10 present results of two scenarios with complete data (result with incomplete data is not shown in figures). These are cleary illustrative examples of inference. We presented only probability distributions with known observation of some variables (Figure 8).

- P($FM|RC_i$): Variable observation in the example is RC_4. We used Bayes formula to calculate the probability failure mode based on this observation. Thus, the FM process is defined (predicted) from the calculation of probabilities. This is the classification model for prognosis (Figure 9).

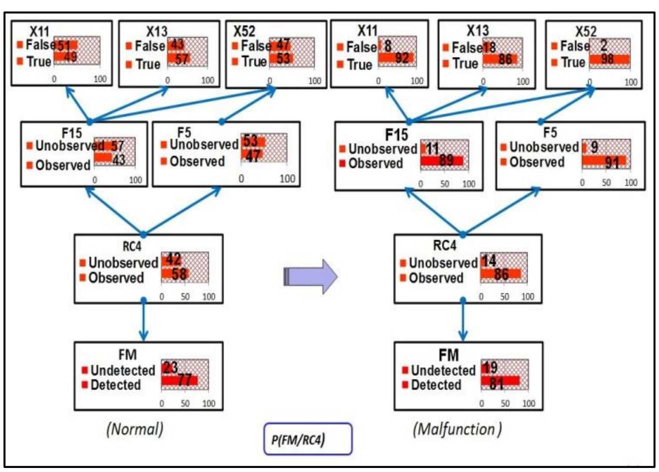


Figure 9. Probability of variables in prognosis case.

- P($RC_i|FM$): Similarly, we establish the diagnosis model when we know the observation of a failure mode. This is to calculate probabilities of the causes (for example RC_4). This is the model of classification for diagnosis (Figure 10).

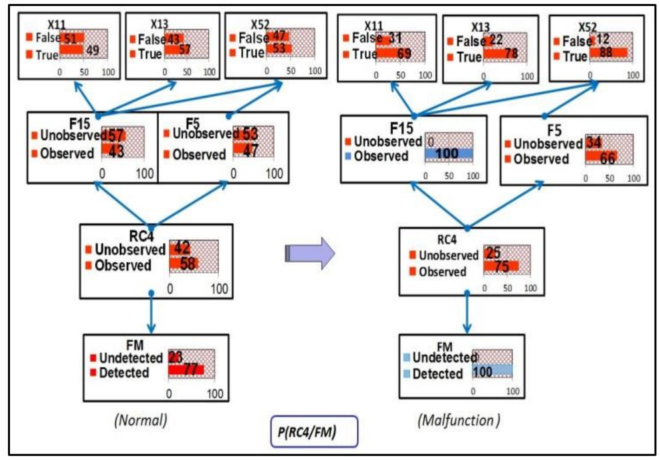


Figure 10. Probability of variables in diagnosis case.

Base on learning results, a predicted result of failure mode of process *FM* is calculated from the observed failures causes RC_4 and F_{15} (see Figure 9). We found similar inferences in both cases. Indeed, probabilistic inference is essentially a matter of calculation. This shows that learning with whether complete or incomplete data (0.81 & 0.84), we also have close probabilities to make a decision. Similarly, in diagnosis case, we found probabilities of these variables (see Figure 10) from a failure mode of process *FM* which is detected. Therefore, we can compare between probabilities to make a correct decision. So these results show that the proposed method performs good detection capability.

However, it should be mentioned that classifiers could not make choice easy if there are too many variables in the manufacturing process. This implies that we must have weights primarily depending on the differences between each variables to propose the optimal distribution.

7. CONCLUSION AND PERSPECTIVES

Our work presented in this paper deal with the identification and classification of failure causes in the context of complex industrial production. We first presented complex industrial manufacturing processes along with detailed steps of our methodology and in particular approaches for Bayes network. In the end, we presented simulation results on our TEP case study.

We showed thereafter an international benchmark that our approach propose a solution in terms of classification. In particular, we have presented a failure causes classification method based on a set of measured parameters. The resulting model, developed using Bayesian approach, allows diagnosis or prognosis in a context of complete/incomplete data. Nevertheless, this proposed model is a testing protocol for failures causes classification. Therefore, certain aspects in this model could be improved. In future, we shall propose the learning of the proposed model on real set of data that requires validation. On the other side, a development will be directed to a new configuration which is the application of a heuristic that quickly finds weights by the optimal structure of VIP classifier. In addition, an extension of the temporal Bayes network will improve dynamic monitoring for decision making.

REFERENCES

Bayes T., (1763). *An Essay towards solving a Problem in the Doctrine of Chances*. Philosophical Transactions of the Royal Society of London, vol. 53, pages 370-418.

Belur V. D, (1991) Nearest Neighbor: *Pattern Classification Techniques* , IEEE Computer Society.

Bouaziz. M.-F, Zamaï. E, Duvivier. F. (2013). *Towards Bayesian Network Methodology for Predicting the equipment Health Factor of Complex Semiconductor Systems*. International Journal of Production Research, Volume 51, Issue 15, 4597-4617.

Bouillaut L., Leray P., Aknin P., François O., Dubois S., (2008). *Dynamic Bayesian Networks Modelling Maintenance Strategies*: Prevention of Broken Rails. WCRR'08 World Congress on Railway Research, Séoul. Corea.

Bishop, C. M. and Tipping M. E., (1998). *A hierarchical latent variable model for data visualization*. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 281–293.

Cover.T.M and Hart. P.E, (1967). *Nearest neighbor pattern classification,* IEEE Transactions on Information Theory, 13 :21–27.

Heckerman D., (1998). *A Tutorial on Learning with Bayesian Network*. JORDAN M. I., Ed., Learning in Graphical Models, Kluwer Academic Publishers, Boston.

Li H. and Xiao D-Y. , (2011). *Faut diagnosic of Tennessee Eastman process using signal geometry matching technique* . EURASIP Journal on Advance in Signal Processing 2011:83.

Jensen F.V., (1996). *Introduction to Bayesian networks*, UCL Press, London.

Kappen H., (2002). *The cluster variation method for approximate reasoning in medical diagnosis*. Eds., Modeling Bio-medical signals,World-Scientic.

Kunio S., Mitsugu K., Yoshifumi K., (1995). *An Advanced step in TPM Implementation.* (pages 64-65). Paris, France.

Lyman. PR, Georgakis. C. (1995). *Plan-wide control of the Tennesse Eastman problem*. Comput Chem Eng, 19(3), 321-331. USA.

Murphy K., (2001). *The BayesNet Toolbox for Matlab*. Computing Science and Statistics: Proceedings of Infence, vol. 33.

Naim P., Wuillemin P.H., Leray P., Pourret O., (2007). *Bayesian Network*. 3e édition, Eyrolles (eds).

Neal R.M., Hinton G.E., (1998). *A View of the EM algorithm that justifies incre-mental, sparse and other variants*. JORDAN M. I., Ed., Learning in Graphical Models, Kluwer Academic Publishers, Boston.

Pearl J., (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan – Kaufmann, San Diego.

Verron. S., Li. J., Tiplica, T. (2010). *Fault detection and isolation of faults in a multivariate process with Bayesian network*. Journal of Process Control 20 (8), 902-911.

Zio, E.. (2009). *Reliability engineering: Old problems and new challenges*. Reliability Engineering and System Safety journal. Volume 94, pp. 125-141.

Zaarour I., Heutte L., Leray P., Labiche J., Eter B., Mellier D., (2004). *Clustering And Bayesian Network Approaches For Discovering Handwriting Strategies Of Primary School Children*. IJPRAI 18(7):1233-1251.

BIOGRAPHIES

Mr. Tran Ngoc Hoang is a PhD student at Grenoble GSCOP Laboratory. He has completed his degree in Mechatronics from Polytechnic University of Danang, Vietnam in 2009. He also holds a Master's degree in engineering of complex system from Grenoble INP in 2013. His research interests include equipment usage diagnosis and multi criteria decision making.

Mr. Mohammed-Farouk Bouaziz was born in Algeria, in 1985. He received a PhD in automation and production from the University of Grenoble (GSCOP), in 2012. He is currently a Postdoctoral fellow at the Research Center for Automatic Control of Nancy (CRAN). His fields of interest include the discrete event systems (DES); probabilistic graphical models (PGM); prognostics and health management (PHM); decision support systems (DSS) and the industrial applications.

Mr. Eric Zamaï was born in France in 1971. He received a Ph.D with distinction in electrical engineering from the University of Toulouse, France, in 1997. He is currently an associated professor at the Grenoble Institute of Technology (Grenoble INP) and does his research at the Laboratory of Grenoble for Sciences of Conception, Optimisation and Production (G-SCOP). His research interests include supervision, diagnostic, prognostic, and management and control of production systems. He also teaches in Grenoble INP engineering school of ENSE3. His subjects are design of real time system, control and management, logic and PLC programming.